Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: http://www.elsevier.com/locate/neuropsychologia

Audio-visual spatial alignment improves integration in the presence of a competing audio-visual stimulus

Justin T. Fleming [a], Abigail L. Noyce [b], Barbara G. Shinn-Cunningham [c,*]

[a] Speech and Hearing Bioscience and Technology Program, Division of Medical Sciences, Harvard Medical School, Boston, MA, USA
[b] Department of Psychological and Brain Sciences, Boston University, Boston, MA, USA
[c] Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA, USA

ARTICLE INFO

Keywords:
Audio-visual integration
Attention
Visual search
Electroencephalography
Temporal coherence
Spatial alignment

ABSTRACT

In order to parse the world around us, we must constantly determine which sensory inputs arise from the same physical source and should therefore be perceptually integrated. Temporal coherence between auditory and visual stimuli drives audio-visual (AV) integration, but the role played by AV spatial alignment is less well understood. Here, we manipulated AV spatial alignment and collected electroencephalography (EEG) data while human subjects performed a free-field variant of the "pip and pop" AV search task. In this paradigm, visual search is aided by a spatially uninformative auditory tone, the onsets of which are synchronized to changes in the visual target. In Experiment 1, tones were either spatially aligned or spatially misaligned with the visual display. Regardless of AV spatial alignment, we replicated the key pip and pop result of improved AV search times. Mirroring the behavioral results, we found an enhancement of early event-related potentials (ERPs), particularly the auditory N1 component, in both AV conditions. We demonstrate that both top-down and bottom-up attention contribute to these N1 enhancements. In Experiment 2, we tested whether spatial alignment influences AV integration in a more challenging context with competing multisensory stimuli. An AV foil was added that visually resembled the target and was synchronized to its own stream of synchronous tones. The visual components of the AV target and AV foil occurred in opposite hemifields; the two auditory components were also in opposite hemifields and were either spatially aligned or spatially misaligned with the visual components to which they were synchronized. Search was fastest when the auditory and visual components of the AV target (and the foil) were spatially aligned. Attention modulated ERPs in both spatial conditions, but importantly, the scalp topography of early evoked responses shifted only when stimulus components were spatially aligned, signaling the recruitment of different neural generators likely related to multisensory integration. These results suggest that AV integration depends on AV spatial alignment when stimuli in both modalities compete for selective integration, a common scenario in real-world perception.

1. Introduction

Vision and audition work synergistically to allow us to sense dynamic events in real-world settings. Often, we identify points of interest in the environment based on simultaneous auditory and visual events, such as when a car horn and flashing lights help us find our car in a crowded parking lot. Such cross-modal convergence helps the brain analyze complex mixtures of multisensory inputs. Indeed, successful integration of information between audition and vision leads to many perceptual benefits, including faster reaction times (Diederich, 1995; Miller and Ulrich, 2003), improved detection rates (Rach et al., 2011), and enhanced salience of weak stimuli (Odgaard et al., 2004).

Temporal synchrony and coherence of audio-visual (AV) inputs over time are important drivers of AV integration that affect both physiological and perceptual responses. Early studies of neural responses to multisensory stimuli noted that neurons in the superior colliculus (SC) respond most strongly to synchronized auditory and visual inputs (Meredith, Nemitz and Stein, 1987; Meredith and Stein, 1986). Similarly, participants tend to perceive auditory and visual stimuli as sharing a common source when events in the two modalities occur close enough in time that they fall within a "temporal binding window" (Stevenson et al., 2012; Wallace and Stevenson, 2014). When listening to an

* Corresponding author. 5000 Forbes Ave., Pittsburgh, PA, 15213. USA.
E-mail address: bgsc@cmu.edu (B.G. Shinn-Cunningham).

https://doi.org/10.1016/j.neuropsychologia.2020.107530
Received 13 September 2019; Received in revised form 8 June 2020; Accepted 8 June 2020

auditory target in a mixture of sounds, a visually displayed disc with a radius that changes in synchrony with the modulation of the target's amplitude enhances detection of acoustic features in the target, even though these features are uncorrelated with the coherent modulation (Atilgan et al., 2018; Maddox et al., 2015). Together, these results establish that cross-modal temporal coherence binds stimuli together perceptually, producing a unified multisensory object that is often representationally enhanced relative to its unisensory components.

The role that spatial alignment plays in real-world multisensory integration is less well understood. Physiologically, neurons in the deep layers of the SC represent space across modalities in retinotopic coordinates (Meredith and Stein, 1990). When presented with spatially misaligned auditory and visual stimuli, many neurons in the SC (Stein and Meredith, 1993) and its avian homologue (Mysore and Knudsen, 2014) exhibit suppressed responses. In non-human primates, neuronal responses in the ventral intraparietal area (VIP) are super- or sub-additively modulated by AV stimuli, but only when the stimuli are spatially aligned (Avillac et al., 2007). Spatial influences on AV integration have also been found behaviorally in humans, such as the AV spatial recalibration observed in the ventriloquist effect: when participants are presented with spatially disparate but temporally coincident AV stimuli, visual stimulus locations bias perception of auditory stimuli, reducing the perceived spatial discrepancy (Bosen, Fleming, Allen, O'Neill and Paige, 2018; Howard and Templeton, 1966; Körding et al., 2007). However, when a task does not specifically require attention to stimulus locations, AV integration is often observed even in the absence of spatial alignment. For instance, a video of a talker's face can bias perception of spoken phonemes (the McGurk effect) whether or not visual and auditory signals are aligned in space (Bertelson et al., 1994). In simple scenes with at most one visual and one auditory source, spatial alignment also has no bearing on the sound-induced flash illusion (also referred to as the flash-beep illusion), in which pairing a single brief flash with multiple auditory stimuli leads to the perception of multiple visual events (Innes-Brown and Crewther, 2009). However, in more complex scenes with competing visual and auditory sources in different locations, spatial alignment does influence the flash-beep illusion (Bizley et al., 2012). In sum, although spatial alignment affects integration in some cases, it is not a universal prerequisite for cross-modal processing or multisensory integration. Especially for relatively simple multisensory scenes, temporal coherence alone can drive multisensory integration.

A striking example of temporally driven AV integration is the pip and pop effect (Van der Burg et al., 2008). In this paradigm, participants search for a visual target (a vertical or horizontal line) amidst a large number of randomly oriented distractor lines. The target and distractors rapidly and randomly change color. If a sequence of tones is played with onsets that are temporally coherent with visual target color changes, the search time required to find the target is significantly reduced, even though these tones provide no explicit information about the target location or orientation. In the initial pip and pop study, tones were presented over headphones and provided no spatial information about the location of the target, and AV integration was driven solely by temporal synchrony. However, the question of whether the spatial relationship between the visual and auditory stimuli modulates the strength of the pip and pop effect has not been thoroughly explored. We hypothesized that the role of AV spatial alignment in the pip and pop effect would depend on the sensory context, just as it does for the flash-beep illusion. Specifically, we postulated that AV spatial alignment takes on greater importance when multiple auditory and visual signals compete for integration.

To test this hypothesis, we conducted two electroencephalography (EEG) experiments. Both experiments used a free-field auditory setup. In Experiment 1, synchronous tones could be either spatially aligned or misaligned with the visual display to test directly whether AV spatial alignment influences the pip and pop effect. If AV integration in the pip and pop effect requires that the auditory and visual stimuli both plausibly fall within the same "spatial binding window," we reasoned that AV search benefits would be weaker when the auditory and visual signals were misaligned. If, on the other hand, the effect is driven purely by temporal synchrony, irrespective of spatial alignment, search times in the two conditions would be unaffected by the spatial position of the auditory stimulus. In Experiment 2, we introduced a second AV stimulus (the AV foil), which participants were instructed to ignore. As with the first experiment, the auditory and visual components of the target and foil could be spatially aligned or misaligned, allowing us to explore whether spatial effects on multisensory integration are stronger when there are competing stimuli in the sensory environment. Briefly, in Experiment 1, we found that search times and neural signatures of AV integration were unaffected by spatially misaligning the auditory and visual stimuli in the pip and pop effect. However, spatial alignment between the senses did promote integration in Experiment 2, signaling an increased role of cross-modal spatial alignment in sensory settings with competing multisensory inputs.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Twenty healthy adults participated in Experiment 1. Two participants were removed due to excessive EEG noise, and so the final data set contained 18 participants (8 female; mean age = 22.8 years, standard deviation = 7.0 years). All participants had normal hearing, defined as thresholds below 20 dB HL at octave frequencies from 250 Hz to 8 kHz, and normal or corrected-to-normal vision. No participants reported any form of color blindness. Participants gave written informed consent, and all experimental procedures were approved by the Boston University Charles River Campus Institutional Review Board.

2.1.2. Experimental setup

The experiment was conducted in an electrically shielded, darkened, sound-treated booth. Visual stimuli were presented on a black background (0.09 cd/m^2) at a refresh rate of 120 Hz using a Benq 1080p LED monitor. The monitor was positioned on a stand directly in front of the participant and at 1.5 m distance, such that the center of the display was at approximately 0° azimuth and 0° elevation relative to eye level. Auditory stimuli were presented using RME Fireface UCX soundcard and three free-field loudspeakers (KEF E301), each driven by a separate amplifier (Crown XLS 1002). The loudspeakers were mounted on stands at 1.5 m distance from the participant and positioned at 0° and ±90° azimuth (Fig. 1A). To prevent the central loudspeaker from being occluded by the display, it was raised in elevation by approximately 5° above the horizontal plane of the eyes; the lateral loudspeakers were raised to the same elevation. This elevation offset is within typical estimates of human auditory vertical localization error (Cui et al., 2010; Razavi, 2009). Experiment control and stimulus presentation were implemented in MATLAB using the Psychtoolbox package (Brainard, 1997).

2.1.3. Task and design

The paradigm used here was similar to that developed by Van der Burg et al. in their original description of the pip and pop effect (2008). On each trial, participants were shown a visual search display with one target item, defined as a perfectly vertical or horizontal line segment, amid many randomly oriented distractor line segments (Fig. 1B). The visual search display could consist of 24, 36, or 48 total line segments, selected randomly on each trial with equal probability. The task was to find and identify the orientation of the visual target as quickly as possible while maintaining fixation on a small white cross subtending 0.07° of visual angle at the center of the display. Participants reported whether the target was vertical or horizontal with a keypress. The trial ended after the participant either responded or failed to find the target

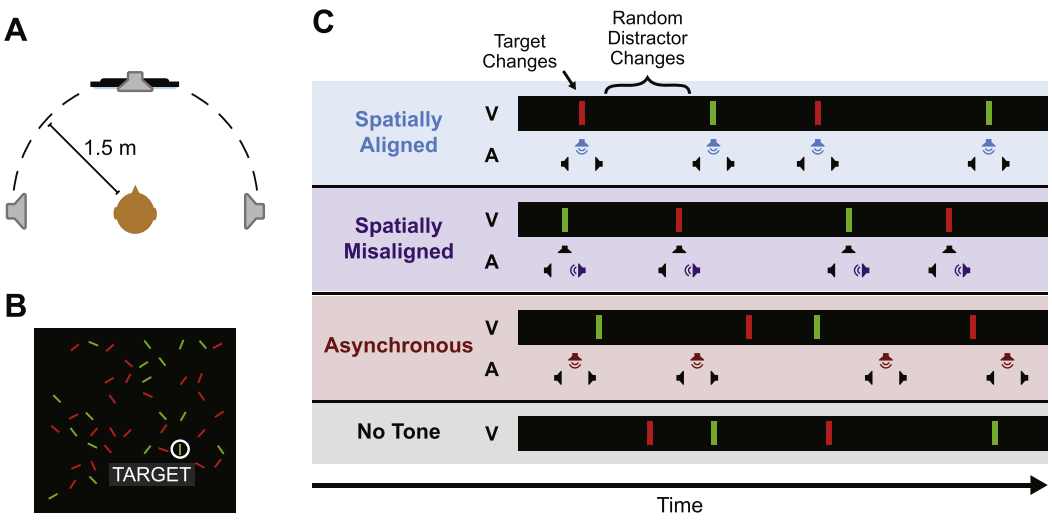

Fig. 1. Exp. 1 setup and design. **A**. A schematic of the experimental setup, as viewed from above. **B**. A snapshot of the visual stimuli during an example trial. The target was either a vertical or horizontal line, and the distractors were oriented at random angles. **C**. Time courses of the four experimental conditions. Visual (V) and (where applicable) auditory (A) stimuli are shown for each condition (for simplicity, ongoing visual distractor changes are not portrayed). The loudspeaker that presents the auditory stimuli is shown in color for each condition. Note that in the Spatially Misaligned condition, tones are portrayed as coming from the right loudspeaker, though in the actual experiment tones were presented from either the left or right loudspeaker, randomly selected on each Spatially Misaligned trial. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

within 12 s.

The search display was dynamic in that, every 50, 100, or 150 ms, either the target in isolation or a subset of distractors changed color between red and green. The intervals in which the target changed were spaced pseudo-randomly throughout the trial such that the target changed every 1.1 s on average. Distractors could change in isolation or in groups that scaled with set size: For 24 search items, 1, 2, or 3 distractors could change at each transition time; for 36 search items, 1, 3 or 5 could change; and for 48 search items, 1, 4, or 7 distractors could change (chosen with equal likelihood). In two of the four experimental conditions, a brief complex tone was played in synchrony with color changes of only the target item. The intervals preceding and following target color changes were fixed at 100 and 150 ms, respectively, to prevent spurious AV interactions between the tone and distractor items.

To test whether AV benefits in the pip and pop effect are stronger when the auditory and visual signals are spatially aligned, we manipulated spatial alignment between the visual target and the auditory tones in two conditions. On *Spatially Aligned* trials, the synchronous tones were played from the central loudspeaker, which was aligned with the azimuthal center of the visual display. On *Spatially Misaligned* trials, the tones were played from either the left or right loudspeaker. The speaker location remained fixed for the duration of each trial.

In a third *Asynchronous* condition, the same tones were played from the central loudspeaker, but with a timing that was uncorrelated with either target or distractor color changes. On these trials, the distribution of the tones' timing was statistically identical to that in the Spatially Aligned and Misaligned trials; however, the Asynchronous tones never fell within 200 ms of visual target changes. Finally, on *No Tone* trials, participants did the task using vision alone. These conditions are illustrated in Fig. 1C.

Asynchronous and No Tone trials were included for two reasons. First, they provided a baseline against which to compare search time improvements when the synchronous AV stimuli were presented. Second, they allowed us to measure neural responses to the isolated auditory and visual components of the AV stimuli, respectively. Comparing multisensory and summed unisensory neural responses can be used to test for nonlinear interactions in the multisensory evoked response (Stein and Stanford, 2008; though see Angelaki et al., 2009). In our paradigm, however, it is likely that participants' detection of the target was more closely aligned to target color changes (and corresponding auditory stimuli) in the AV conditions than in No Tone condition. Thus, neural responses to the AV stimuli may include visual contributions not present in the No Tone responses. Nonetheless, we present the summed unisensory responses here as a reference against which to compare the AV ERPs.

The experiment was divided into 12 blocks of trials: 3 Asynchronous blocks, 3 No Tone blocks, and 6 blocks of randomly intermixed Spatially Aligned and Misaligned trials. One of each type of block was presented (the order was randomized independently for each subject) before any were repeated to mitigate any potential ordering effects. In total, participants completed 108 trials per condition. Participants were given untimed breaks between blocks. Before the start of a block, an instruction screen informed participants about the upcoming block type, stating either "tones will be synced to target," "tones and target will be asynchronous," or "there will be no tone this block." Before data collection began, participants performed one practice block of intermixed Spatially Aligned and Misaligned trials and one of No Tone trials. The ratio of search times between trials with and without the synchronous tone was computed, and if this ratio was not clearly below one, participants were allowed another set of practice blocks.

2.1.4. Stimuli

Each visual stimulus in the search display was positioned within a randomly selected square in a 10 by 10 search grid. The exact positions of the line segments were then jittered to further randomize their locations. Adjacent search items were not allowed to fall within 5 pixels of each other. Each edge of the entire search grid subtended 8.5° of visual angle. The long edge of all the line segments subtended 0.44° visual angle for a stimulus directly in the line of sight. Distractor line segments' orientations were randomly assigned, but were not allowed to fall within 16° rotational angle of the cardinal vertical or horizontal orientations. Luminance values were 4.56 cd/m^2 for the red stimuli and 26.7 cd/m^2 for the green stimuli.

The auditory tones were 60 ms in duration, with onsets and offsets ramped by a 6 ms cosine-squared window to limit spectral splatter. Each "tone" was actually a chord comprising three complex tones with fundamental frequencies of 350, 467, and 556 Hz. Each component of the chord was composed of the first 10 harmonics, all set to the same intensity, to ensure that the stimulus had enough spectral bandwidth to be localized accurately. The sounds were presented at 71 dB SPL, as

measured at the location of the participant's head.

2.1.5. Passive listening follow-up experiment

To determine whether auditory stimuli presented from the central and lateral loudspeakers elicited different neural responses, a post hoc, passive listening control experiment was also conducted during a separate experimental session. We attempted to recruit the same 18 individuals who participated in Experiment 1, but only 10 were available to return for the follow-up. To increase statistical power, five additional participants were added (one of whom participated in Exp. 1 but whose data was rejected due to excessive artifacts), making for a total of 15 normal-hearing participants. EEG was collected while participants maintained fixation on a cross at the center of a black screen. Tone complexes identical to those used in Exp. 1 were presented from the three loudspeakers (left, center, and right). Tones were presented isochronously, with an inter-stimulus interval of 1 s, and in a randomized order. In total, 100 stimuli were played from each of the three loudspeakers.

2.1.6. Behavioral analysis

For all analyses (excluding error rate calculations), trials on which the participant failed to correctly identify the target as vertical or horizontal were excluded. The search time (ST) was defined as the elapsed time between the appearance of the search display and the participant's response. To calculate grand average STs for each condition, the median STs for individual participants were averaged. All comparisons were Bonferroni corrected for multiple comparisons.

2.1.7. EEG analysis

During task performance, 64-channel EEG data were collected at a sampling rate of 2048 Hz using a BioSemi ActiveTwo system. EEG data were recorded using a dedicated PC, separate from the one used for stimulus presentation. To ensure that participants did not make any major eye movements away from the fixation cross, electrooculography (EOG) data was collected using external electrodes placed on both temples and above and below the participant's left eye. The EEG preprocessing pipeline consisted of the following steps: downsampling the data to 512 Hz; bandpass filtering between 0.5 and 20 Hz; visually rejecting stochastic motion artifacts; conducting independent component analysis (ICA) to remove components of the data generated by blinks and small saccades; epoching the data between the 200 ms before and the 400 ms following each auditory stimulus (Asynchronous condition), visual target color change (No Tone), or AV stimulus (Spatially Aligned and Misaligned); rejecting epochs in which the signal exceeded a 100 μV peak-to-peak threshold; and baseline correcting each epoch to the 200-ms long time window immediately preceding stimulus presentation. All analyses were carried out using the FieldTrip MATLAB package (Oostenveld et al., 2011).

A zero-phase FIR filter with a transition width of 0.2 Hz, order of 9274, and stopband attenuation of −60 dB was used. We chose a 20 Hz low-pass cutoff because our primary aim was to study the effects of AV integration on attention in a multisensory search task. We had no strong a priori expectations of how attention might affect high-frequency EEG components in this task. We therefore focused our analysis on relatively low-frequency components of the event-related potential (ERP) known to be modulated by attention. While oscillatory activity in higher frequency bands is known to play a role in AV integration (see Keil and Senkowski, 2018), this was beyond the scope of the present work.

In one Exp. 1 dataset, an excessively noisy frontal channel was replaced with the weighted average of neighboring channels; otherwise, no channels were removed or interpolated. ERPs from trials on which the participant failed to report the correct target orientation were excluded from all analyses. Average ERP counts for all analyses, as well as the number of removed independent components for each experiment, can be found in Supplemental Table 1.

Non-parametric cluster-based permutation testing was used to assess whether differences in evoked EEG responses between experimental conditions were statistically significant. This approach limits the multiple comparison problem for high-dimensionality EEG datasets by testing the significance of "clusters" of time-channel points, rather than the highly co-dependent individual data points. For a given comparison between two conditions, average ERPs were first calculated in both conditions for each participant. Next, T-tests were computed between conditions at each time-channel sample, and clusters were formed from significant samples ($p < 0.01$) contiguous in time and within a scalp distance limit of 40 mm. For each cluster, the T-values were summed across all member samples, giving a single value (the "cluster mass") for the strength of the effect captured by the cluster (Maris and Oostenveld, 2007).

Permutation tests were used to assess the significance of cluster mass values, in which we randomly reassigned each subjects' condition labels and repeated the above procedure. On each of 2000 permutation runs, the mass of the largest cluster was calculated, forming a null distribution of cluster masses. For each cluster in the true data, its mass was compared to this distribution; p-values indicate the proportion of permutations on which a cluster of equal or greater mass to the real one was formed (Maris and Oostenveld, 2007).

3. Results

3.1. The pip and pop effect is impervious to AV spatial misalignment

The behavioral results of Exp. 1 replicate the pip and pop effect, with search times (STs) improved by a tone played in synchrony with visual target color changes (Fig. 2A). Without the synchronous tone, STs increased monotonically with set size in the Asynchronous and No Tone conditions, suggesting that participants had to search through the display serially to find the target. In contrast, when the synchronous tones were present, STs were minimally affected by increasing the number of search items, consistent with a parallelization of search (Wolfe et al., 1989). These observations are supported by the results of a two-way repeated measures ANOVA, which showed a significant interaction between the effects of condition and number of search items on ST ($F(6,102) = 3.90$, $p = 0.015$). Tukey's HSD post-hoc tests showed no significant differences between conditions on trials with 24 search items. However, AV Aligned and AV Misaligned search were both faster than Asynchronous and No Tone search on trials with 36 ($p < 0.05$ for all comparisons; see Supplemental Table 2 for full list) and 48 ($p < 0.01$ for all comparisons) search items. AV Aligned and AV Misaligned STs did not increase significantly with the addition of more search items, indicative of search "popout" in these conditions. The spatial alignment between the visual display and the loudspeaker playing the synchronous tones had no effect on the strength of the pip and pop effect. These ST effects are shown in greater detail in the distributions in Fig. 2B, collapsed across participants and number of search items. The average time of the first target color change is indicated by the dashed grey line at 715 ms. The AV Aligned and AV Misaligned distributions peak shortly after this time, indicating that participants were often able to find the target after the first presentation of the synchronized tone and visual target color change.

In most cases, participants correctly identified the target as vertical or horizontal (95% of all trials). However, error rates also varied systematically with condition and number of search items (Fig. 2C). In agreement with the ST data, error rates increased with number of search items in the Asynchronous and No Tone conditions, but not in the two AV conditions. Consistent with these observations, a two-way ANOVA revealed a significant interaction between the effects of condition and number of search items on error rate ($F(6,102) = 3.80$, $p = 0.0019$). Post-hoc tests showed that when the number of search items was largest (48), AV Aligned and AV Misaligned error rates were both significantly lower than Asynchronous and No Tone error rates ($p < 0.0001$ for all comparisons, Bonferroni corrected). Similar to the ST results, AV spatial

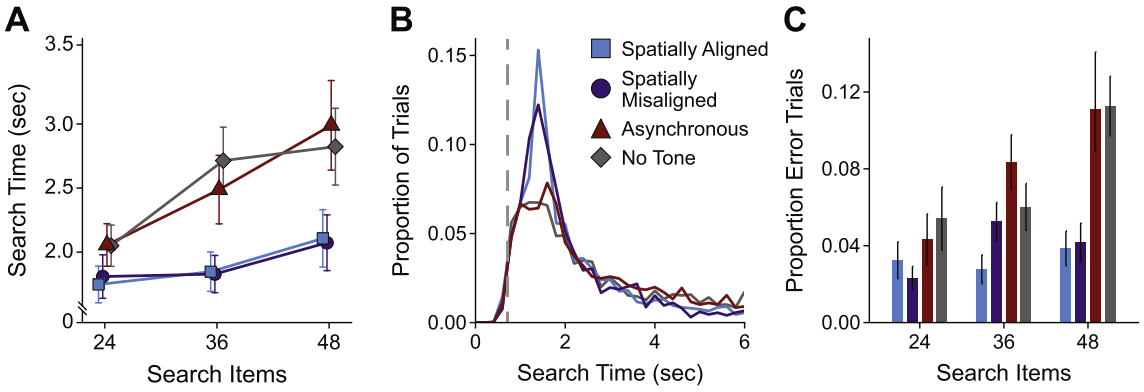

Fig. 2. Behavioral Results from Experiment 1. **A**. Average median search times across participants. **B**. Search time distributions collapsed across participants and number of search items. The grey dashed line indicates the average time of the first visual target color change, synchronous with the tone in the AV conditions. **C**. Mean proportion of error trials, defined as those on which participants incorrectly identified the target as vertical or horizontal or failed to make a response before the time limit (12 s). All error bars represent the standard error of the mean (S.E.M). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

alignment did not significantly affect error rates for any number of search items ($p > 0.99$ for each after multiple comparisons correction).

In sum, we observed consistent behavioral benefits in both search time and error rate when the tone was synchronized to the visual target color changes. Behavioral results were the same whether the visual and auditory stimuli were Spatially Aligned or Misaligned, arguing against an obligatory role of spatial alignment in the AV interactions driving the pip and pop effect.

3.2. *Spatially Aligned and Misaligned AV ERPs are both enhanced*

To examine the neural correlates of search enhancement in the pip and pop effect, we first analyzed the last auditory, visual, or AV event before participants responded on each trial. We reasoned that on AV trials, this final stimulus was the one most likely to have undergone AV integration, facilitating target detection. In the Asynchronous condition, evoked responses were measured relative to the final auditory stimulus, and in the No Tone condition, measurements were time-locked to the

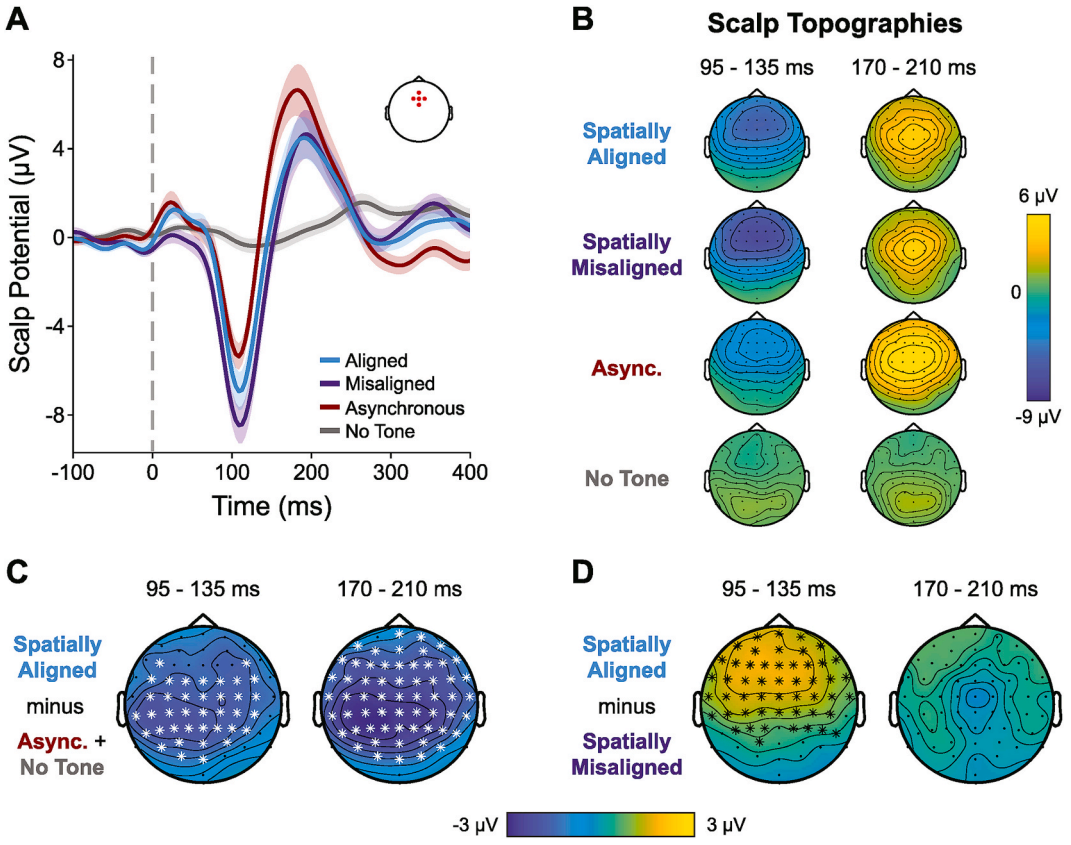

Fig. 3. Exp. 1 ERP results. **A**. Average waveforms of the last ERPs before participants responded across five fronto-central electrode sites (Fz, FCz, Cz, FC1, and FC2). **B**. Scalp topographies of evoked responses in the N1 (95-135 ms) and P2 (170-210 ms) time ranges. **C**. Difference between the Spatially Aligned responses and the sum of the isolated auditory (Asynchronous) and visual (No Tone) responses. **D.** Difference between the Spatially Aligned and Spatially Misaligned AV responses. Asterisks denote channels that are part of a significant cluster ($p < 0.01$) in permutation testing in the indicated time range.

final target color change before the behavioral response. We will refer to ERPs elicited by these final stimulus events as "Last" ERPs.

Fig. 3A shows grand average ERP waveforms averaged across 5 fronto-central electrode sites (Fz, FCz, Cz, FC1, and FC2 on the standard 10-20 layout) elicited by these stimuli. Evoked responses in the visual-only No Tone condition were relatively weak across the electrode montage; this was not surprising, as visual ERPs are known to be greatly weakened in cluttered, dynamic visual displays (Leblanc et al., 2008; Martens et al., 2006). However, both the auditory N1 component, a broadly fronto-central negative-going deflection that peaks approximately 100 ms post-stimulus, and the P2 component, a positive-going potential that peaks around 200 ms post-stimulus, showed clear differences across the other conditions. The scalp topographies of evoked responses in these time windows are shown in Fig. 3B.

To statistically assess differences between AV and unisensory responses, we used cluster-based permutation testing to compare the Spatially Aligned AV responses to the sum of the auditory-only (Asynchronous) and visual-only (No Tone) ERPs. We first performed this test over the entire post-stimulus epoch (0–400 ms) to determine when the earliest significant multisensory effects occurred. This analysis revealed a broad cluster of significant time-channel points, with the Spatially Aligned AV response diverging from the summed unisensory model by 70 ms post-stimulus ($p < 0.001$). Interpretation of this comparison may be complicated by that fact that, in the No Tone condition, participants likely detected the visual target without attending changes in its color. Because any evoked responses in this condition would be poorly time-locked to color changes, the resulting ERP would be weaker than in the AV conditions. However, since visual evoked activity was weak even in the AV conditions (see occipital electrode sites in the right panels of Fig. 3B; which encompass the expected visual N1 time range), such temporal dispersion cannot account for the observed AV enhancement over the summed unisensory model.

Because differences in both the auditory N1 and P2 could contribute to the observed effect, we next conducted separate versions of the test restricted to the N1 (95-135 ms, left panel of Fig. 3C) and P2 (170-210 ms, right panel) time ranges. These time ranges were selected by calculating the average N1 and P2 peak latencies across participants, and then expanding the windows to the 20 ms before and after each grand average peak to allow for individual differences in component latencies. Testing separately in these two windows allowed us to assess the degree to which these two distinct ERP components contributed to the broad spatiotemporal cluster.

In the N1 time window, a single cluster biased toward left parietal electrode channels was identified in which the Spatially Aligned AV responses were significantly larger (more negative) than the summed unisensory model ($p = 0.006$). This result was qualitatively similar when comparing the Spatially Misaligned AV responses to the summed unisensory model ($p < 0.001$; see Supplemental Fig. 1). In some cases, such differences between multisensory and summed unisensory responses have been interpreted as a signature of multisensory integration, but this "super-additive" property is rarely observed in human scalp recordings (Stanford and Stein, 2007; also see Angelaki et al., 2009). Since the auditory N1 is known to be strongly modulated by attention (Choi et al., 2014; Hillyard et al., 1973), we instead interpret these results as a neural correlate of the increased bottom-up salience of the multisensory target.

In the P2 time range, the Spatially Aligned AV responses were significantly smaller (more negative) than the summed unisensory model ($p < 0.001$). This effect was driven by the especially large P2 component in the Asynchronous condition. Prior studies have shown that non-target auditory stimuli that are to be ignored elicit stronger P2 responses in both pitch discrimination (Novak et al., 1992) and oddball detection (García-Larrea et al., 1992) tasks. Participants likely ignored the tones in the Asynchronous condition, as these tones had no temporal relation to the visual target and therefore could not aid search, which could explain the larger P2 response in this condition.

3.3. *Differences in Aligned and Misaligned AV responses can be explained by stimulus characteristics*

Despite a lack of behavioral differences between the Spatially Aligned and Misaligned conditions, Aligned AV stimuli evoked smaller (less negative) N1 responses than Misaligned stimuli, (Fig. 3D, $p = 0.002$). The P2 components were indistinguishable between these conditions, consistent with participants attending the AV stimuli in both conditions. Previous studies using spatialized sounds have found larger ERPs in response to lateralized sounds than central sounds (Dai et al., 2018; Palomäki et al., 2005). Since the Spatially Aligned tones were always presented from the central loudspeaker and the Spatially Misaligned tones from the left or right loudspeaker, we reasoned that the N1 differences we observed could be due to these inherent differences in auditory spatial encoding. In a brief follow-up experiment, we measured neural responses to tone complexes played from the central and lateral loudspeakers during passive listening, and then compared these to responses elicited during the AV search task.

Consistent with previous findings, The N1 and P2 components of ERPs elicited by lateral free-field stimuli tended to be larger than those elicited by central stimuli (Fig. 4A). Of particular interest is the difference in the N1 component. To calculate the N1 amplitude for each participant, average ERPs were first computed over five fronto-central electrode sites (Fz, FCz, Cz, FC1, and FC2). The peak N1 amplitude was defined as the average of the ERP minimum in the N1 time range

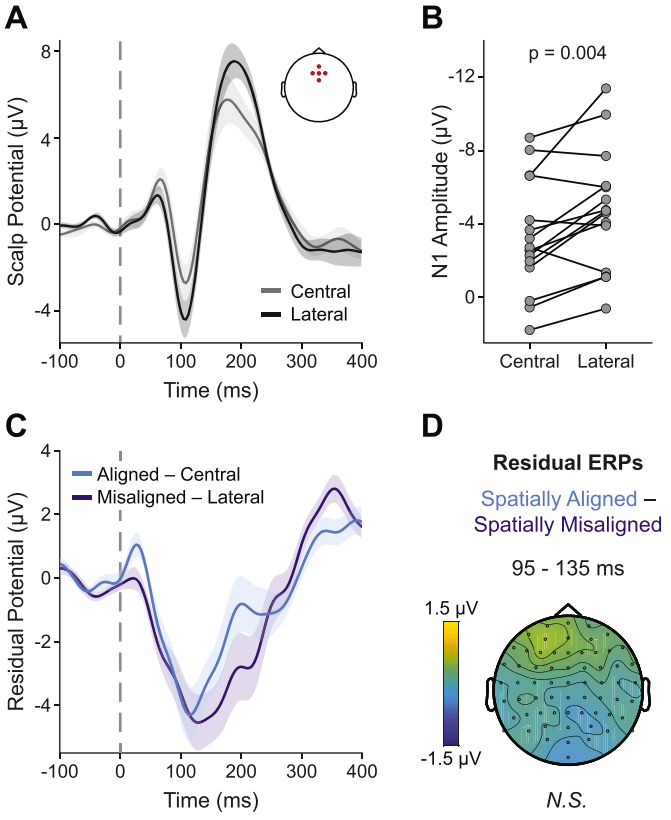

Fig. 4. ERPs elicited by passive listening to central and lateral tone complexes in the follow-up experiment. **A.** Average ERP traces (N = 15) across five fronto-central electrode sites (Fz, FCz, Cz, FC1, and FC2). Each subject's lateral ERP is the average of ERPs elicited by left and right tone complexes. **B.** Central and lateral N1 amplitudes from waveforms averaged across the same five channels. **C.** Residual ERPs from the search task, calculated by subtracting the average central passive ERP from the Spatially Aligned ERPs, and the average lateral passive ERP from the Spatially Misaligned ERPs. **D.** Scalp topography of the difference between Aligned and Misaligned residual ERPs (not significant) in the N1 time range.

(95-135 ms) and the two samples flanking it. Lateral N1 amplitudes (mean = −4.76 μV, SD = 3.29) were significantly larger than central (mean = −3.31 μV, SD = 3.10), based on a paired T-test (T = 3.39, df = 14, p = 0.004; Fig. 4B).

To test whether this difference could account for the difference between Spatially Misaligned and Spatially Aligned responses in Exp. 1, we computed residual ERPs by subtracting these passively elicited ERPs from those elicited during the search task. For the Spatially Aligned ERPs, the central passive ERPs were subtracted, and for the Spatially Misaligned ERPs, the lateral passive ERPs were subtracted; that is, we subtracted off ERPs elicited by physically identical sound sources heard passively. Since only 11 participants in the passive follow-up had participated in Exp. 1, we could not match individual responses, and so we used the grand average passive ERPs. The results of this analysis are shown in panels C and D of Fig. 4. The significant N1 difference between Spatially Aligned and Misaligned responses was abolished after accounting for the inherent differences between central and lateral tone responses. Thus, we conclude that the Spatially Aligned and Misaligned AV responses were enhanced roughly equally as compared to the summed unisensory model, mirroring the behavioral search time improvements in both AV conditions.

3.4. ERP enhancement is caused by a combination of bottom-up and top-down attention

While participants used the tone to guide search in the AV conditions, they most likely ignored it in the Asynchronous condition, as it had no relation to the visual target. Since selective attention has well-established effects on N1 amplitudes (Choi et al., 2014; Hillyard et al., 1973), the AV N1 enhancement we observed could have been caused by differences in top-down attention between the conditions, by heightened salience of the multisensory stimuli, or by a combination of these factors. To test for effects beyond sustained attentional differences, we compared ERPs evoked by the last stimulus before the participant's response on each trial (Last ERP) with the Other ERPs throughout the trial. If the N1 differences we observed were due to increased top-down attention to the tones in the AV conditions, one would expect ERPs elicited by synchronous tones to be strengthened throughout these trials, as participants were using the tones to guide search even before they found the AV target. If, on the other hand, the enhancement was driven by integration of visual targets and synchronous tones and a subsequent increase in stimulus salience, the Last ERP might be enhanced relative to the Other ERPs, as this last stimulus was likely the only one actively detected as an AV event.

For this analysis, a Last ERP was removed if the participant's response fell within 150 ms after tone presentation, as such responses are too soon after the tone to have been guided by the AV stimulus. At the level of individual participant average ERPs, there tended to be more Other ERP epochs than Last ERP epochs (see Supplemental Table 1). This is because each trial could have at most one Last ERP, whereas depending on the search time, there could be multiple Other ERPs. To test whether these unequal ERP counts would bias our comparison of Last and Other ERPs, we used a bootstrap procedure in which, over 1000 iterations, we randomly downsampled ERP counts in each condition to the minimum count obtained across participants and conditions (54 ERPs) before computing group average ERPs. On each iteration, we compared Last vs. Other ERPs in each condition using cluster-based permutation testing. We found that the effect sizes reported below were larger than those typically found using the ERP downsampling procedure (Supplemental Fig. 2, panels A and B). Critically however, the downsampling did not bias the ERP amplitude in either direction (Supplemental Fig. 2, panels C and D), and so the increased statistical strength when all ERPs were used likely resulted from reduced noise in individual participant average ERP estimates. Thus, for all analyses, we kept the maximum possible number of epochs in each participant average ERP.

We used cluster-based permutation tests, restricted to a window spanning 95–135 ms post-stimulus, to compare N1 components between the Last and Other ERPs in each of the four experimental conditions. Fig. 5 shows the grand average Last and Other ERP waveforms at channel Fz, near the strongest auditory N1 response. At this channel, Last ERPs were significantly enhanced relative to Other ERP in the N1 time range in both AV conditions (cluster p = 0.030 for Spatially Aligned, p < 0.001 for Spatially Misaligned), but not the Asynchronous or No Tone conditions. Since participants attended the tones throughout the trial in the AV conditions – even prior to finding the target – this result is inconsistent with the Last ERP differences being driven by top-down attention alone. Instead, multisensory integration of the AV events likely contributes to the observed N1 enhancements, reflecting the increased perceptual salience of the AV stimuli.

In the Asynchronous and No Tone conditions, this analysis also revealed a separate cluster, biased toward more posterior electrode sites, in which Last ERP N1s were significantly more positive (smaller) than Other ERP N1s. This positivity appeared to strengthen after the N1 time range, so we examined it using a second cluster test focused on a later time window (250–400 ms post-stimulus). In this time window, Last evoked responses were more positive than Other responses in all conditions (Supplemental Fig. 3). We attribute this effect to a weakly time-locked P3 ERP component, which is related to the detection of target stimuli that require a motor response (Conroy and Polich, 2007; Squires et al., 1975).

Finally, as an explicit test for effects of top-down attention, we compared Other (not trial-final) ERPs in the AV conditions to a summed unisensory model (Asynchronous + No Tone), also composed of Other ERPs. Since the Other ERPs preceded target detection, effects here reflect sustained attention and not integration of the AV target. As was the case with the Last ERPs (Fig. 3), in the N1 time range Other ERPs were larger in the AV conditions than the summed unisensory model

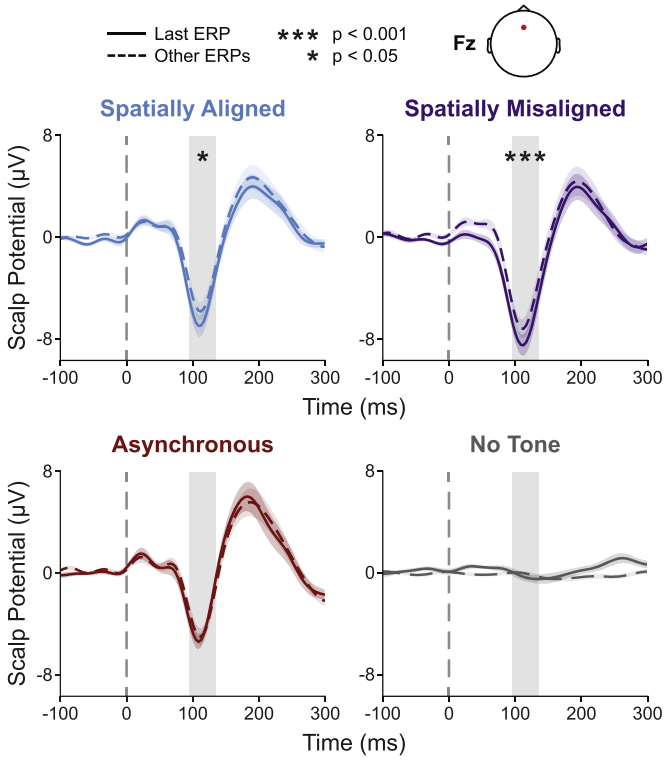

Fig. 5. Last ERPs in the trial compared to Other ERPs in Exp. 1. All traces are shown at channel Fz. Shaded regions indicate the N1 time window examined in cluster-based permutation testing (95-135 ms post-stimulus). Asterisks denote the significance level of resulting clusters. Error clouds around the waveforms represent SEM.

(cluster $p = 0.006$ for Spatially Aligned, $p < 0.001$ for Spatially Misaligned). While this enhancement might reflect a sub-threshold signature of AV integration, a more likely explanation is that top-down attentional control and an exogenous increase in multisensory stimulus salience both exerted influences on the observed neural responses, enhancing AV ERPs throughout the trial when the tones were behaviorally relevant.

4. Experiment 2

When searching for a single AV target in the Exp. 1 pip and pop task, spatial misalignment between the auditory and visual signals did not diminish the behavioral benefits or neural signatures of AV integration. However, the auditory stimuli in Exp. 1 consisted of a single stream of tones. On trials in which these tones were present, listeners knew they were either synchronized with the visual target (AV conditions) or not useful (Asynchronous condition). We hypothesized that, as in the flash-beep illusion, the spatial alignment of auditory and visual stimuli might be irrelevant for AV integration when there is strong temporal coherence between auditory and visual events and no competing sound. However, in a more complex auditory scene, in which irrelevant sounds must be ignored, the spatial alignment of auditory and visual events may become important. We therefore undertook a second experiment to explore the possibility that in more complex, ambiguous multisensory scenes, AV spatial alignment would influence the pip and pop effect.

We created a version of the pip and pop task with competing auditory and visual stimuli. On each trial, two potential visual targets were presented, one in each hemifield, as well as two sets of tones, one in each hemifield. Participants were cued to search for a visual target in one hemifield and ignore a visual "foil" in the other. The tones in one hemifield were temporally matched to the target, while the tones in the other hemifield temporally matched the foil. Importantly, the visual target could either be Spatially Aligned or Spatially Misaligned with the informative, temporally coherent tones. Thus, in the Spatially Misaligned condition, the tones synchronized to the visual target came from the hemifield opposite that of the target, and the same was true for the foil.

4.1. Methods

Experimental methods of Exp. 2 were largely the same as those for Exp. 1. Specifically, the Stimuli, Behavioral analysis, and EEG analysis were generally the same and are fully described in the Methods for Exp. 1. For brevity, here we describe only those methods that differed between the experiments.

4.1.1. Participants

Twenty healthy adults participated in Experiment 2, none of whom had completed Exp. 1. Two participants in Exp. 2 were removed due to excessive EEG noise, and another was removed due a large number of no-response trials and anomalously slow reaction times (more than three standard deviations slower than the group average across conditions). Thus, the final data set comprised 17 participants (11 female; mean age = 21.3 years, standard deviation = 2.6 years). All had normal hearing, no reported colorblindness, and gave written consent as in Exp. 1.

4.1.2. Task and design

The task in Exp. 2 was similar to Exp. 1, with the addition of an AV foil stimulus. Dynamic visual stimuli were presented as in Exp. 1, and auditory stimuli were presented from two free-field loudspeakers directly on either side of the monitor (at ±10° visual angle; see Fig. 6A). On each trial, the display contained two stimuli that could be the target, defined as horizontal or vertical line segments (as in Exp. 1). The potential visual targets always appeared on opposite hemifields of the display (Fig. 6B). The trial timeline is illustrated in Fig. 6C. Participants first received a 2-s visual cue indicating the side on which the visual *target* would appear, and they were instructed to ignore the *foil* in the opposite visual hemifield. The participant then performed the search task until they found the target or timed out (12 s). On half the trials, the target changed color first, and on the other half the foil changed color first. The target and foil were not allowed to change color within 250 ms of each other.

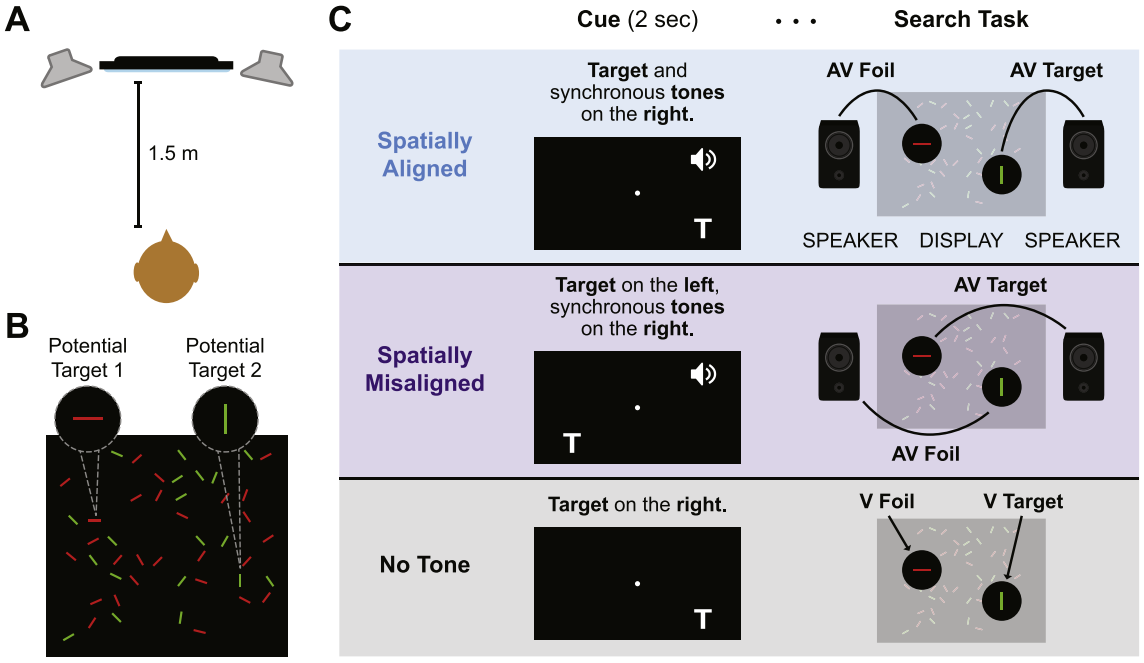

Fig. 6. Exp. 2 setup and design. **A.** A schematic of the experimental setup, as viewed from above. **B.** A snapshot of the visual stimuli during an example trial. The two potential targets always appeared on opposite hemifields of the display. **C.** Experimental conditions. A visual pre-trial cue informed participants whether the actual target would be located on the left or right, and whether tones synchronized to the target would come from the loudspeaker on the same side (Spatially Aligned), the opposite side (Spatially Misaligned), or not be present (No Tone). Tones synchronized to the foil were played from the loudspeaker not playing the target tones.

On AV trials, tones were played in synchrony with *both* target and foil color changes. The synchronous tones were identical tone complexes (as described in Exp. 1), separable only by their spatial locations (from symmetrically placed speakers). On trials containing synchronous tones, the cue informed participants which loudspeaker would play the target-synchronized tones, in addition to the hemifield of the visual target (Fig. 2C). When the tones and visual target were *Spatially Aligned*, synchronous tones played from the loudspeaker in the same hemifield as the visual target. Tones synchronized to the foil were played from the other loudspeaker, in the same hemifield as the foil. On *Spatially Misaligned* trials, the relationship was reversed; tones synchronized to the target were played from the loudspeaker in the opposite hemifield, and the same was true for the foil. Finally, *No Tone* trials required participants to do the same cued search task using vision alone.

The search grid was widened for Exp. 2, subtending 11.2° of visual angle, and targets and foils were restricted to appear in the left- and rightmost 3.2° of the display. These changes were made to minimize the physical distance between key visual stimuli and the loudspeakers, while still allowing participants to maintain central fixation while performing the task. Participants were asked after the experiment if they recognized any patterns in where targets appeared; no participants noticed the restricted target zones. Since participants were instructed to limit their search to half the display, the total number of search items was increased to 84 (spread evenly between hemifields) to increase task difficulty. The orientation of the foil was randomly chosen to be vertical or horizontal, irrespective of the target orientation, such that the target orientation could not be inferred by finding the foil. This does mean that, on some trials, the target and foil had the same orientation, making it impossible to know for certain to which stimulus participants responded. However, given the very high average hit rate (92%), we can conclude that participants were not responding to the foil with any regularity.

Prior to starting the actual experiment, all participants first completed the same training described for Exp. 1, and then completed one additional training block of the Exp. 2 task including intermixed Spatially Aligned and Misaligned trials. During this training, all participants indicated that they could clearly discriminate the identical tones presented from the left and right loudspeakers.

The actual experiment was divided into 15 blocks, with untimed breaks between blocks. 3 blocks contained only No Tone trials, while the other 12 contained intermixed Spatially Aligned and Misaligned trials. Participants completed a total of 108 No Tone trials, 216 AV Aligned trials (108 target-leading, 108 foil-leading), and 216 AV Misaligned trials (108 target-leading, 108 foil-leading). As in Exp. 1, one of each type of block was presented (order randomized independently for each subject) before any were repeated to mitigate any ordering effects.

4.1.3. EEG analysis

The addition of a second stream of tone complexes shortened the average interval between successive auditory stimuli to 550 ms. Thus, to limit contamination of ERPs by previous evoked responses, the baseline window was shortened to the 100 ms immediately preceding stimulus onset. Otherwise, the same preprocessing steps and statistical analyses described in Exp. 1 were used here. An additional analysis was performed to assess the topographic distribution of evoked potentials across the scalp, which reflects the combination of neural generators underlying the response. In order to separate ERP topography and amplitude differences between conditions, we divided individual subject average ERPs in each condition by the instantaneous global field power (GFP). This procedure normalizes the response amplitude while preserving the relative topographic distribution of the response (McCarthy and Wood, 1985). To quantify the degree of topographic difference between two responses, we then calculated a measure of global dissimilarity (DISS; see Cappe et al., 2010; Murray et al., 2008). At each time point, DISS is calculated as the root mean square of the differences in strength-normalized responses across the electrode montage. DISS was not calculated for the first 40 ms of post-stimulus time, as low evoked power in this early window rendered the metric unstable. Finally, permutation testing was used to determine time windows of statistically significant DISS values: over 2000 iterations, the labels of the conditions were randomly shuffled and DISS was recalculated, forming a null distribution. P-values at each time point were determined as the percentage of iterations on which the randomly permuted DISS value was greater than the actual DISS value. Only statistically significant regions at least 20 ms in duration were considered reliable.

5. Results

5.1. AV spatial alignment affects pip and pop effect strength when multiple AV stimuli are present

In the presence of a competing AV stimulus, search times (STs) were faster in the Spatially Aligned condition than the Misaligned condition (Fig. 7A). In addition, when the target color change and synchronous tone were presented before the foil (target-leading), search times (STs) were faster than when the foil was leading. The magnitude of this ST difference is similar to the average time difference between the first AV target latency in these conditions (529 ms). A repeated-measures two-way ANOVA supports these observations, revealing main effects of leading stream ($F(1,16) = 23.02$, $p < 0.001$) and AV spatial alignment ($F(1,16) = 13.54$, $p = 0.002$). There was no significant interaction between these factors ($F(1,16) = 2.98$, $p = 0.1$). Due to the difference in first target presentation time between the target-leading and foil-leading conditions, behavioral benefits of AV integration (reduced STs relative to the No Tone control) were only observed in the target-leading condition. Thus, the remaining analyses were limited to target-leading trials. On these trials, we observed that just over half of participants showed a substantial (>300 ms) ST benefit from Spatially Aligned tones. Within the target-leading trials, Tukey's HSD post-hoc testing revealed that STs were significantly faster in the Spatially Aligned condition than in the No Tone ($p = 0.006$) or Spatially Misaligned ($p = 0.015$) conditions. (Fig. 7B).

5.2. Separation between ERPs elicited by targets and foils depends on AV behavioral benefit

To examine neural correlates of selectively attending the AV target, we focused our EEG analysis on comparing ERPs elicited by the AV target and the AV foil. This analysis circumvented the need to sum unisensory ERPs, which can result in a doubling of electrical noise and

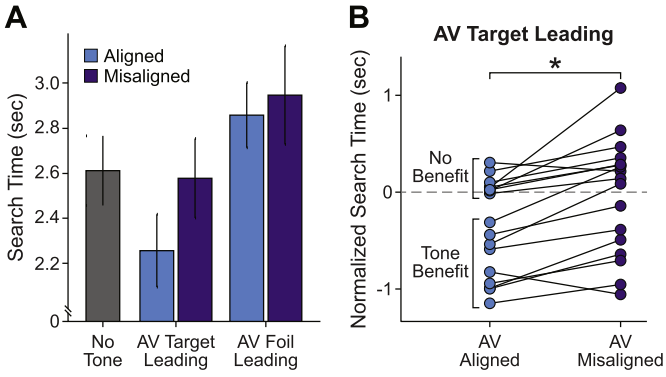

Fig. 7. Behavioral results from the pip and pop task with two AV stimuli (Exp. 2). **A.** Search times in all experimental conditions. Error bars represent S.E.M. **B.** Individual participant data from the target-leading condition only, normalized to search times in the No-Tone condition. Negative values indicate faster AV than visual-only search. Participants with search time benefits of at least 300 ms in the Aligned condition were termed Tone Benefit participants; their EEG data was analyzed separately from No Benefit participants.

common neural signals (e.g. potentials related to target detection and motor preparation) in the summed response (Gondan and Röder, 2006; Teder-Sälejärvi et al., 2002). Fig. 8A shows the scalp topographies of grand average target and foil ERPs, as well as the difference between them, in two time windows and both AV spatial alignment conditions. At the group level, cluster-based permutation testing revealed significant differences between responses to the AV target and AV foil in both spatial alignment conditions. In a time range centered on the auditory N1 (95-135 ms), responses to the target were larger than responses to the foil in both the Spatially Aligned ($p = 0.018$) and Spatially Misaligned ($p = 0.023$) conditions (Fig. 8A, left panels). In the auditory P2 time range (170-210 ms), responses to the target and foil had markedly different scalp topographies, resulting in significant clusters in both spatial alignment conditions ($p = 0.015$ for Spatially Aligned, $p = 0.011$ for Misaligned; right panels of Fig. 8A). AV foils elicited a fronto-central positivity similar to the P2 responses observed in Exp. 1, whereas AV targets elicited a bilateral negativity over occipital channels on average. A likely explanation for this difference is that, while participants had to find the visual target to perform the task, the visual component of AV foils generally went undetected. This later time range also corresponds to the visual N1 component, which typically peaks 150-200 ms post-stimulus, and is stronger contralateral to the hemifield containing the visual stimulus (Makeig et al., 1999; Mangun and Hillyard, 1991). Separately plotting trials in which the AV target was on the left and right reveals that the responses were indeed lateralized contralaterally to the visual target hemifield (Supplemental Fig. 4). Therefore, ERP effects in this time range can be explained by differences in detection of the visual stimuli. However, differences between ERPs elicited by targets and foils in the auditory N1 time range indicate that – on average – participants were able to selectively attend the AV target whether its unisensory components were spatially aligned or misaligned.

The difference in the strength of N1 responses elicited by AV targets and foils was weakly correlated with individual participants' AV search benefit ($R^2 = 0.115$, $p = 0.05$; Fig. 8B). Here, N1 response strength in each condition was calculated as the average GFP in the N1 time range so that particular channels did not have to be selected a priori. Given this modest correlation, we re-analyzed the ERP data separately for participants who benefitted from the AV synchrony (Tone Benefit participants) and those who did not (No Benefit participants). These groups were delineated based on whether search times were at least 300 ms faster in the AV Spatially Aligned condition than the No Tone condition. We chose this cutoff empirically, as there was a clear divide between participants who benefitted and those who did not (see Fig. 7B); it also approximates a median split of the data, with 9 Tone Benefit and 8 No Benefit participants.

For the Tone Benefit participants, there was a clear separation between AV target and foil responses across most of the electrode montage, in both spatial alignment conditions (Fig. 9A). This separation indicates that these participants were able to selectively attend the AV target, even if its unisensory components were misaligned in space. In keeping with this, the Tone Benefit participants generally experienced some behavioral benefit from the Spatially Misaligned synchronous tones (although less so than in the Spatially Aligned condition). The topography of the Spatially Aligned target responses was shifted in a left-posterior direction. Consistent with this, the left-posterior electrode channels showed the greatest difference between target and foil responses in this condition, whereas the maximum difference occurred at more central channels in the Spatially Misaligned condition.

In the Tone Benefit group, Cluster-based permutation testing showed significantly stronger N1 responses to AV targets than foils in both the Spatially Aligned ($p = 0.04$) and Misaligned ($p = 0.032$) conditions (Fig. 9B, left panels). Interestingly, a broad region of significant difference between the target and foil responses in the P2 time range was observed in the Spatially Aligned ($p = 0.034$) condition, but not the Misaligned condition. This effect appears to have been driven by the combination of a stronger response to the Spatially Aligned target in sensors over visual brain areas, and a stronger attentionally driven P2 response to the Aligned foil. Since AV integrative responses are known to occur in visual cortex (Allman and Meredith, 2007; Morgan et al., 2008; Murray et al., 2016), this result suggests a combination of more robust

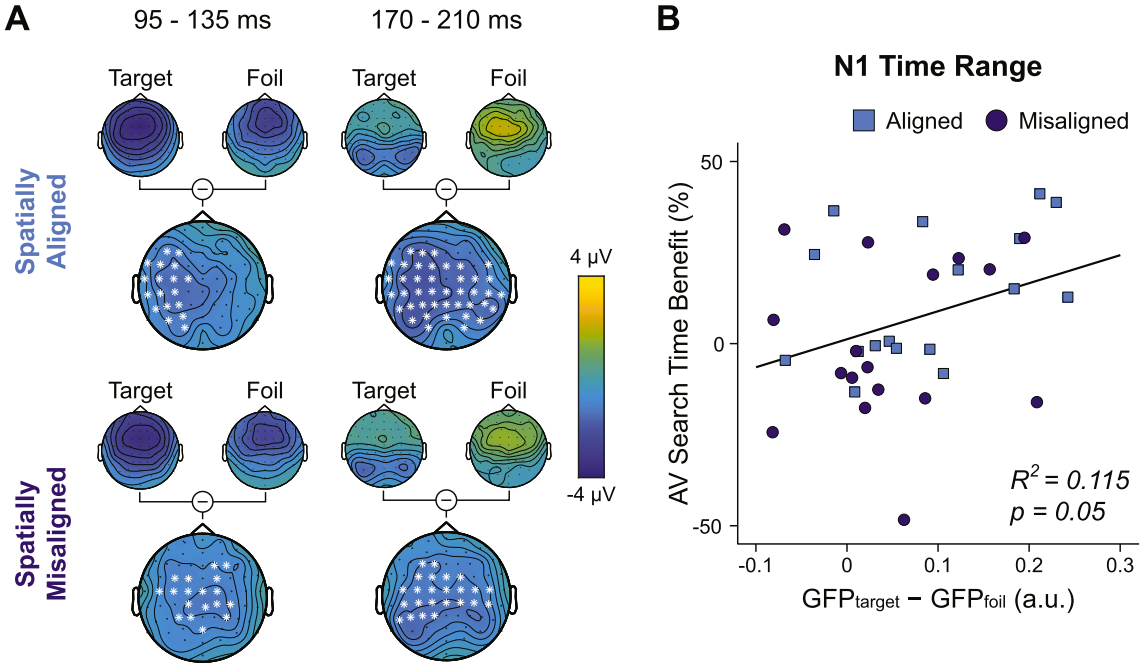

Fig. 8. Group-level ERP results from Experiment 2. **A**. Scalp topographies of ERPs elicited by AV targets and foils are shown in time ranges corresponding to the auditory N1 (95-135 ms, left column) and the auditory P2/visual N1 (170-210 ms, right column). Spatially Aligned responses are shown in the top row, and Misaligned responses are in the bottom row. Difference topographies (AV target – AV foil) are shown beneath each pair of target and foil topoplots. Asterisks indicate significance in cluster-based permutation testing. **B**. Correlation between the global field power difference between target and foil responses in the auditory N1 time range and AV behavioral benefit. Positive percentage values on the y-axis indicate faster search relative to the No Tone condition.

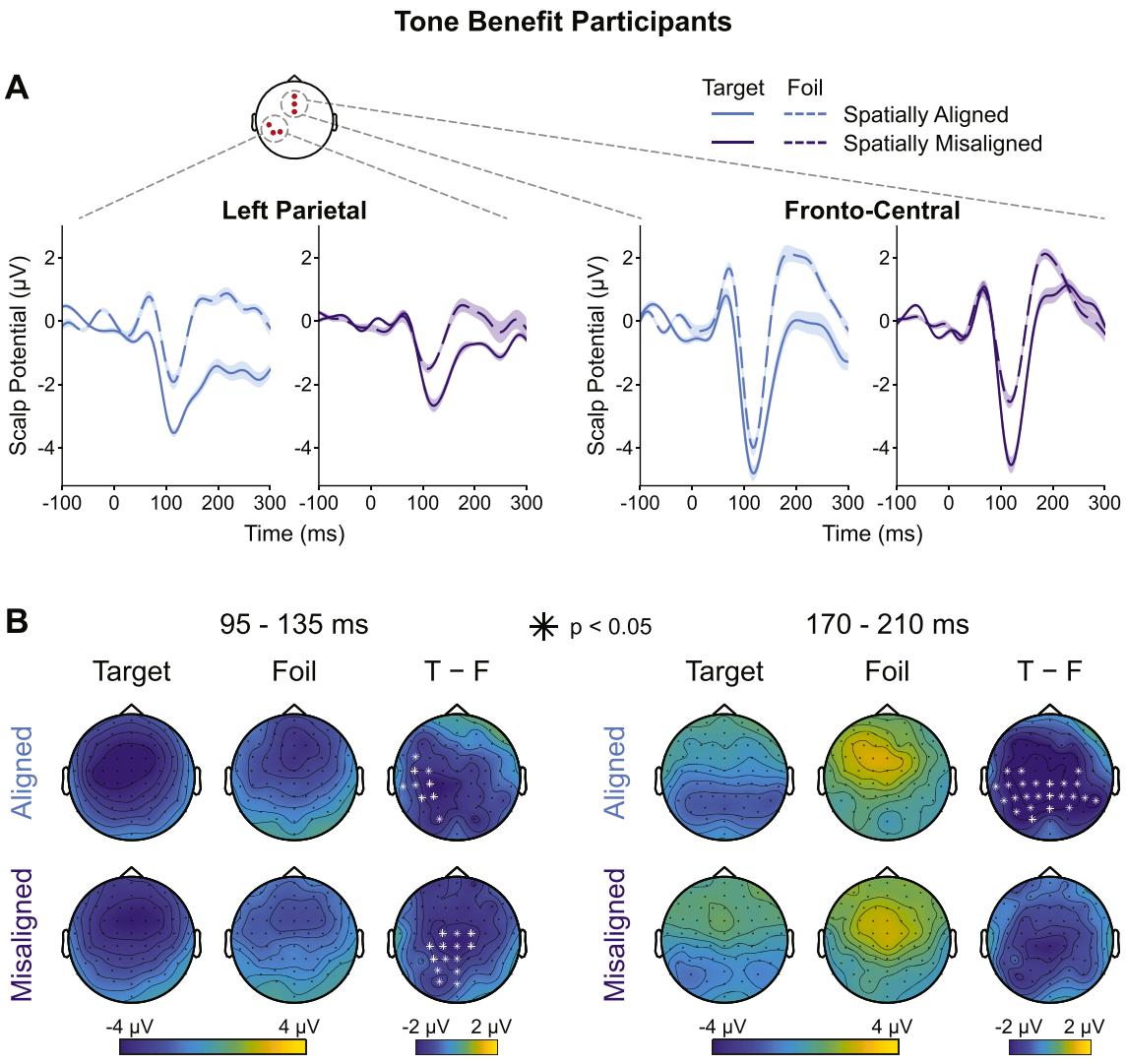

Fig. 9. Comparison of ERPs elicited by AV targets and foils for Tone Benefit participants in Experiment 2. **A.** ERP waveforms, each averaged over a group of three electrodes. The left parietal group includes electrodes P3, P5, and CP5, selected to capture the early, left-biased cluster in the Spatially Aligned Target versus Foil comparison. The fronto-central group includes electrodes Fz, FCz, and Cz, where auditory responses were strongest. Error clouds indicate S.E.M. **B.** Response topographies in the auditory N1 and P2 time ranges. Note the different color scales used to visualize effects. Asterisks indicate significant differences between target and foil responses in cluster-based permutation testing. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

encoding of the Spatially Aligned AV targets and an increased suppression of responses to Spatially Aligned foils.

We lacked sufficient statistical power to directly compare the Tone Benefit and No Benefit groups in terms of the separation between their target and foil evoked responses. This is partly because of the low participant counts after dividing participants into groups, and partly because of the noise introduced by subtracting the target and foil ERPs. For the No Benefit participants, although the response topographies were qualitatively similar to those in the Tone Benefit group, differences between target and foil responses were not statistically significant in either the N1 or P2 time windows (Fig. 10). These neural measures suggest that the No Benefit participants struggled to perceptually segregate the AV target and foil in both spatial alignment conditions, consistent with the lack of search facilitation in this group.

The No Benefit group had one fewer participant than the Tone Benefit group. To address the possibility that the null results for the No Benefit group were due to reduced statistical power, we re-ran the cluster tests on the Tone Benefit data 9 times, each time removing a different participant from the dataset. The Tone Benefit results described above held on all runs; N1s elicited by targets were larger than those elicited by foils in the Aligned and Misaligned conditions, and the P2 elicited by the foil was larger than that elicited by the target in only the Aligned condition. Cluster p-values were less than (or in one case equal to) 0.05 for each of these comparisons on each run. Thus, the slight difference in sample sizes does not account for response differences between the Tone Benefit and No Benefit participants.

5.3. The early ERP topography is shifted when AV stimuli are spatially aligned

In the Tone Benefit group, the differential evoked responses in the 170–210 ms time range represent one possible neural correlate of the enhanced behavioral performance in the Spatially Aligned condition. In this group, the topography of AV target responses in the earlier N1 window also appeared to differ between the Aligned and Misaligned conditions, suggesting differential activation of brain areas engaged in processing the AV stimuli. To test the significance of this topographic shift separately from ERP amplitude differences, we strength-normalized target and foil-evoked ERPs and then compared their scalp topographies using the DISS metric (Fig. 11; see EEG analysis in Exp. 2

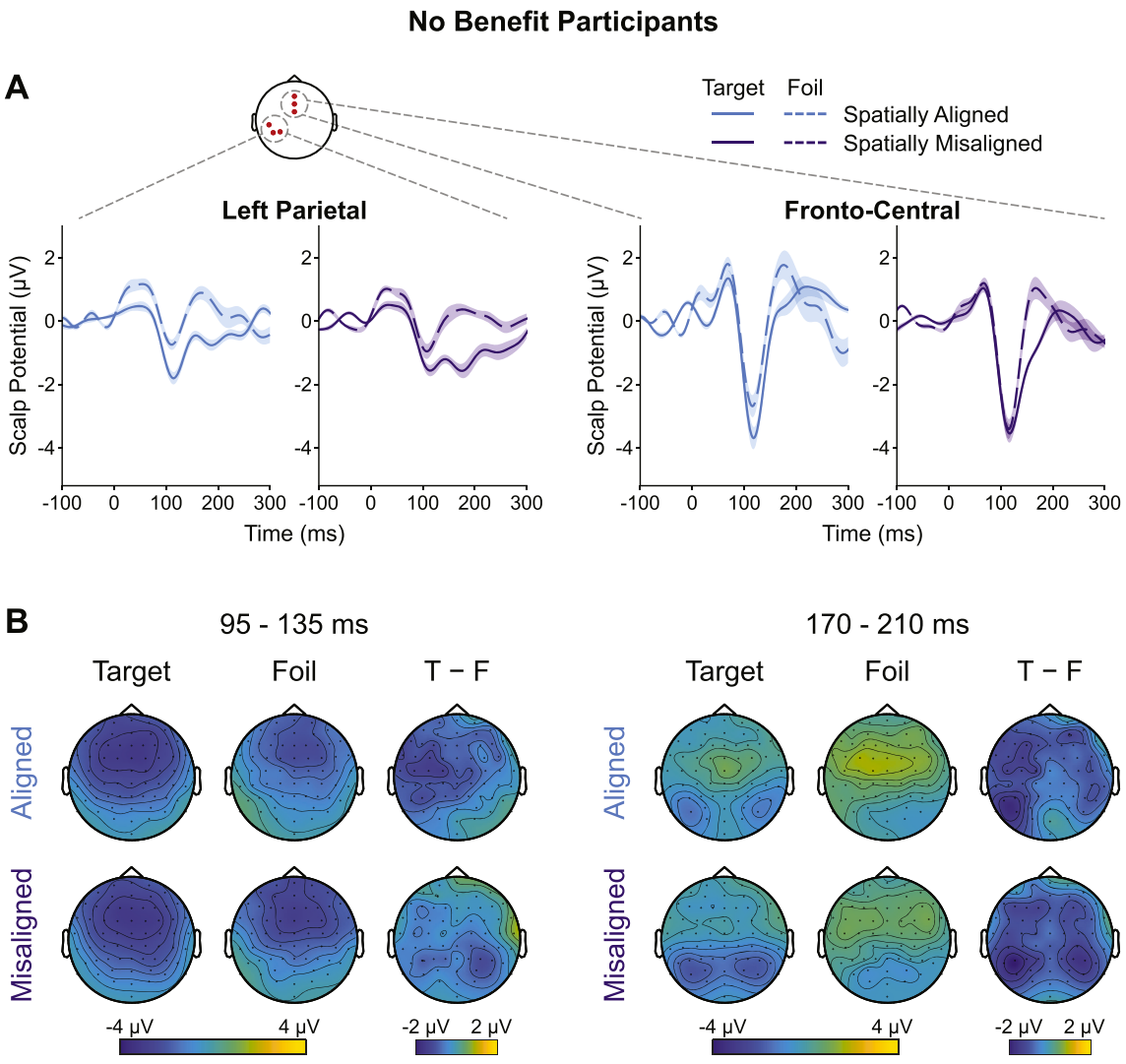

Fig. 10. Comparison of ERPs elicited by AV targets and foils for No Benefit participants in Experiment 2. **A**. ERP waveforms, each averaged over a group of three electrodes: P3, P5, and CP5 on the left, and Fz, FCz, and Cz on the right. These match the electrode groups plotted in Fig. 9. Error clouds indicate S.E.M. **B**. Response topographies in the N1 and P2 time ranges. No significant differences were found between target and foil responses in any condition or time window.

Methods).

For the Tone Benefit participants, DISS indicated that the topographic difference between target and foil responses was elevated throughout the post-stimulus epoch in the Spatially Aligned condition (Fig. 11A). Spatially Aligned target and foil response topographies were significantly different between 62 and 93 ms post-stimulus, as well as between 127 and 179 and 189 and 234 ms post-stimulus. In contrast to these broad differences, in the Spatially Misaligned condition, target and foil topographies were significantly different only between 146 and 168 ms post-stimulus (Fig. 11B). This time window corresponds to the onset of clear visual responses evoked by the target stimulus, but not the foil in the unattended hemifield. Thus, topographic differences were expected in this time window regardless of AV spatial alignment. Of particular interest, however, is the early topographic difference in the Spatially Aligned condition, which was not present in the Misaligned condition. Multiple studies have reported early multisensory effects on evoked responses around 40-60 ms post-stimulus (Ghazanfar et al., 2005; Molholm et al., 2002; Murray et al., 2016; Schwartz et al., 2004). We also conducted a separate DISS analysis to directly compare the topographies of ERPs elicited by AV targets between the Aligned and Misaligned conditions. This analysis revealed that the Aligned and Misaligned target response topographies differed significantly from each other in a similar early time window, between 50 and 78 ms post-stimulus ($p < 0.05$ for all time points). Although these DISS analyses do not specifically implicate multisensory brain regions, the early topographic shifts are consistent with AV integration of the target being stronger when its unisensory components were spatially aligned.

For the No Benefit participants, the early topographic difference between target and foil responses was absent, consistent with weaker AV integration (Fig. 11, panels C and D). In the Spatially Misaligned condition, DISS reached significance in a later time window (134–193 ms) corresponding to visual responses elicited by the AV target. Though DISS did not reach significance at any point in the Spatially Aligned condition, a peak is visible around 165 ms post-stimulus, as expected. A direct DISS comparison between Aligned and Misaligned target ERP topographies found no differences in the No Benefit group.

Finally, we revisited the Exp. 1 data to test for similar topographic differences between AV and summed unisensory (A + V) responses. Indeed, DISS revealed significant topographic differences between both the AV Aligned and Misaligned conditions and the summed unisensory models (Supplemental Fig. 5). This reflects the patterns observed in the behavioral data; when a single AV stimulus was present, AV integration was unaffected by AV spatial misalignment (Exp. 1), whereas when multiple AV stimuli were present, AV integration was aided by spatial

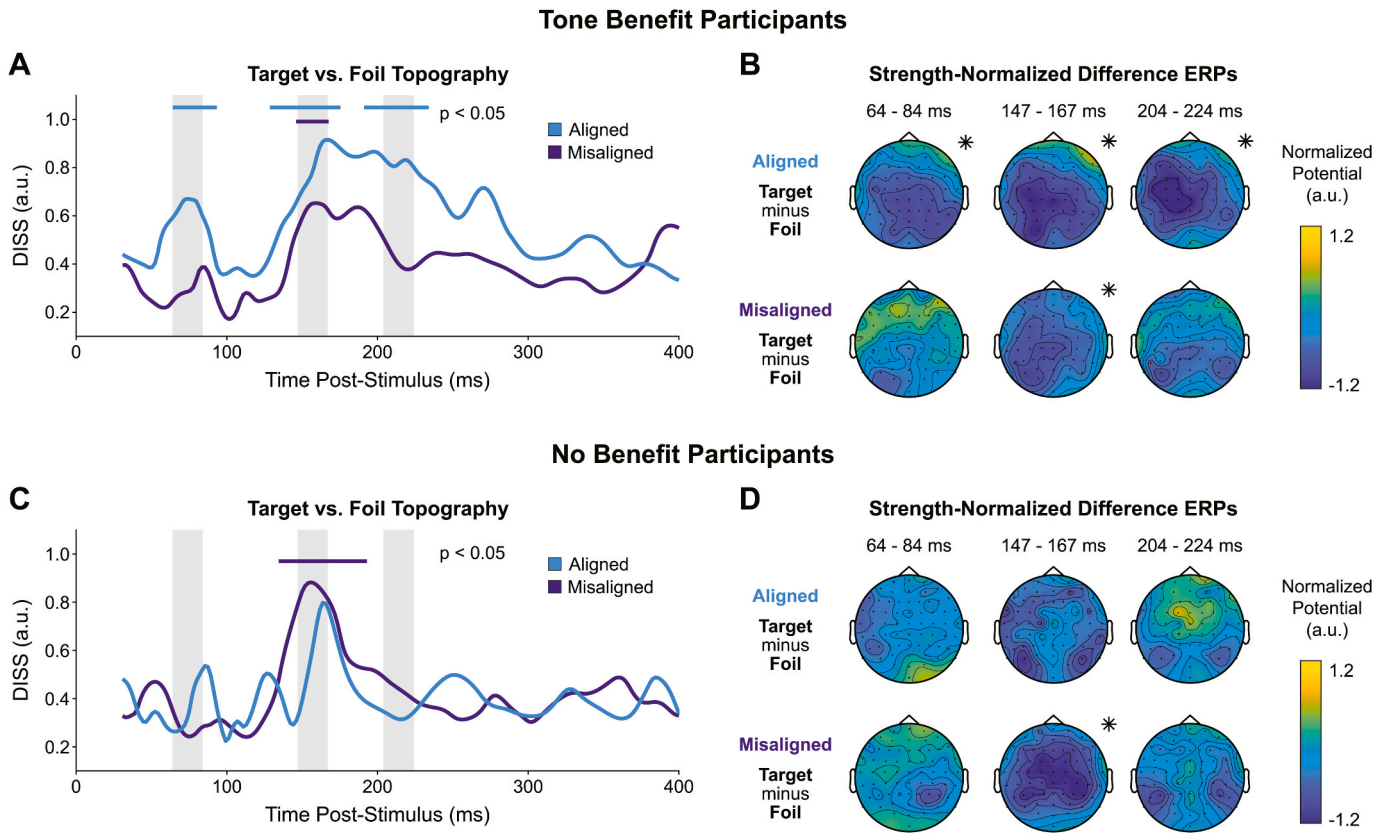

Fig. 11. Analysis of AV topographic shift in Exp. 2. DISS time courses reflect the degree of topographic dissimilarity between the AV target and foil responses in the AV Aligned (light blue) and Misaligned (purple) conditions. DISS was calculated separately for Tone Benefit (**A**) and No Benefit (**C**) participants. Horizontal bars above the time courses represent time regions in which DISS reached statistical significance in permutation testing. Grey time regions correspond to the topoplots in B and D. **B,D**. Grand average difference topographies between AV target and foil ERPs, normalized by the global field power to visualize regions of differential topography. Asterisks indicate DISS significance in the corresponding condition and time region.

alignment between corresponding auditory and visual inputs (Exp. 2).

6. Discussion

In two experiments, we measured the behavioral and neurophysiological effects of audio-visual integration in search tasks using variants on the pip and pop paradigm. Of particular interest to us was elucidating whether spatial alignment between the auditory and visual stimuli influenced the strength of the pip and pop effect, and whether the impact of AV spatial alignment depended on the presence of multisensory competition in the scene.

In Exp. 1, we first replicated the pip and pop effect (Van der Burg et al., 2008; Zou et al., 2012). We found that the addition of an auditory stimulus synchronized to changes in a visual target feature (color) suppressed errors and sped visual search. Importantly, these behavioral benefits were observed regardless of whether the auditory stimuli were Spatially Aligned or Spatially Misaligned with the visual search display. In accord with this, we observed an enhancement of the ERP N1 component in AV conditions, irrespective of AV spatial alignment. We demonstrated that the final auditory stimulus before the participant's response on each trial, which likely was integrated with the visual target, tended to evoke the largest N1 response. This finding is consistent with AV synchrony increasing the salience of the AV target.

In Exp. 2 we added an AV foil, so that the task required selective integration of the visual target with its synchronous tones, while ignoring competing stimuli in both the auditory and visual modalities. In contrast to Exp. 1, under these conditions search was faster when the auditory and visual stimuli were Spatially Aligned than when they were Misaligned. Regardless of spatial alignment, AV targets elicited significantly larger N1 ERP components than physically identical AV foils for participants who benefitted from the synchronous tones; for participants who did not benefit behaviorally, this difference was not significant. After normalizing the ERPs to account for differences in response strength, we found that the ERP scalp topography differed between target and foil responses only for participants for whom AV search was faster than visual-only search, and only in the Spatially Aligned condition. For these participants, the Aligned and Misaligned target response topographies also differed from each other in an early post-stimulus time window. This indicates that the neural generators of the ERPs differed in these two conditions, possibly reflecting cortical multisensory processing mechanisms being driven more strongly in the Spatially Aligned condition.

Taken together, our results indicate that in simple scenes, temporal synchrony alone is sufficient to drive AV integration. However, in more complex and ambiguous environments with competing sources, spatial relationships between the auditory and visual stimuli can aid in parsing the sensory scene.

6.1. The influence of spatial factors in AV integration depends on stimulus context

A preponderance of evidence points to the importance of temporal coherence between auditory and visual signals in driving AV integration and binding. Van der Burg and colleagues demonstrated this in the context of the pip and pop effect by introducing temporal offsets between the tones and visual targets, which reduced search benefits in a manner consistent with estimates of the AV temporal binding window (Van der Burg et al., 2008). On the other hand, the functional role of AV

spatial alignment, in both the pip and pop effect and AV integration in general, remains a topic of substantial debate.

In most previous investigations of the pip and pop effect, spatial relationships between the visual stimuli and the tones have been ambiguous as tones were presented over headphones (Van der Burg, Awh and Olivers, 2013; Van der Burg, Olivers and Cass, 2017; Van der Burg et al., 2010). One study did find stronger behavioral benefits when synchronous tones were presented from a free-field loudspeaker placed near the visual display compared to when the tones were presented over headphones (Ngo and Spence, 2010). However, it is not clear if this difference was due to AV spatial alignment per se, as opposed to the externalization of stimuli in both modalities into physical space.

In the experiments presented here, we manipulated AV spatial alignment of physical sound sources, and found effects of spatial alignment only in Exp. 2, when both visual and auditory stimuli were in competition for integration resources. One explanation for these results is that spatial information about the stimuli was task-relevant in Exp. 2, as visual search was restricted to a cued hemifield and aided by synchronous tones in a separately cued hemifield. This has some experimental precedent; in one study, spatial misalignment between visual and somatosensory signals was shown to affect integration only when participants had to restrict their responses to stimuli presented in one hemifield (Girard et al., 2011). However, others have failed to find spatial effects on multisensory integration even when spatial information was directly task-relevant (e.g. Sperdin et al., 2010).

The addition of a competing stream of tones in Exp. 2 altered the requirements of the pip and pop task; participants had to find an AV target amongst a mixture of unisensory *and* multisensory distractors – a common situation in everyday perception. One intriguing possibility is that the spatial effects we observed in Exp. 2 were due to the addition of this cross-modal competition. A number of past studies support this idea. Auditory spatial cues are often secondary to other sources of auditory information (e.g. pitch, talker identity, etc.) in forming auditory objects, and only become influential when other cues are ambiguous or the listening environment is particularly complex (Bizley et al., 2012; Mehraei et al., 2018). Given that the relatively poor spatial acuity of audition limits the precision of AV spatial estimates, it stands to reason that spatial cues would also play a secondary role in AV integration, helping to resolve scenes only when other cues to integration are ambiguous or the multisensory environment is cluttered.

It should also be noted that the relative sparsity of possible spatial locations in our experimental setup and others could lead to a down-weighting of spatial cues to AV integration. The brain integrates sensory cues in a near-optimal fashion, giving most credence to highly reliable cues (Ernst and Banks, 2002; Fetsch et al., 2012; Hillis et al., 2004; Jacobs and Fine, 1999). In ventriloquism, for example, spatial perception is normally biased toward vision, but the balance can be shifted toward audition by making the visual stimulus spatially unreliable (e.g., a blurred light instead of a small dot; Alais and Burr, 2004). Our setup comprised a single visual display and two (Exp. 2) or three (Exp. 1) possible auditory stimulus locations. The very simplistic auditory spatial arrangements we employed did little to drive participants to make fine use of auditory spatial cues; multisensory integration may have therefore favored precise AV temporal coherence over coarse auditory spatial information. Future studies investigating the role of AV spatial alignment using more immersive multisensory experimental environments will be required to fully resolve the role that spatial information plays in real-world AV processing.

6.2. *The pip and pop effect reveals interactions between top-down and bottom-up attention and AV integration*

When comparing AV neural responses to summed unisensory responses in Exp. 1, we observed an enhancement of AV ERP components, particularly the in the time range of the auditory N1. The strength of this component was also strongly modulated in Exp. 2, when AV target and AV foil responses were compared. Such "super-additive" responses have been considered a hallmark of AV integration in the cat superior colliculus, but mounting evidence indicates that super-additivity does not constitute a general property of multisensory processing (Angelaki et al., 2009; Meredith and Stein, 1986; Stein and Meredith, 1993). More recent studies in humans and animals have demonstrated that super-additive multisensory responses are restricted to near-threshold stimulus levels, and that at supra-threshold levels, multisensory responses are more commonly sub-additive (Perrault et al., 2003; Stanford, Quessy and Stein, 2005; Stanford and Stein, 2007; Van der Burg, Talsma, Olivers, Hickey and Theeuwes, 2011). In addition, modeling and empirical studies have shown that sub-additive multisensory interactions can account for near-optimal cue integration across sensory modalities (Anastasio et al., 2000; Gu et al., 2008; Ma et al., 2006).

Given that super-additive multisensory responses are rare, a more probable explanation is that the N1 modulations we observed were influenced by attentional differences between conditions. Attentionally driven modulation of the auditory N1 has been extensively reported in the literature, lending weight to this interpretation (Alho, 1992; Choi et al., 2014; Hillyard et al., 1973; Woldorff et al., 1993). Regarding top-down attention, AV targets were attended in all conditions, whereas asynchronous auditory-only stimuli (Exp. 1) and AV foils (Exp. 2) were ignored. While these differences did modulate the ERPs in both experiments, in Exp. 1 we also found a selective enhancement of the last AV target ERP before target detection relative to earlier auditory ERPs. This is inconsistent with the enhancement being fully explained by top-down attention, as participants presumably sustained attention to synchronous tones throughout the trials, even before the visual target was detected. Thus, AV integration in the pip and pop effect likely increased the allocation of bottom-up attention to the multisensory targets.

A growing body of research highlights the fact that multisensory integration and attentional mechanisms are intimately intertwined, but substantial debate surrounds the particular roles of bottom-up and top-down attention in integration (De Meo, Murray, Clarke and Matusz, 2015; Hartcher-O'Brien et al., 2017; Macaluso et al., 2016; Talsma et al., 2010). Previous EEG studies focusing on ERPs (Talsma et al., 2007) and gamma responses (Senkowski et al., 2005) have demonstrated that only AV stimuli subject to top-down attention elicit enhanced multisensory responses. In non-human primates, it has been shown that attended visual stimuli can reset the phase of ongoing oscillatory activity in primary auditory cortex, and vice versa (Lakatos et al., 2009). In the pip and pop effect, the size of participants' endogenous spatial attention window modulates the ability of salient AV events to capture attention. Specifically, AV integration appears to be weakened when stimuli are presented far from a focused spatial window of top-down attention (Van der Burg, et al., 2012).

On the other hand, numerous accounts exist of multisensory integration enhancing the salience of sensory events, leading to capture of exogenous attention (see Tang et al., 2016 for review). Such a salience account fits well with AV reaction time studies, which demonstrate that AV RTs are faster than would be predicted from probability summation of unisensory RTs, and that RTs are slowed when AV integration is disrupted by introducing spatial or temporal offsets between the stimuli (Frens et al., 1995). Salient stimuli in general also produce neural responses that are in accord with the AV N1 enhancements observed in the present study. For instance, EEG studies have shown increased N1 amplitudes to salient somatosensory stimuli (Iannetti et al., 2008) and visual stimuli that are distinct compared to distractors in visual search (Töllner et al., 2011). A parsimonious explanation of the search benefits observed in the pip and pop effect is that AV integration causes a bottom-up increase in salience, resulting in the attentional capture of synchronous AV stimuli.

6.3. Neural underpinnings of multisensory integration in the pip and pop effect

A great deal of previous multisensory research has focused on detection tasks using single, isolated stimuli. These studies have elucidated foundational principles of AV integration in humans, but an important open question concerns whether the same neural processes operate under more challenging sensory circumstances. To this end, electrophysiological studies using more complex paradigms with sensory competition, including the pip and pop effect, can be enlightening.

To date, few studies have investigated the neural underpinnings of the pip and pop effect. One diffusion tensor imaging study found that the strength of white matter connections between subcortical auditory structures and auditory cortex predicted the magnitude of the pip and pop effect in individual participants, hinting at subcortical contributions to the effect (Van den Brink et al., 2014). A particularly relevant study involved a modified version of the pip and pop paradigm in which tones could be synchronized to target or distractor orientation changes (Van der Burg et al., 2011). As with the present work, that study found that AV stimuli evoked significantly different responses than their summed unisensory components in an early time window (50-70 ms post-stimulus), with the largest multisensory differences clustered at left-posterior electrode sites. This could signal a topographic shift of the AV ERPs, similar to those we identified using the DISS metric (McCarthy and Wood, 1985; Murray et al., 2008). Explicit analyses of shifts in the topography of ERPs elicited by AV stimuli have been reported previously. In one such study, regions of significant AV topographic change were source-localized to primary auditory and visual cortices, as well as the posterior superior temporal sulcus, known for multisensory response properties (Cappe et al., 2010). The use of a different task and the presence of stronger visually evoked responses in that study confound direct comparison to the present experiments, and our lack of sufficient electrode density and registration of individual participants' electrode positions precluded source localization of the data presented here. Still, it is plausible that the topographic shifts observed were driven by similar multisensory neural processes.

7. Conclusions

Using variants of the pip and pop search paradigm, we found differential effects of AV spatial alignment on AV integration depending on the presence of a competing AV stimulus. Taken together, our results suggest a context-dependence of the role of spatial cues in AV integration. In a simple scene with only a single AV object, AV temporal coherence alone can drive integration, but when selective attention is required to suppress irrelevant AV stimuli, AV spatial relationships play a more decisive role in selective integration of the correct inputs across the senses.

CRediT authorship contribution statement

Justin T. Fleming: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Visualization, Writing - original draft. **Abigail L. Noyce:** Conceptualization, Funding acquisition, Methodology, Supervision, Validation, Writing - review & editing. **Barbara G. Shinn-Cunningham:** Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Validation, Writing - review & editing.

Acknowledgements

The authors would like to thank Alex Hammerman and Sucheta Tamragouri for assistance with data collection and preprocessing, and to Erik van der Burg for his helpful correspondence about the pip and pop effect. The authors declare no competing interests. This work was supported by the National Institute on Deafness and Other Communication Disorders (grant numbers R01-DC013825 and T32-DC000038), the National Eye Institute (grant number F32-EY026796) and the Office of Naval Research (grant number N000141812069).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuropsychologia.2020.107530.

References

Alais, D., Burr, D., 2004. The ventriloquist effect results from near-optimal bimodal integration. Curr. Biol. 14 (3), 257–262. https://doi.org/10.1016/S0960-9822(04)00043-0.

Alho, K., 1992. Selective attention in auditory processing as reflected by event-related brain potentials. Psychophysiology 29 (3), 247–263.

Allman, B.L., Meredith, M.A., 2007. Multisensory processing in "unimodal" neurons: cross-modal subthreshold auditory effects in cat extrastriate visual cortex. J. Neurophysiol. 98 (1), 545–549. https://doi.org/10.1152/jn.00173.2007.

Anastasio, T.J., Patton, P.E., Belkacem-Boussaid, K., 2000. Using Bayes' rule to model multisensory enhancement in the superior colliculus. Neural Comput. 12 (5), 1165–1187. https://doi.org/10.1162/089976600300015547.

Angelaki, D.E., Humphreys, G., DeAngelis, G.C., 2009. Multisensory integration: psychophysics, neurophysiology, and computation. Curr. Opin. Neurobiol. 19 (4), 452–458. https://doi.org/10.1016/j.conb.2009.06.008.Multisensory.

Atilgan, H., Town, S.M., Wood, K.C., Jones, G.P., Maddox, R.K., Lee, A.K.C., Bizley, J.K., 2018. Integration of visual information in auditory cortex promotes auditory scene analysis through multisensory binding. Neuron 97 (3), 640–655. https://doi.org/10.1016/j.neuron.2017.12.034.

Avillac, M., Ben Hamed, S., Duhamel, J.-R., 2007. Multisensory integration in the ventral intraparietal area of the macaque monkey. J. Neurosci. 27 (8), 1922–1932. https://doi.org/10.1523/JNEUROSCI.2646-06.2007.

Bertelson, P., Vroomen, J., Wiegeraad, G., de Gelder, B., 1994. Exploring the relation between McGurk interference and ventriloquism. Proc. ICSLP 559–562. Retrieved from. https://www.isca-speech.org/archive/icslp_1994/i94_0559.html.

Bizley, J.K., Shinn-Cunningham, B.G., Lee, A.K.C., 2012. Nothing is irrelevant in a noisy world: sensory illusions reveal obligatory within-and across-modality integration. J. Neurosci. 32 (39), 13402–13410. https://doi.org/10.1523/JNEUROSCI.2495-12.2012.

Bosen, A.K., Fleming, J.T., Allen, P.D., O'Neill, W.E., Paige, G.D., 2018. Multiple time scales of the ventriloquism aftereffect. PloS One 13 (8), 1–20. https://doi.org/10.1371/journal.pone.0200930.

Brainard, D.H., 1997. The psychophysics toolbox. Spatial Vis. 10 (4), 433–436. https://doi.org/10.1163/156856897X00357.

Cappe, C., Thut, G., Romei, V., Murray, M.M., 2010. Auditory-visual multisensory interactions in humans: timing, topography, directionality, and sources. J. Neurosci. 30 (38), 12572–12580. https://doi.org/10.1523/jneurosci.1099-10.2010.

Choi, I., Wang, L., Bharadwaj, H., Shinn-Cunningham, B., 2014. Individual differences in attentional modulation of cortical responses correlate with selective attention performance. Hear. Res. 314, 10–19. https://doi.org/10.1016/j.heares.2014.04.008.

Conroy, M.A., Polich, J., 2007. Normative variation of P3a and P3b from a large sample: gender, topography, and response time. J. Psychophysiol. 21 (1), 22–32. https://doi.org/10.1027/0269-8803.21.1.22.

Cui, Q.N., Razavi, B., O'Neill, W.E., Paige, G.D., 2010. Perception of auditory, visual and egocentric spatial alignment adapts differently to changes in eye position. J. Neurophysiol. 103 (2), 1020–1035. https://doi.org/10.1152/jn.00500.2009.

Dai, L., Best, V., Shinn-Cunningham, B.G., 2018. Sensorineural hearing loss degrades behavioral and physiological measures of human spatial selective auditory attention. Proc. Natl. Acad. Sci. Unit. States Am. 115 (14), E3286–E3295. https://doi.org/10.1073/pnas.1721226115.

De Meo, R., Murray, M.M., Clarke, S., Matusz, P.J., 2015. Top-down control and early multisensory processes: chicken vs. egg. Front. Integr. Neurosci. 9 https://doi.org/10.3389/fnint.2015.00017.

Diederich, A., 1995. Intersensory facilitation of reaction time: evaluation of counter and diffusion coactivation models. J. Math. Psychol. 39, 197–215. https://doi.org/10.1006/jmps.1995.1020.

Ernst, M.O., Banks, M.S., 2002. Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415 (6870), 429–433. https://doi.org/10.1038/415429a.

Fetsch, C.R., Pouget, A., Deangelis, G.C., Angelaki, D.E., 2012. Neural correlates of reliability-based cue weighting during multisensory integration. Nat. Neurosci. 15 (1), 146–154. https://doi.org/10.1038/nn.2983.

Frens, M. a, Opstal, a J. Van, Willigen, R. F. Van Der, 1995. Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. Percept. Psychophys. 57 (6), 802–816. https://doi.org/10.3758/BF03206796.

García-Larrea, L., Lukaszewicz, A.C., Mauguiére, F., 1992. Revisiting the oddball paradigm. Non-target vs neutral stimuli and the evaluation of ERP attentional effects. Neuropsychologia 30 (8), 723–741. https://doi.org/10.1016/0028-3932(92)90042-K.

Ghazanfar, A.A., Joost, X.M., Hoffman, K.L., Logothetis, N.K., 2005. Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. J. Neurosci. 25 (20), 5004–5012. https://doi.org/10.1523/jneurosci.0799-05.2005.

Girard, S., Collignon, O., Lepore, F., 2011. Multisensory gain within and across hemispaces in simple and choice reaction time paradigms. Exp. Brain Res. 214 (1), 1–8. https://doi.org/10.1007/s00221-010-2515-9.
Gondan, M., Röder, B., 2006. A new method for detecting interactions between the senses in event-related potentials. Brain Res. 1073–1074 (1), 389–397. https://doi.org/10.1016/j.brainres.2005.12.050.
Gu, Y., Angelaki, D.E., DeAngelis, G.C., 2008. Neural correlates of multisensory cue integration in macaque MSTd. Nat. Neurosci. 11 (10), 1201–1210. https://doi.org/10.1038/nn.2191.
Hartcher-O'Brien, J., Soto-Faraco, S., Adam, R., 2017. Editorial: a matter of bottom-up or top-down processes: the role of attention in multisensory integration. Front. Integr. Neurosci. 11 https://doi.org/10.3389/fnint.2017.00005.
Hillis, J.M., Watt, S.J., Landy, M.S., Banks, M.S., 2004. Slant from texture and disparity cues: optimal cue combination. J. Vis. 12 (4) https://doi.org/10.1167/4.12.1.
Hillyard, S.A., Hink, R.F., Schwent, V.L., Picton, T.W., 1973. Electrical signs of selective attention in the human brain. Science 182 (4108), 177–180. https://doi.org/10.1126/science.182.4108.177.
Howard, P., Templeton, W.B., 1966. Human spatial orientation. I. P. Howard, and W. B. Templeton. J. Royal Aeronaut. Soc. 70 (670), 960–961. https://doi.org/10.1017/S0368393100082778.
Iannetti, G.D., Hughes, N.P., Lee, M.C., Mouraux, A., 2008. Determinants of laser-evoked EEG responses: pain perception or stimulus saliency? J. Neurophysiol. 100 (2), 815–828. https://doi.org/10.1152/jn.00097.2008.
Innes-Brown, H., Crewther, D., 2009. The impact of spatial incongruence on an auditory-visual illusion. PloS One 4 (7). https://doi.org/10.1371/journal.pone.0006450.
Jacobs, R.A., Fine, I., 1999. Experience-dependent integration of texture and motion cues to depth. Vis. Res. 39 (24), 4062–4075. https://doi.org/10.1016/S0042-6989(99)00120-0.
Keil, J., Senkowski, D., 2018. Neural oscillations orchestrate multisensory processing. Neuroscientist 24 (6), 609–626. https://doi.org/10.1177/1073858418755352.
Körding, K.P., Beierholm, U., Ma, W.J., Quartz, S., Tenenbaum, J.B., Shams, L., 2007. Causal inference in multisensory perception. PloS One 2 (9). https://doi.org/10.1371/journal.pone.0000943.
Lakatos, P., O'Connell, M.N., Barczak, A., Mills, A., Javitt, D.C., Schroeder, C.E., 2009. The leading sense: supramodal control of neurophysiological context by attention. Neuron 64 (3), 419–430. https://doi.org/10.1016/j.neuron.2009.10.014.
Leblanc, É., Prime, D.J., Jolicoeur, P., 2008. Tracking the location of visuospatial attention in a contingent capture paradigm. J. Cognit. Neurosci. 20 (4), 657–671. https://doi.org/10.1162/jocn.2008.20051.
Ma, W.J., Beck, J.M., Latham, P.E., Pouget, A., 2006. Bayesian inference with probabilistic population codes. Nat. Neurosci. 9 (11), 1432–1438. https://doi.org/10.1038/nn1790.
Macaluso, E., Noppeney, U., Talsma, D., Vercillo, T., Hartcher-O'Brien, J., Adam, R., 2016. The curious incident of attention in multisensory integration: bottom-up vs. Top-down. Multisensory Res. 29, 557–583. https://doi.org/10.1163/22134808-00002528.
Maddox, R.K., Atilgan, H., Bizley, J.K., Lee, A.K., 2015. Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. ELife. https://doi.org/10.7554/eLife.04995.001.
Makeig, S., Westerfield, M., Townsend, J., Jung, T.P., Courchesne, E., Sejnowski, T.J., 1999. Functionally independent components of early event-related potentials in a visual spatial attention task. Philosophical Trans. Royal Soc. 354 (1387), 1135–1144. https://doi.org/10.1098/rstb.1999.0469.
Mangun, G.R., Hillyard, S.A., 1991. Modulation of sensory-evoked brain potentials provide evidence for changes in perceptual processing during visual-spatial priming. J. Exp. Psychol. Hum. Percept. Perform. 17 (4), 1057–1074.
Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164, 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024.
Martens, S., Munneke, J., Smid, H., Johnson, A., 2006. Quick minds don't blink: electrophysiological correlates of individual differences in attentional selection. J. Cognit. Neurosci. 18 (9), 1423–1438. https://doi.org/10.1162/jocn.2006.18.9.1423.
McCarthy, G., Wood, C.C., 1985. Scalp distributions of event-related potentials: an ambiguity associated with analysis of variance models. Electroencephalogr. Clin. Neurophysiol. 62 (3), 203–208.
Mehraei, G., Shinn-Cunningham, B., Dau, T., 2018. Influence of talker discontinuity on cortical dynamics of auditory spatial attention. Neuroimage 179, 548–556. https://doi.org/10.1016/j.neuroimage.2018.06.067.
Meredith, M.A., Nemitz, J.W., Stein, B.E., 1987. Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. J. Neurosci. 7 (10), 3215–3229. https://doi.org/10.1523/JNEUROSCI.07-10-03215.1987.
Meredith, M.A., Stein, B.E., 1986. Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. J. Neurophysiol. 56 (3), 640–662.
Meredith, M. Alex, Stein, B.E., 1990. The visuotopic component of the multisensory map in the deep laminae of the cat superior colliculus. J. Neurosci. 10 (11), 3727–3742. https://doi.org/10.1002/cne.903120304.
Miller, J., Ulrich, R., 2003. Simple reaction time and statistical facilitation: a parallel grains model. Cognit. Psychol. 46 (2), 101–151. https://doi.org/10.1016/S0010-0285(02)00517-0.
Molholm, S., Ritter, W., Murray, M.M., Javitt, D.C., Schroeder, C.E., Foxe, J.J., 2002. Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study. Cognit. Brain Res. 14 (1), 115–128. https://doi.org/10.1016/S0926-6410(02)00066-6.
Morgan, M.L., DeAngelis, G.C., Angelaki, D.E., 2008. Multisensory integration in macaque visual cortex depends on cue reliability. Neuron 59 (4), 662–673. https://doi.org/10.1016/j.neuron.2008.06.024.
Murray, M.M., Brunet, D., Michel, C.M., 2008. Topographic ERP analyses: a step-by-step tutorial review. Brain Topogr. 20 (4), 249–264. https://doi.org/10.1007/s10548-008-0054-5.
Murray, M.M., Thelen, A., Thut, G., Romei, V., Martuzzi, R., Matusz, P.J., 2016. The multisensory function of the human primary visual cortex. Neuropsychologia 83, 161–169. https://doi.org/10.1016/j.neuropsychologia.2015.08.011.
Mysore, S.P., Knudsen, E.I., 2014. Descending control of neural bias and selectivity in a spatial attention network: rules and mechanisms. Neuron 84 (1), 214–226. https://doi.org/10.1016/j.neuron.2014.08.019.
Ngo, M.K., Spence, C., 2010. Auditory, tactile, and multisensory cues facilitate search for dynamic visual stimuli. Atten. Percept. Psychophys. 72 (6), 1654–1665. https://doi.org/10.3758/APP.72.6.1654.
Novak, G., Ritter, W., Vaughan, H.G., 1992. Mismatch detection and the latency of temporal judgments. Psychophysiology 29 (4), 398–411.
Odgaard, E.C., Arieh, Y., Marks, L.E., 2004. Brighter noise: Sensory enhancement of perceived loudness by concurrent visual stimulation. Cognit. Affect Behav. Neurosci. 4 (2), 127–132. https://doi.org/10.3758/CABN.4.2.127.
Oostenveld, R., Fries, P., Maris, E., Schoffelen, J.M., 2011. FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. Comput. Intell. Neurosci. https://doi.org/10.1155/2011/156869.
Palomäki, K.J., Tiitinen, H., Mäkinen, V., May, P.J.C., Alku, P., 2005. Spatial processing in human auditory cortex: The effects of 3D, ITD, and ILD stimulation techniques. Cognit. Brain Res. 24 (3), 364–379. https://doi.org/10.1016/j.cogbrainres.2005.02.013.
Perrault, T.J., Vaughan, J.W., Stein, B.E., Wallace, M.T., 2003. Neuron-Specific response characteristics predict the magnitude of multisensory integration. J. Neurophysiol. 90 (6), 4022–4026. https://doi.org/10.1152/jn.00494.2003.
Rach, S., Diederich, A., Colonius, H., 2011. On quantifying multisensory interaction effects in reaction time and detection rate. Psychol. Res. 75 (2), 77–94. https://doi.org/10.1007/s00426-010-0289-0.
Razavi, B., 2009. Factors Influencing Human Sound Localization. University of Rochester. Retrieved from. https://urresearch.rochester.edu/fileDownloadForInstitutionalItem.action?itemId=8320&itemFileId=17539%5Cnpapers2://publication/uuid/D8019049-31CF-4EC0-A372-025022CBEFBC.
Schwartz, J.L., Berthommier, F., Savariaux, C., 2004. Seeing to hear better: evidence for early audio-visual interactions in speech identification. Cognition 93 (2), B69–B78. https://doi.org/10.1016/j.cognition.2004.01.006.
Senkowski, D., Talsma, D., Herrmann, C.S., Woldorff, M.G., 2005. Multisensory processing and oscillatory gamma responses: effects of spatial selective attention. Exp. Brain Res. 166 (3–4), 411–426. https://doi.org/10.1007/s00221-005-2381-z.
Sperdin, H.F., Cappe, C., Murray, M.M., 2010. Auditory-somatosensory multisensory interactions in humans: dissociating detection and spatial discrimination. Neuropsychologia 48 (13), 3696–3705. https://doi.org/10.1016/j.neuropsychologia.2010.09.001.
Squires, N.K., Squires, K.C., Hillyard, S.A., 1975. Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. Electroencephalogr. Clin. Neurophysiol. 38 (4), 387–401. https://doi.org/10.1016/0013-4694(75)90263-1.
Stanford, T.R., Quessy, S., Stein, B.E., 2005. Evaluating the operations underlying multisensory integration in the cat superior colliculus. J. Neurosci. 25 (28), 6499–6508. https://doi.org/10.1523/jneurosci.5095-04.2005.
Stanford, Terrence R., Stein, B.E., 2007. Superadditivity in multisensory integration: putting the computation in context. Neuroreport 18 (8), 787–792. https://doi.org/10.1097/WNR.0b013e3280c1e315.
Stein, B.E., Meredith, M.A., 1993. The Merging of the Senses. The MIT Press, Cambridge, MA, US.
Stein, B.E., Stanford, T.R., 2008. Multisensory integration: current issues from the perspective of the single neuron. Nat. Rev. Neurosci. 9 (4), 255–266. https://doi.org/10.1038/nrn2331.
Stevenson, R.A., Zemtsov, R.K., Wallace, M.T., 2012. Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. J. Exp. Psychol. Hum. Percept. Perform. 38 (6), 1517–1529. https://doi.org/10.1037/a0027339.
Talsma, D., Doty, T.J., Woldorff, M.G., 2007. Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? Cerebr. Cortex 17 (3), 679–690. https://doi.org/10.1093/cercor/bhk016.
Talsma, D., Senkowski, D., Soto-Faraco, S., Woldorff, M.G., 2010. The multifaceted interplay between attention and multisensory integration. Trends Cognit. Sci. 14 (9), 400–410. https://doi.org/10.1016/j.tics.2010.06.008.
Tang, X., Wu, J., Shen, Y., 2016. The interactions of multisensory integration with endogenous and exogenous attention. Neurosci. Biobehav. Rev. 61, 208–224. https://doi.org/10.1016/j.neubiorev.2015.11.002.
Teder-Sälejärvi, W.A., McDonald, J., Di Russo, F., Hillyard, S., 2002. An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. Cognit. Brain Res. 14 (1), 106–114. https://doi.org/10.1016/S0926-6410(02)00065-4.
Töllner, T., Zehetleitner, M., Gramann, K., Müller, H.J., 2011. Stimulus saliency modulates pre-attentive processing speed in human visual cortex. PloS One 6 (1). https://doi.org/10.1371/journal.pone.0016276.
Van den Brink, R.L., Cohen, M.X., Van der Burg, E., Talsma, D., Vissers, M.E., Slagter, H. A., 2014. Subcortical, modality-specific pathways contribute to multisensory processing in humans. Cerebr. Cortex 24 (8), 2169–2177. https://doi.org/10.1093/cercor/bht069.

Van der Burg, E., Awh, E., Olivers, C.N.L., 2013. The capacity of audiovisual integration is limited to one item. Psychol. Sci. 24 (3), 345–351. https://doi.org/10.1177/0956797612452865.

Van der Burg, E., Cass, J., Olivers, C.N.L., Theeuwes, J., Alais, D., 2010. Efficient visual search from synchronized auditory signals requires transient audiovisual events. PloS One 5 (5). https://doi.org/10.1371/journal.pone.0010664.

Van der Burg, E., Olivers, C.N.L., Bronkhorst, A.W., Theeuwes, J., 2008. Pip and pop: nonspatial auditory signals improve spatial visual search. J. Exp. Psychol. Hum. Percept. Perform. 34 (5), 1053–1065. https://doi.org/10.1037/0096-1523.34.5.1053.

Van der Burg, E., Olivers, C.N.L., Cass, J., 2017. Evolving the keys to visual crowding. J. Exp. Psychol. Hum. Percept. Perform. 43 (4), 690–699. https://doi.org/10.1037/xhp0000337.

van der Burg, E., Olivers, C.N.L., Theeuwes, J., 2012. The attentional window modulates capture by audiovisual events. PloS One 7 (7), e39137. https://doi.org/10.1371/journal.pone.0039137.

Van der Burg, E., Talsma, D., Olivers, C.N.L., Hickey, C., Theeuwes, J., 2011. Early multisensory interactions affect the competition among multiple visual objects. Neuroimage 55 (3), 1208–1218. https://doi.org/10.1016/j.neuroimage.2010.12.068.

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.

Woldorff, M.G., Gallen, C.C., Hampson, S.A., Hillyard, S.A., Pantev, C., Sobel, D., Blooms, F.E., 1993. Modulation of early sensory processing in human auditory cortex during auditory selective attention. Neurobiology 90 (18), 8722–8726. https://doi.org/10.1073/pnas.90.18.8722.

Wolfe, J.M., Cave, K.R., Franzel, S.L., 1989. Guided search: an alternative to the feature integration model for visual search. J. Exp. Psychol. 15 (3), 419–433.

Zou, H., Müller, H.J., Shi, Z., 2012. Non-spatial sounds regulate eye movements and enhance visual search. J. Vis. 12 (5), 1–18. https://doi.org/10.1167/12.5.2, 2.