Neuroscience

Inverted encoding of neural responses to audiovisual stimuli reveals super-additive multisensory enhancement

Reviewed Preprint
v2 • December 6, 2024
Revised by authors

Reviewed Preprint
v1 • May 17, 2024

Zak Buhmann ✉, **Amanda K Robinson, Jason B Mattingley, Reuben Rideaux**

Queensland Brain Institute, The University of Queensland, Brisbane, Australia • School of Psychology, The University of Queensland, Brisbane, Australia • School of Psychology, University of Sydney, Sydney, Australia

eLife Assessment

Despite the well-established facilitatory effects of multisensory integration on behavioural measures, standard neuroimaging approaches have yet to reliably and precisely identify the corresponding neural correlates. In this **valuable** paper, Buhmann et al. leverage EEG decoding methods, moving beyond traditional univariate analyses, to capture these correlates. They present **solid** evidence that this approach can effectively estimate multisensory integration in humans across a broad range of contexts.

https://doi.org/10.7554/eLife.97230.2.sa3

Abstract

A central challenge for the brain is how to combine separate sources of information from different sensory modalities to optimally represent objects and events in the external world, such as combining someone's speech and lip movements to better understand them in a noisy environment. At the level of individual neurons, audiovisual stimuli often elicit super-additive interactions, where the neural response is greater than the sum of auditory and visual responses. However, investigations using electroencephalography (EEG) to record brain activity have revealed inconsistent interactions, with studies reporting a mix of super- and sub-additive effects. A possible explanation for this inconsistency is that standard univariate analyses obscure multisensory interactions present in EEG responses by overlooking multivariate changes in activity across the scalp. To address this shortcoming, we investigated EEG responses to audiovisual stimuli using inverted encoding, a population tuning approach that uses multivariate information to characterise feature-specific neural activity. Participants (n = 41) completed a spatial localisation task for both unisensory stimuli (auditory clicks, visual flashes) and combined audiovisual stimuli (spatiotemporally congruent clicks and flashes). To assess multivariate changes in EEG activity, we used inverted encoding to recover stimulus location information from event-related potentials (ERPs). Participants localised audiovisual stimuli more accurately than unisensory stimuli alone. For univariate ERP analyses we found an additive multisensory interaction. By contrast, multivariate analyses revealed a super-additive interaction ~180 ms following stimulus onset, such that the location of audiovisual stimuli was decoded more accurately

than that predicted by maximum likelihood estimation. Our results suggest that super-additive integration of audiovisual information is reflected within multivariate patterns of activity rather than univariate evoked responses.

Introduction

We exist in a complex, dynamically changing sensory environment. Vertebrates, including humans, have evolved sensory organs that transduce relevant sources of physical information, such as light and changes in air pressure, into patterns of neural activity that support perception (vision and audition) and adaptive behaviour. Such activity patterns are noisy, and often ambiguous, due to a combination of external (environmental) and internal (transduction) factors. Critically, information from the different sensory modalities can be highly correlated because it is often elicited by a common external source or event. For example, the sight and sound of a hammer hitting a nail produces a single, unified perceptual experience, as does the sight of a person's lips moving as we hear their voice. To improve the reliability of neural representations, the brain leverages these sensory relationships by combining information in a process referred to as *multisensory integration*. The existence of such processes heighten perception, e.g., by making it easier to understand a person's speech in a noisy setting by looking at their lip movements (Sumby & Pollack, 1954).

Multisensory integration of audiovisual cues improves performance across a range of behavioural outcomes, including detection accuracy (Bolognini et al., 2005; Frassinetti et al., 2002; Lovelace et al., 2003), response speed (Arieh & Marks, 2008; Cappe et al., 2009; Colonius & Diederich, 2004; Rach & Diederich, 2006; Senkowski et al., 2011), and saccade speed and accuracy (Corniel et al., 2002; Van Wanrooij et al., 2009). Successful integration requires the constituent stimuli to occur at approximately the same place and time (Leone & McCourt, 2015). The degree to which behavioural performance is improved follows the principles of maximum likelihood estimation (MLE), wherein sensory information from each modality is weighted and integrated according to its relative reliability (Alais & Burr, 2004; Ernst & Banks, 2002; although other processing schemes have also been identified; Rideaux & Welchman, 2018). As such, behavioural performance that matches MLE predictions is often seen as a benchmark of successful, optimal integration of relevant unisensory cues.

The ubiquity of behavioural enhancements for audiovisual stimuli suggests there are fundamental neural mechanisms that facilitate improved precision. Recordings from single multisensory (audiovisual) neurons within cat superior colliculus have revealed the principle of inverse effectiveness, whereby the increased response to audiovisual stimuli is larger when the constituent unisensory stimuli are weakly stimulating (Corniel et al., 2002; Meredith & Stein, 1983). Depending on the intensity of the integrated stimuli, the neural response can be either *super-additive,* where the multisensory response is greater than the sum of the unisensory responses, *additive,* equal to the sum of responses, or *sub-additive*, where the combined response is less than the sum of the unisensory responses (see Stein & Stanford, 2008). Inverse effectiveness has also been observed in human behavioural experiments, with low intensity audiovisual stimuli eliciting greater multisensory enhancements in response precision than those of high intensity (Colonius & Diederich, 2004; Corniel et al., 2002; Rach & Diederich, 2006; Rach et al., 2010).

Neuroimaging methods, such as electroencephalography (EEG) and functional magnetic resonance imaging (fMRI), have been used to investigate neural population-level audiovisual integration in humans. These studies have typically applied an additive criterion to quantify multisensory integration, wherein successful integration is marked by a non-linear enhancement of audiovisual responses relative to unisensory responses (Besle et al., 2004). The super- or sub-additive nature of this enhancement, however, is often inconsistent. In fMRI, neural super-additivity in blood-

oxygen-level dependent (BOLD) responses to audiovisual stimuli has been found in a variety of regions, primarily the superior temporal sulcus (STS; Calvert et al., 2000; Calvert et al., 2001; Stevenson et al., 2007; Stevenson & James, 2009; Werner & Noppeney, 2010, 2011). However, other studies have failed to replicate audiovisual super-additivity in the STS (Joassin et al., 2011; Porada et al., 2021; Ross et al., 2022; Venezia et al., 2015), or have found sub-additive responses (see Scheliga et al., 2023 for review). As such, some have argued that BOLD responses are not sensitive enough to adequately characterise super-additive audiovisual interactions within populations of neurons (Beauchamp, 2005; James et al., 2012; Laurienti et al., 2005). In EEG, meanwhile, the evoked response to an audiovisual stimulus typically conforms to a sub-additive principle (Cappe et al., 2010; Fort et al., 2002; Giard & Peronnet, 1999; Murray et al., 2016; Puce et al., 2007; Stekelenburg & Vroomen, 2007; Teder-Sälejärvi et al., 2002; Vroomen & Stekelenburg, 2010). However, when the principle of inverse effectiveness is considered and relatively weak stimuli are presented together, there has been some evidence for super-additive responses (Senkowski et al., 2011).

It is important to consider the differences in how super-additivity is classified between neural and behavioural measures. At the level of single neurons, super-additivity is defined as a non-linear response enhancement, with the multisensory response exceeding the sum of the unisensory responses. In behaviour, meanwhile, it has been observed that the performance improvement from combining two senses is close to what is expected from optimal integration of information across the senses (Alais & Burr, 2004; Stanford & Stein, 2007). Critically, behavioural enhancement of this kind does not require non-linearity in the neural response, but can arise from a reliability-weighted average of sensory information. In short, behavioural performance that conforms to MLE is not necessarily indicative of neural super-additivity, and the MLE model can be considered a linear baseline for multisensory integration.

While behavioural outcomes for multisensory stimuli can be predicted by MLE, and single neuron responses follow the principles of inverse effectiveness and super-additivity, among others (Rideaux et al., 2021), how audiovisual super-additivity manifests within populations of neurons is comparatively unclear given the mixed findings from relevant fMRI and EEG studies. This uncertainty may be due to biophysical limitations of human neuroimaging techniques, but it may also be related to the analytic approaches used to study these recordings. For instance, super-additive responses to audiovisual stimuli in EEG studies are often reported from very small electrode clusters (Molholm et al., 2002; Senkowski et al., 2011; Talsma et al., 2007), suggesting that neural super-additivity in humans may be highly specific. However, information encoded by the brain can be represented as increased activity in some areas, accompanied by decreased activity in others, so simplifying complex neural responses to the average rise and fall of activity in specific sensors may obscure relevant multivariate patterns of activity evoked by a stimulus.

Inverted encoding is a multivariate analytic method that can reveal how sensory information is encoded within the brain by recovering patterns of neural activity associated with different stimulus features. This method has been successfully used in fMRI, EEG, and magnetoencephalography studies to characterise the neural representations of a range of stimulus features, including colour (Brouwer & Heeger, 2009), spatial location (Bednar & Lalor, 2020; Robinson et al., 2021) and orientation (Brouwer & Heeger, 2011; Harrison et al., 2023; Kok et al., 2017). A multivariate approach may capture potential non-linear enhancements associated with audiovisual responses and thus could reveal super-additive interactions that would otherwise be hidden within the brain's univariate responses. The sensitivity of inverted encoding analyses to multivariate neural patterns may provide insight into how audiovisual information is processed and integrated at the population level.

In the present study, we investigated neural super-additivity in human audiovisual sensory processing using inverted encoding of EEG responses during a task where participants had to spatially localise visual, auditory, and audiovisual stimuli. In a separate behavioural experiment, we monitored response accuracy to characterise behavioural improvements to audiovisual relative to unisensory stimuli. Although there was no evidence for super-additivity in response to audiovisual stimuli within univariate ERPs, we observed a reliable non-linear enhancement of multivariate decoding performance at ~180 ms following stimulus onset when auditory and visual stimuli were presented concurrently as opposed to alone. These findings suggest that population-level super-additive multisensory neural responses are present within multivariate patterns of activity rather than univariate evoked responses.

Results

Behavioural performance

Participants performed well in discriminating stimulus location across all conditions in both the behavioural and EEG sessions (**Figure 1**). For the behavioural session, the psychometric curves for responses as a function of stimulus location showed stereotypical relationships for the auditory, visual, and audiovisual conditions (**Figure 1A**). A quantification of the behavioural sensitivity (i.e., steepness of the curves) revealed significantly higher sensitivity for the audiovisual stimuli (M = .04, SD = .02) than for the auditory stimuli alone (M = .03, SD = .01; Z = -3.09, p = .002), and than for the visual stimuli alone (M = .02, SD = .01; Z = -5.28, p = 1.288e-7; **Figure 1B**). Sensitivity for auditory stimuli was also significantly higher than sensitivity for visual stimuli (Z = 2.02, p = .044). To test for successful integration of stimuli in the audiovisual condition, we calculated the predicted MLE sensitivity from the unisensory auditory and visual results. We found no evidence for a significant difference between the predicted and actual audiovisual sensitivity (Z = - 1.54, p = 0.125).

We repeated these analyses for behavioural performance in the EEG session (**Figure 1C**). We found a similar pattern of results to those in the behavioural session; sensitivity for audiovisual stimuli (M = .85, SD = .33) was significantly higher than for auditory (M = .69, SD = .41; Z = -2.27, p = .023) and visual stimuli alone (M = .61, SD = .29; Z = -3.52, p = 4.345e-4), but not significantly different from the MLE prediction (Z = -1.07, p = .285). However, sensitivity for auditory stimuli was not significantly different from sensitivity to visual stimuli (Z = 1.12, p = .262).

Event-related potentials

We plotted the ERPs for auditory, visual, and audiovisual conditions at each stimulus location from -100ms to 500ms around stimulus presentation (**Figure 2**). For each stimulus location, cluster corrected t-tests were conducted to assess significant differences in ERP amplitude between the unisensory (auditory and visual) and audiovisual conditions. While auditory ERPs did not significantly differ from the audiovisual, visual ERPs were significantly lower in amplitude than audiovisual ERPs at all stimulus locations (typically from ~80-130 ms following stimulus presentation).

To test whether the enhancement in response amplitude to audiovisual stimuli was super-additive, we compared this response with the sum of the response amplitudes for visual and auditory conditions, averaged over stimulus location. We found no significant difference between the additive and audiovisual ERPs (**Figure 2**, bottom right). This result suggests that, using univariate analyses, the audiovisual response was additive and did not show any evidence for super- or sub-additivity.

Figure 1.

Behavioural performance is improved for audiovisual stimuli.

A) Average accuracy of responses across participants in the behavioural session at each stimulus location for each stimulus condition, fitted to a psychometric curve. Steeper curves indicate greater sensitivity in identifying stimulus location. **B**) Average sensitivity across participants in the behavioural task, estimated from psychometric curves, for each stimulus condition. The red cross indicates estimated performance assuming optimal (MLE) integration of unisensory cues. **C**) Average behavioural sensitivity across participants in the EEG session for each stimulus condition. Error bars indicate ±1 SEM.

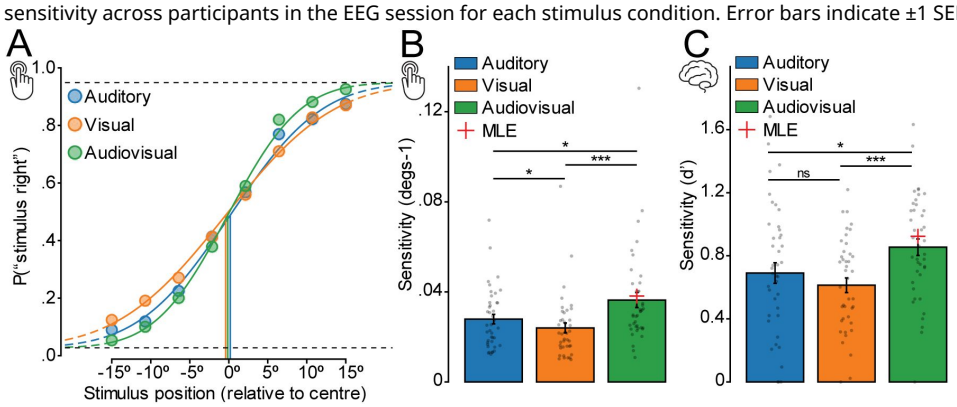
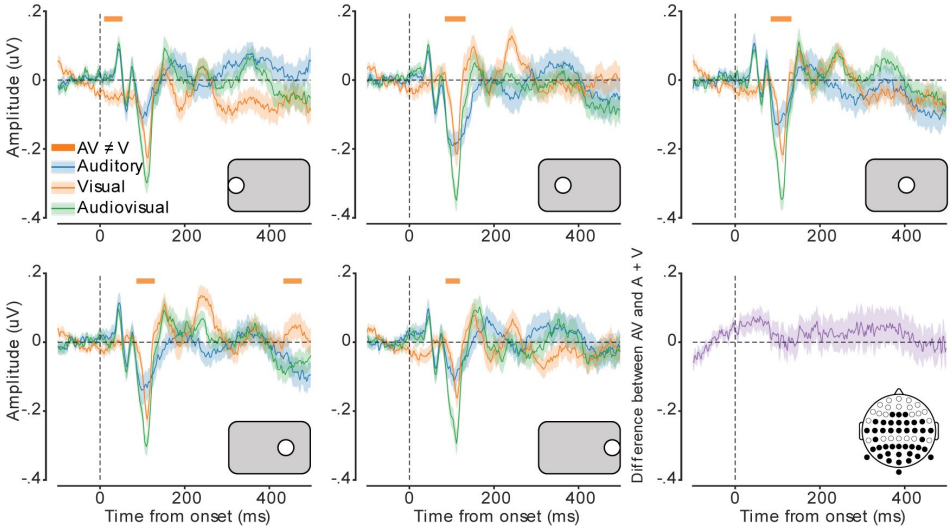


Figure 2.

Audiovisual ERPs follow an additive principle.

Average ERP amplitude for each modality condition. Five plots represent the different stimulus locations, as indicated by the grey inset, and the final plot (bottom-right) shows the difference between the summed auditory and visual responses and the audiovisual response. Shaded error bars indicate ±1 SEM. Orange horizontal bars indicate cluster corrected periods of significant difference between visual and audiovisual ERP amplitudes.



Inverted encoding results

We next used inverted encoding to calculate the spatial decoding accuracy for auditory, visual, and audiovisual stimuli (**Figure 3A**). For all conditions, we found that spatial location could be reliably decoded from approximately ~100-150 ms after stimulus onset. Decoding for all conditions was consistent for most of the epoch, indicating that location information within the neural signal was relatively persistent and robust.

To assess the spatial representation of the neural signal containing location-relevant information, we computed decoding accuracy at each electrode from 150-250 ms post-stimulus presentation (**Figure 3B**). For auditory stimuli, information was primarily based over bilateral temporal regions, whereas for visual and audiovisual stimuli, the occipital electrodes carried the most information.

Multivariate super-additivity

Although the univariate response did not show evidence for super-additivity, we expected the multivariate measure would be more sensitive to nonlinear audiovisual integration. To test whether a super-additive interaction was present in the multivariate response, we calculated the sensitivity of the decoder in discriminating stimuli presented on the left and right side. The pattern of decoding sensitivity for auditory, visual, and audiovisual stimuli (**Figure 4A**) was similar to that in decoding accuracy (**Figure 3A**). Notably, audiovisual sensitivity was significantly greater than sensitivity to auditory and visual stimuli alone, particularly ~180 ms following stimulus onset. To test whether this enhanced sensitivity reflected super-additivity, we compared decoding sensitivity for audiovisual stimuli with two estimates of linearly combined unisensory stimuli: 1) MLE predicted sensitivity based on auditory and visual sensitivity and 2) aggregate responses of auditory and visual stimuli (**Figure 4B**). We found that audiovisual sensitivity significantly exceeded both estimates of linear combination (MLE, ~160-220 ms post-stimulus; aggregate, ~150-250 ms). These results provide evidence of non-linear audiovisual integration in the multivariate pattern of EEG recordings. Taken together with the ERP results, our findings suggest that super-additive integration of audiovisual information is reflected in multivariate patterns of activity, but not univariate evoked responses.

Neurobehavioural correlations

As behavioural and neural data violated assumptions of normality, we calculated rank-order correlations (Spearman's rho) between the average decoding sensitivity for each participant from 150-250 ms post-stimulus onset and behavioural performance on the EEG task. As Spearman's rho is resistant to outliers (Wilcox, 2016), we did not perform outlier rejection. We found that decoding sensitivity was significantly positively correlated with behavioural sensitivity for audiovisual stimuli (r = .43, p = .003), but not for auditory (r = -.04, p = .608) or visual stimuli (r = .15, p = .170) alone.

Discussion

We tested for super-additivity in multivariate patterns of EEG responses to audiovisual stimuli. Participants judged the location of auditory, visual, and audiovisual stimuli while their brain activity was measured using EEG. As expected, participants' behavioural responses to audiovisual stimuli were more precise than that for unisensory auditory and visual stimuli. ERP analyses showed that although audiovisual stimuli elicited larger responses than visual stimuli, the overall response followed an additive principle. Critically, our multivariate analyses revealed that

Figure 3.

Spatiotemporal representation of audiovisual location.

A) Accuracy of locations decoded from neural responses for each stimulus condition. Shaded error bars indicate ±1 SEM. Coloured horizontal bars indicate cluster corrected periods that showed a significant difference from chance (0). **B**) Topographic decoding performance in each condition during critical period (grey inset in (**A**)).

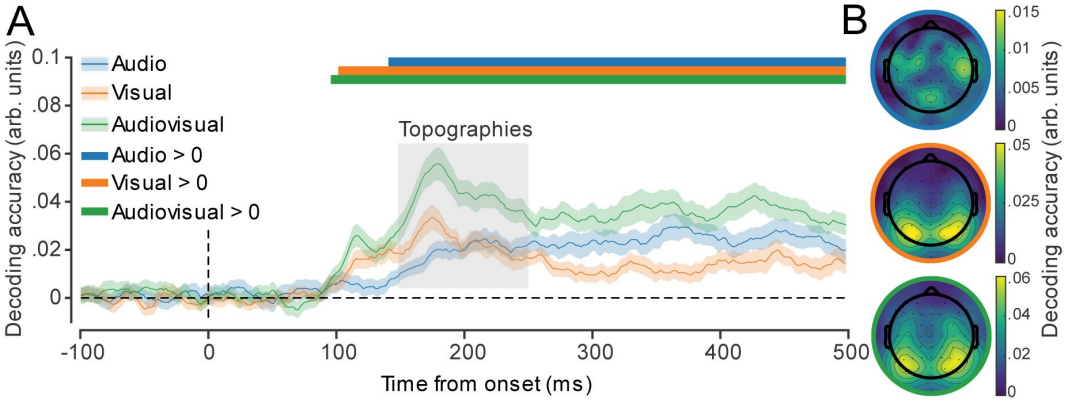
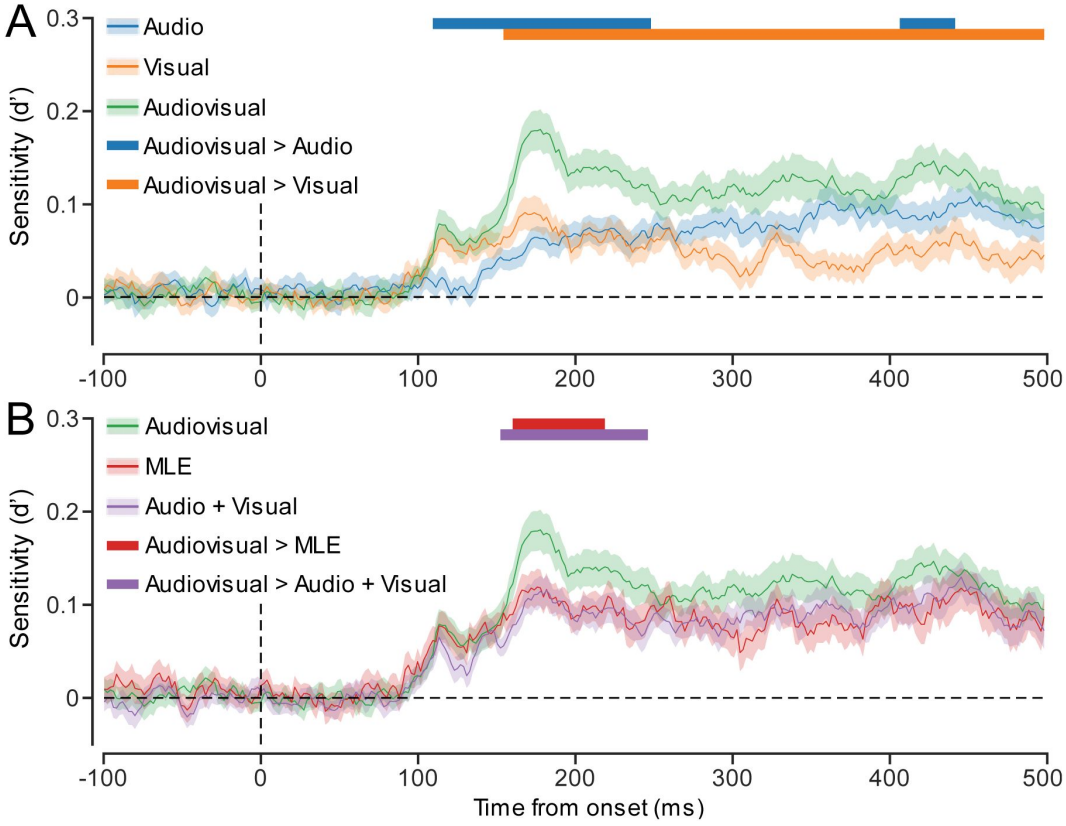


Figure 4.

Super-additive multisensory interaction in multivariate patterns of EEG activity.

A) Decoding sensitivity in each stimulus condition across the epoch. Overall trends closely matched decoding accuracy. **B**) Predicted (optimal sensitivity through MLE and aggregate A+V) and actual audiovisual sensitivity across the epoch. Coloured horizontal bars indicate cluster corrected periods where actual sensitivity significantly exceeded that which was predicted. Shaded error bars indicate ±1 SEM.



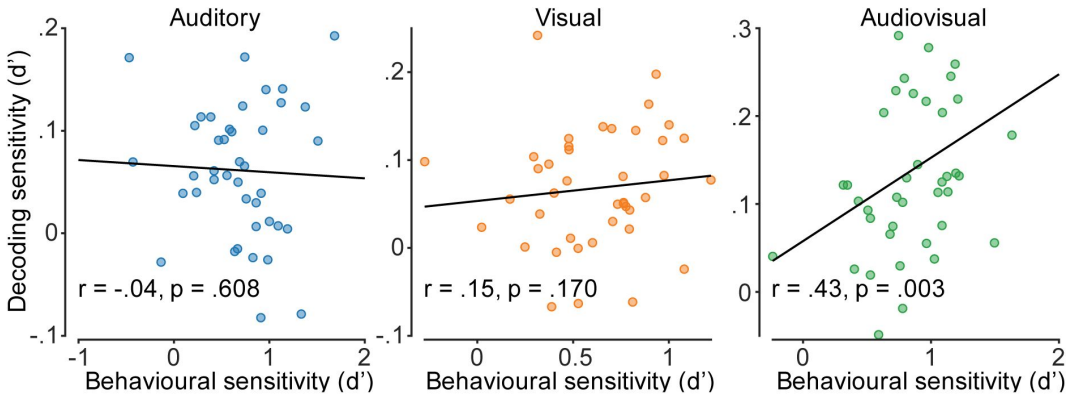


Figure 5.

Audiovisual decoding sensitivity is significantly positively correlated with behavioural sensitivity.

Correlations (Spearman's rho) are shown between decoding and behavioural sensitivity from the EEG session (150-250 ms post-stimulus onset) for each stimulus condition, with a line of best fit.

decoding sensitivity for audiovisual stimuli exceeded predictions of both MLE and aggregate auditory and visual information, indicating non-linear multisensory enhancement (i.e., super-additivity).

Participants localised audiovisual stimuli more accurately than unisensory in both the behavioural and EEG sessions. This behavioural facilitation in response to audiovisual stimuli is well-established within the literature (Bolognini et al., 2005; Frassinetti et al., 2002; Lovelace et al., 2003; Meredith & Stein, 1983; Senkowski et al., 2011). In accordance with theories of optimal cue integration, we found participants' performance for audiovisual stimuli in both sessions matched that predicted by MLE (Ernst & Banks, 2002). Matching this 'optimal' prediction of performance indicates that the auditory and visual cues were successfully integrated when presented together in the audiovisual condition (Fetsch et al., 2013).

Our EEG analyses revealed that for most spatial locations, audiovisual stimuli elicited a significantly greater neural response than exclusively visual stimuli approximately 80-120 ms after stimulus onset. Despite numerically larger ERPs to audiovisual than auditory stimuli, this effect failed to reach significance, most likely due to greater inter-trial variability in the auditory ERPs. Critically, however, the audiovisual ERPs consistently matched the sum of visual and auditory ERPs. Sub- or super-additive interaction effects in neural responses to multisensory stimuli are a hallmark of successful integration of unisensory cues in ERPs (Besle et al., 2004; Stevenson et al., 2014). An additive ERP in this context cannot imply successful multisensory integration, as the multisensory 'enhancement' may be the result of recording from distinct populations of unisensory neurons responding to the two unisensory sensory modalities (Besle et al., 2009). This invites the question of why we see evidence for integration at the behavioural level, but not in the amplitude of neural responses. One explanation could be that the signals measured by EEG simply do not contain evidence of non-linear integration because the super-additive responses are highly spatiotemporally localized and filtered out by the skull before reaching the EEG sensors. Another possibility, however, is that evidence for non-linear integration is only observable within the changing pattern of ERPs across sensors. Indeed, Murray et al. (2016) found that multisensory interactions followed from changes in scalp topography rather than net gains to ERP amplitude.

Our decoding results reveal that not only do audiovisual stimuli elicit more distinguishable patterns of activity than visual and auditory stimuli, but this enhancement exceeds that predicted by both optimal integration and the aggregate combination of auditory and visual responses. Critically, the non-linear enhancement of decoding sensitivity for audiovisual stimuli indicates the presence of an interactive effect for the integration of auditory and visual stimuli that was not evident from the univariate analyses. This indicates super-additive enhancement of the neural representation of integrated audiovisual cues, and supports the interpretation that increased behavioural performance for multisensory stimuli is related to a facilitation of the neural response (Fetsch et al., 2013). This interaction was absent from univariate analyses (Nikbakht et al., 2018), suggesting that the neural facilitation of audiovisual processing is more nuanced than net increased excitation, and may be associated with a complex pattern of excitatory and inhibitory neural activity, e.g., divisive normalization (Ohshiro et al., 2017).

The non-linear neural enhancement in decoding sensitivity for audiovisual stimuli occurred ~180 ms after stimulus onset, which is later than previously reported audiovisual interactions (<150 ms; Cappe et al., 2010; Fort et al., 2002; Giard & Peronnet, 1999; Molholm et al., 2002; Murray et al., 2016; Senkowski et al., 2011; Talsma et al., 2007; Teder-Sälejärvi et al., 2002). As stimulus characteristics and task requirements are likely to have a significant influence over the timing of multisensory interaction effects in EEG activity (Calvert & Thesen, 2004; De Meo et al., 2015), our use of peripheral spatial locations (where previous studies only used stimuli centrally) may explain the slightly later timing of our audiovisual effect. Indeed, our finding is

consistent with previous multivariate studies which found that location information in EEG data, for both visual (Rideaux, 2024; Robinson et al., 2021) and auditory (Bednar & Lalor, 2020) stimuli, is maximal at ~190 ms following stimulus presentation.

An interesting aspect of our results is the apparent mismatch between the behavioural and neural responses. While the behavioural results meet the optimal statistical threshold predicted by MLE, the decoding analyses suggest that the neural response exceeds it. Though non-linear neural responses and statistically optimal behavioural responses are reliable phenomena in multisensory integration (Alais & Burr, 2004; Ernst & Banks, 2002; Stanford & Stein, 2007), the question remains – if neural super-additivity exists to improve behavioural performance, why is it not reflected in behavioural responses? A possible explanation for this neurobehavioural discrepancy is the large difference in timing between sensory processing and behavioural responses. A motor response would typically occur some time after the neural response to a sensory stimulus (e.g., 70-200 ms), with subsequent neural processes between perception and action that introduce noise (Heekeren et al., 2008) and may obscure super-additive perceptual sensitivity. In the current experiment, participants reported either the distribution of 20 serially presented stimuli (EEG session) or compared the positions of two stimuli (behavioural session), whereas the decoder attempts to recover the location of every presented stimulus. While stimulus location could be represented with higher fidelity in multisensory relative to unisensory conditions, this would not necessarily result in better performance on a binary behavioural task in which multiple temporally separated stimuli are compared. One must also consider the inherent differences in how super-additivity is measured at the neural and behavioural levels. Neural super-additivity should manifest in responses to each individual stimulus. In contrast, behavioural super-additivity is often reported as proportion correct, which can only emerge between conditions after being averaged across multiple trials. The former is a biological phenomenon, while the latter is an analytical construct. In our experiment, we recorded neural responses for every presentation of a stimulus, but behavioural responses were only obtained after multiple stimulus presentations. Thus, the failure to find super-additivity in behavioural responses might be due to their operationalisation, with between-condition comparisons lacking sufficient sensitivity to detect super-additive sensory improvements. Future work should focus on experimental designs that can reveal super-additive responses in behaviour.

We also found a significant positive correlation between participants' behavioural judgements in the EEG session and decoding sensitivity for audiovisual stimuli. This result suggests that participants who were better at identifying stimulus location also had more reliably distinct patterns of neural activity. The lack of neurobehavioural correlation in the unisensory conditions might suggest a poor correspondence between the different tasks, perhaps indicative of the differences between behavioural and neural measures explained previously. However, multisensory stimuli have consistently been found to elicit stronger neural responses than unisensory stimuli (Meredith & Stein, 1983; Puce et al., 2007; Senkowski et al., 2011; Vroomen & Stekelenburg, 2010), which has been associated with behavioural performance (Frens & Van Opstal, 1998; Wang et al., 2008). Thus, the weaker signal-to-noise ratio in unisensory conditions may prevent correlations from being detected.

Any experimental design that varies stimulus location needs to consider the potential contribution of eye movements. We computed correlations between participants' average eye position and each stimulus position between the three sensory conditions (auditory, visual and audiovisual; Figure S1) and found evidence that participants made eye movements toward stimuli. A re-analysis of the data with a very strict eye-movement criterion (i.e., removing trials with eye movements >1.875°) revealed that the super-additive enhancement in decoding accuracy no longer survived cluster correction, suggesting that our results may be impacted by the consistent motor activity of saccades towards presented stimuli. Further investigation, however, suggests this is unlikely. Though the correlations were significantly different from 0, they were not significantly different from each other. If consistent saccades to audiovisual stimuli were responsible for the nonlinear

multisensory benefit we observed, we would expect to find a higher positive correlation between horizontal eye position and stimulus location in the audiovisual condition than in the auditory or visual conditions. Interestingly, eye movements corresponded more to stimulus location in the auditory and audiovisual conditions than in the visual condition, indicating that it was the presence of a sound, rather than a visual stimulus, that drove small eye movements. This could indicate that participants inadvertently moved their eyes when localising the origin of sounds. We also re-ran our analyses using the activity measured from the frontal electrodes alone (Figure S2). If the source of the nonlinear decoding accuracy in the audiovisual condition was due to muscular activity produced by eye movements, there should be better decoding accuracy from sensors closer to the source. Instead, we found that decoding accuracy of stimulus location from the frontal electrodes (peak d' = 0.08) was less than half that of decoding accuracy from the more posterior electrodes (peak d' = 0.18). These results suggest that the source of neural activity containing information about stimulus position was located over occipito-parietal areas, consistent with our topographical analyses (inset of **Figure 3**).

In summary, here we have shown a non-linear enhancement in the neural representation of audiovisual stimuli relative to unisensory (visual/auditory) stimuli. This enhancement was obscured within univariate ERP analyses focusing exclusively on response amplitude but was revealed through inverted encoding analyses in feature-space, suggesting that super-additive integration of audiovisual information is reflected within multivariate patterns of activity rather than univariate evoked responses. Further research on the multivariate representation of audiovisual integration may shed light on the neural mechanisms that facilitate this non-linear enhancement. In particular, future work may consider the influence of different stimulus features and task requirements on the timing and magnitude of the audiovisual enhancement. How and when auditory and visual information are integrated to enhance multisensory processing remains an open question, with evidence for a complex combination of top-down and bottom-up interactions (Delong & Noppeney, 2021; Keil & Senkowski, 2018; Rohe & Noppeney, 2018). Our study highlights the importance of considering multivariate analyses in multisensory research, and the potential loss of stimulus-relevant neural information when relying solely on univariate responses.

Methods

Participants

Seventy-one human adults were recruited in return for payment. The study was approved by The University of Queensland Human Research Ethics Committee, and informed consent was obtained in all cases. Participants were first required to complete a behavioural session with above 60% accuracy in all conditions to qualify for the EEG session (see Behavioural session for details). Twenty-nine participants failed to meet this criterion and were excluded from further participation and analyses, along with one participant who failed to complete the EEG session with above chance behavioural accuracy. This left a total of 41 participants (*M* = 27.21 yrs; min = 20 yrs; max = 64 yrs; 24 females; 41 right-handed). Participants reported no neurological or psychiatric disorders, and had normal visual acuity (assessed using a standard Snellen eye chart).

Materials and procedure

The experiment was split into two separate sessions, with participants first completing a behavioural session followed by an EEG session. Each session had three conditions, in which the presented stimuli were either visual, auditory, or combined audio and visual (audiovisual). Conditions were not interleaved, but the order in which conditions were presented was counterbalanced across participants. Before each task, participants were given instructions and completed two rounds of practice for each condition.

Apparatus

The experiment was conducted in a dark, acoustically and electromagnetically shielded room. For the EEG session, stimuli were presented on a 24-inch ViewPixx monitor (VPixx Technologies Inc., Saint-Bruno, QC) with 1920×1080-pixel resolution and a refresh rate of 144 Hz. Viewing distance was maintained at 54 cm using a chinrest. For the behavioural session, stimuli were presented on a 32-inch Cambridge Research Systems Display++ LCD monitor with a 1920×1080-pixel resolution, hardware gamma correction and a refresh rate of 144Hz. Viewing distance was maintained at 59.5 cm using a chinrest. Stimuli were generated in MATLAB v2021b (The MathWorks Inc., 2021) using the Psychophysics Toolbox (Brainard, 1997). Auditory stimuli were played through two loudspeakers placed either side of the display (80 cm apart for the behavioural session, 58cm apart for the EEG session). The loudspeakers had a power handling capacity of 25-75 W and a nominal impedance of 6 Ω. In both experiments, an EyeLink 1000 infrared eye tracker recorded gaze direction (SR Research Ltd., 2009) at a sampling rate of 1000 Hz.

Stimuli

The EEG and behavioural paradigms used the same stimuli within each condition. Visual stimuli were gaussian blobs (0.2 contrast, 16° diameter) presented for 16 ms on a mid-grey background. Auditory stimuli were 100 ms, 850 Hz tones with a decay function (sample rate = 44, 100 Hz; volume = 60 dBA SPL, as measured at the ears). Audiovisual stimuli were spatially and temporally matched combinations of the audio and visual stimuli, with no changes to stimuli properties. To manipulate spatial location, target stimuli were presented from multiple horizontal locations along the display, centred on linearly spaced locations from 15° visual angle to the left and right of the display centre (eight locations for behavioural, five for EEG). Auditory stimuli were played through two speakers placed equidistantly either side of the display. The perceived source location of auditory stimuli was manipulated via changes to interaural level and timing (Whitworth & Jeffress, 1961; Wightman & Kistler, 1992). The precise timing of when each speaker delivered an auditory stimulus was calculated from the following formula:

$$ITD_{L,R} = \frac{\sqrt{(x \pm r)^2 + z^2}}{s}$$

where x and z are the horizontal and forward distances in metres between the ears and the source of the sound on the display, respectively, r is the head radius, and s is the speed of sound. We used a constant approximate head radius of 8 cm for all participants. r was added to x for the left speaker and subtracted for the right speaker to produce the interaural time difference. For ±15° source locations, interaural timing difference was 1.7 ms. To simulate the decrease in sound intensity as a function of distance, we calculated interaural level differences for the left and right speakers by dividing the sounds by the left and right distance vectors. Finally, we resampled the sound using linear interpolation based on the calculations of the interaural level and timing differences. This process was used to calculate the soundwaves played by the left and right speakers for each of the possible stimulus locations on the display. The maximum interaural level difference between speakers was 0.14 A for ±15° auditory locations, and 0.07 A for ±7.5°.

Behavioural Session

During pilot testing, stimulus features (size, luminance, volume, frequency etc.) were manipulated to make visual and auditory stimuli similarly difficult to spatially localize. These values were held constant in the main experiment. We employed a two-interval forced choice design to measure participants' audiovisual localization sensitivity. Participants were presented with two consecutive stimuli and tasked with indicating, via button press, whether the first ('1' number-pad key) or second ('2' number-pad key) interval contained the more leftward stimulus. Each trial consisted of a central reference stimulus, and a target stimulus presented at one of eight locations along the horizontal azimuth on the display. The presentation order of the reference and target stimuli was

randomised across trials. Stimulus modality was auditory, visual, or audiovisual, presented in separate blocks with short breaks (~2 min) between conditions (see **Figure 6A** for an example trial). The order of conditions was counterbalanced across participants. Each condition consisted of 384 target presentations across the eight origin locations, leading to 48 presentations at each location.

EEG Session

In this session, the experimental task was changed slightly from the behavioural session to increase the number stimulus presentations required for inverted encoding analyses of EEG data. Participants viewed and/or listened to a sequence of 20 stimuli, each of which were presented at one of five horizontal locations along the display (selected at random). At the end of each sequence, participants were tasked with indicating, via button press, whether more presentations appeared on the right ('right' arrow key) or the left ('left' arrow key) of the display. To minimize eye movements, participants were asked to fixate on a black dot presented 8° above the display centre (see **Figure 6B** for an example trial). The task used in the EEG session included the same blocked conditions as in the behavioural session, i.e., visual, auditory, and (congruent) audiovisual stimuli. As the locations of stimuli were selected at random, some sequences had an equal number of presentations on each side of the display, and thus had no correct "left" or "right" response; these trials were not included in the analysis of behavioural performance. Each block consisted of 10 trials, followed by a feedback display indicating the number of trials participants answered correctly. Each condition consisted of 12 blocks, yielding a total of 2400 presentations for each.

EEG data pre-processing

EEG data were recorded using a 64-channel BioSemi system at a sampling rate of 1024 Hz, which was down-sampled to 512 Hz during preprocessing. Signals were recorded with reference to the CMS/DRL electrode loop, with bipolar electrodes placed above and below the eye, at the temples, and at each mastoid to monitor for eye movements and muscle artifacts. EEG preprocessing was undertaken in MATLAB using custom scripts and the EEGLAB toolbox (Delorme & Makeig, 2004). Data were high-pass filtered at 0.25 Hz to remove baseline drifts, and re-referenced according to the average of all 64 channels. Analyses were stimulus locked, with ERP responses segmented into 600 ms epochs from 100 ms before stimulus presentation to 500 ms after stimulus presentation. We removed trials with substantial eye movements (>3.75 away from fixation) from the analyses. After the removal of eye movements, on average 2365 (*SD* = 56.94), 2346 (*SD* = 152.87) and 2350 (*SD* = 132.47) trials remained for auditory, visual and audiovisual conditions, respectively, from the original 2400 per condition.

Forward model

To describe the neural representations of sensory stimuli, we used an inverted modelling approach to reconstruct the location of stimuli based upon recorded ERPs (Brouwer & Heeger, 2011; Harrison et al., 2023). Analyses were performed separately for visual, auditory, and audiovisual stimuli. We first created an encoding model that characterised the patterns of activity in the EEG electrodes given the five locations of the presented stimuli. The encoding model was then used to obtain the inverse decoding model that described the transformation from electrode activity to stimulus location. We used a 10-fold cross-validation approach where 90% of the data were used to obtain the inverse model on which the remaining 10% of the data were decoded. Cross-validation was repeated 10 times such that all the data were decoded. For the purposes of these analyses, we assume that EEG electrode noise is isotropic across locations and additive with the signal.

Prior to the neural decoding analyses, we established the sensors that contained the most location information by treating time as the decoding dimension and obtaining the inverse models from each electrode, using 10-fold cross-validation. This revealed that location was primarily

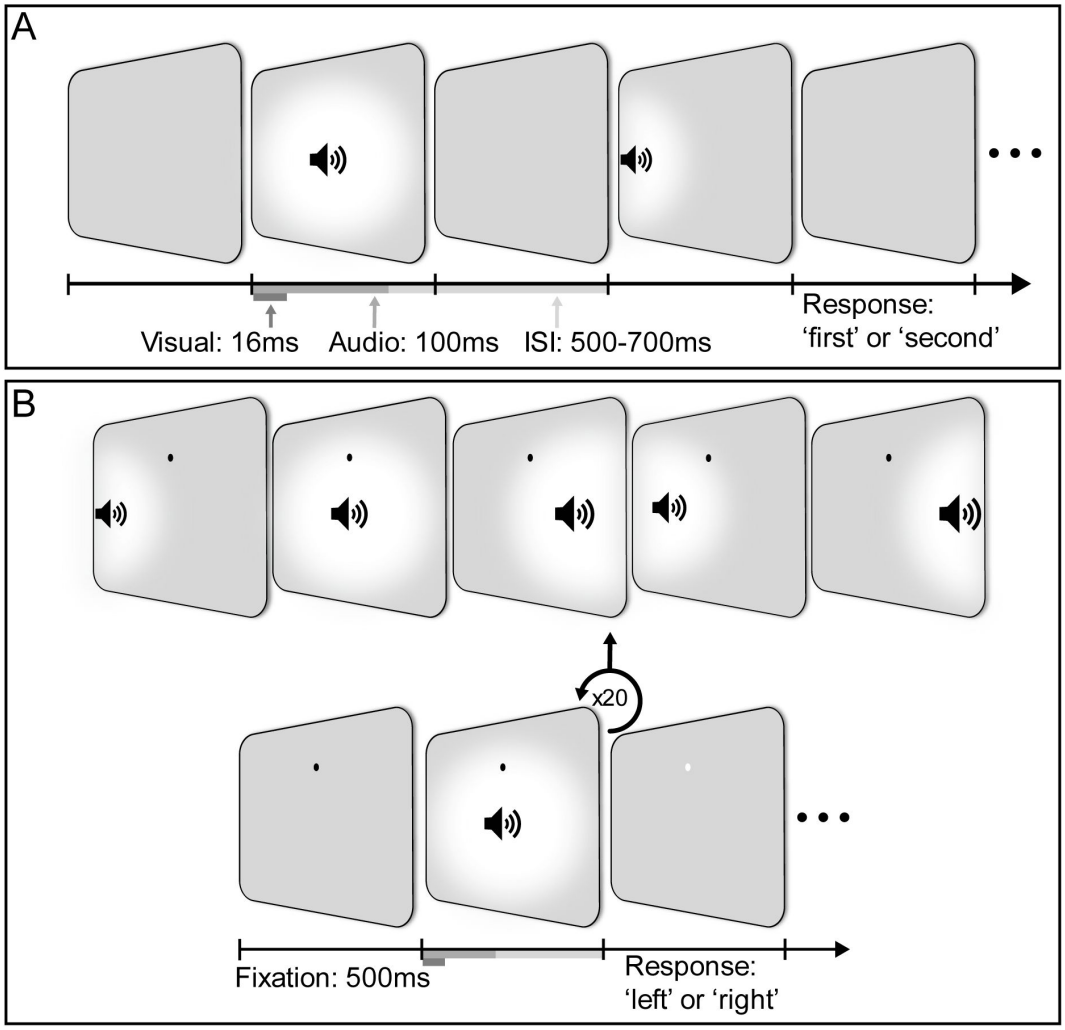


Figure 6.

Experimental design of behavioural and EEG sessions.

A) An example trial for the audiovisual condition in the behavioural session. Each trial consisted of a (centred) reference stimulus and a target stimulus presented at one of eight locations along the horizontal meridian of the display. **B)** An example trial for the audiovisual condition in the EEG session. The top row displays the possible locations of stimuli. In each trial, participants were presented with 20 stimuli that were each spatially localised to one of five possible locations along the horizontal meridian. The task was to determine if there were more stimuli presented to the left or right of fixation.

represented in posterior electrodes for visual and audiovisual stimuli, and in central electrodes for auditory stimuli. Thus, for all subsequent analyses we only included signals from the central-temporal, parietal-occipital, occipital and inion sensors for computing the inverse model (see final inset of **Figure 2**). As the model was fitted for multiple electrodes, subtle patterns of neural information contained within combinations of sensors could be detected.

The encoding model contained five hypothetical channels, with evenly distributed idealised location preferences between -15° to +15° viewing angles to the left and right of the display centre. Each channel consisted of a half-wave rectified sinusoid raised to the fifth power. The channels were arranged such that an idealised tuning curve of each location preference could be expressed as a weighted sum of the five channels. The observed activity for each presentation can be described by the following linear model:

$$\mathbf{B} = \mathbf{WC} + \mathbf{E}$$

where **B** indicates the EEG data (*m* electrodes x *n* presentations), **W** is a weight matrix (*m* electrodes x 5 channels) that describes the transformation from EEG activity to stimulus location, **C** denotes the hypothesized channel activities (5 channels x *n* presentations), and **E** indicates the residual errors.

To compute the inverse model we estimated the weights that, when applied to the data, would reconstruct the channel activities with the least error. Due to the correlation between neighbouring electrodes, we took noise covariance into account when computing the model to optimize it for EEG data (Harrison et al., 2023; Kok et al., 2017; Rideaux et al., 2023). We then used the inverse model to reconstruct the stimulus location from the recorded ERP responses.

To assess how well the forward model captured location information in the neural signal per modality, two measures of performance were analysed. First, decoding accuracy was calculated as the similarity of the decoded location to the presented location, represented in arbitrary units. To test whether a super-additive interaction was present in the multivariate response, an additive threshold against which to compare the audiovisual response was required. However, it is unclear how the arbitrary units used to represent decoding accuracy translate to a measure of the linear summation of auditory and visual accuracy. As used for the behavioural analyses, MLE provides a framework for calculating the estimated optimal sensitivity of the combination of two sensory signals, according to signal detection theory principles. To compute decoding sensitivity (d'), required to apply MLE, we omitted trials where stimuli appeared in the centre of the display. The decoder's reconstructions of stimulus location were grouped for stimuli appearing on the left and right side of the display, respectively. The proportion of hits and misses was derived by comparing the decoded side to the presented side, which was then used to calculate d' for each condition (Stanislaw & Todorov, 1999). The d' from the auditory and visual conditions can be used to estimate the predicted 'optimal' sensitivity to audiovisual signals as calculated through the following formula:

$$d'AV = \sqrt{(d'A)^2 + (d'V)^2}$$

We can then compare actual audiovisual sensitivity to this auditory + visual sensitivity and test for super-additivity in the audiovisual condition as evidenced by the presence of a nonlinear combination of auditory and visual stimuli. A similar method was previously employed to investigate depth estimation from motion and binocular disparity cues, decoded from BOLD responses (Ban et al., 2012).

To represent an additional 'additive' multivariate signal with which to compare the decoding sensitivity derived through MLE, we first matched the EEG data between unisensory conditions such that the order of presented stimulus locations was the same for the auditory and visual

conditions. The auditory and visual condition data were then concatenated across sensors, and inverted encoding analyses were performed on the resulting 'additive' audiovisual dataset. This additive condition was designed to represent neural activity evoked by both the auditory and visual conditions, without any non-linear neural interaction, and served as a baseline for the audiovisual condition.

Statistical analyses

Statistical analyses were performed in MATLAB v2021b. Two metrics of accuracy were calculated to assess behavioural performance. For the behavioural session we calculated participants' sensitivity separately for each modality condition by fitting psychometric functions to the proportion of rightward responses per stimulus location. In the EEG session participants responded to multiple stimuli rather than individual presentations, so behavioural performance was assessed via d'. We derived d' in each condition from the average proportion of hits and misses for each participant's performance in discriminating the side of the display on which more stimuli were presented (Stanislaw & Todorov, 1999). A one-sample Kolmogorov-Smirnov test for each condition revealed all conditions in both sessions violated assumptions of normality. A non-parametric two-sided Wilcoxon signed-rank test was therefore used to test for significant differences in behavioural accuracy between all conditions.

For the neural data, univariate ERPs were calculated by averaging EEG activity across presentations and channels for each stimulus location from -100 to 500 ms around stimulus onset. A conservative mass-based cluster correction was applied to account for spurious differences across time (Pernet et al., 2015). To test for significant differences between conditions, a paired-samples *t*-test was conducted between each condition at each time point. A one-sample *t*-test was used when comparing decoding accuracy against chance (i.e., zero). Next, the summed value of computed *t* statistics associated with each comparison (separately for positive and negative values) was calculated within contiguous temporal clusters of significant values. We then simulated the null distribution of the maximum summed cluster values using permutation (n = 5000) of the location labels, from which we derived the 95% percentile threshold value. Clusters identified in the data with a summed effect-size value less than the threshold were considered spurious and removed.

Data availability

The behavioural and EEG data, and the scripts used for analysis and figure creation, are available at *https://doi.org/10.17605/OSF.IO/8CDRA*.

Acknowledgements

We thank R. West for data collection, and D. Lloyd for technical assistance. This work was supported by Australian Research Council (ARC) Discovery Early Career Researcher Awards awarded to RR (DE210100790) and AKR (DE200101159). RR was also supported by a National Health and Medical Research Council (Australia) Investigator Grant (2026318).

References

Alais D., Burr D. (2004) **The ventriloquist effect results from near-optimal bimodal integration** *Current Biology* **14**:257–262 https://doi.org/10.1016/j.cub.2004.01.029

Arieh Y., Marks L. E. (2008) **Cross-modal interaction between vision and hearing: a speed-accuracy analysis. Attention, Perception** *& Psychophysics* **70**:412–421 https://doi.org/10.3758/pp.70.3.412

Ban H., Preston T. J., Meeson A., Welchman A. E. (2012) **The integration of motion and disparity cues to depth in dorsal visual cortex** *Nature Neuroscience* **15**:636–643 https://doi.org/10.1038/nn.3046

Beauchamp M. S. (2005) **Statistical criteria in fMRI studies of multisensory integration** *Neuroinformatics* **3**:93–113 https://doi.org/10.1385/NI:3:2:093

Bednar A., Lalor E. C. (2020) **Where is the cocktail party? Decoding locations of attended and unattended moving sound sources using EEG** *Neuroimage* **205** https://doi.org/10.1016/j.neuroimage.2019.116283

Besle J., Bertrand O., Giard M. H. (2009) **Electrophysiological (EEG, sEEG, MEG) evidence for multiple audiovisual interactions in the human auditory cortex** . *Hearing Research* **258**:143–151 https://doi.org/10.1016/j.heares.2009.06.016

Besle J., Fort A., Giard M. H. (2004) **Interest and validity of the additive model in electrophysiological studies of multisensory interactions** *Cognitive Processing* **5**:189–192 https://doi.org/10.1007/s10339-004-0026-y

Bolognini N., Frassinetti F., Serino A., Ladavas E. (2005) **"Acoustical vision" of below threshold stimuli: interaction among spatially converging audiovisual inputs** *Experimental Brain Research* **160**:273–282 https://doi.org/10.1007/s00221-004-2005-z

Brainard D. H. (1997) **The Psychophysics Toolbox** *Spatial Vision* **10**:433–436 https://doi.org/10.1163/156856897X00357

Brouwer G. J., Heeger D. J. (2009) **Decoding and Reconstructing Color from Responses in Human Visual Cortex** *The Journal of Neuroscience* **29** https://doi.org/10.1523/JNEUROSCI.3577-09.2009

Brouwer G. J., Heeger D. J. (2011) **Cross-orientation suppression in human visual cortex** *Journal of Neurophysiology* **106**:2108–2119 https://doi.org/10.1152/jn.00540.2011

Calvert G. A., Campbell R., Brammer M. J. (2000) **Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex** *Current Biology* **10**:649–657 https://doi.org/10.1016/S0960-9822(00)00513-3

Calvert G. A., Hansen P. C., Iversen S. D., Brammer M. J. (2001) **Detection of Audio-Visual Integration Sites in Humans by Application of Electrophysiological Criteria to the BOLD Effect** *Neuroimage* **14**:427–438 https://doi.org/10.1006/nimg.2001.0812

Calvert G. A., Thesen T. (2004) **Multisensory integration: methodological approaches and emerging principles in the human brain** *Journal of Physiology-Paris* **98**:191–205 https://doi.org/10.1016/j.jphysparis.2004.03.018

Cappe C., Thut G., Romei V., Murray M. M. (2009) **Selective integration of auditory-visual looming cues by humans** *Neuropsychologia* **47**:1045–1052 https://doi.org/10.1016/j.neuropsychologia.2008.11.003

Cappe C., Thut G., Romei V., Murray M. M. (2010) **Auditory–Visual Multisensory Interactions in Humans: Timing, Topography, Directionality, and Sources** *The Journal of Neuroscience* **30** https://doi.org/10.1523/JNEUROSCI.1099-10.2010

Colonius H., Diederich A. (2004) **Multisensory Interaction in Saccadic Reaction Time: A Time-Window-of-Integration Model** *Journal of Cognitive Neuroscience* **16**:1000–1009 https://doi.org/10.1162/0898929041502733

Corniel B. D., van Wanrooij M. M., Munoz D. P., van Opstal J. (2002) **Auditory-Visual Interactions Subserving Goal-Directed Saccades in a Complex Scene** *Journal of Neurophysiology* **88**:438–454 https://doi.org/10.1152/jn.00699.2001

De Meo R., Murray M. M., Clarke S., Matusz P. J. (2015) **Top-down control and early multisensory processes: chicken vs. egg** *Frontiers in integrative neuroscience* **9** https://doi.org/10.3389/fnint.2015.00017

Delong P., Noppeney U. (2021) **Semantic and spatial congruency mould audiovisual integration depending on perceptual awareness** *Scientific Reports* **11** https://doi.org/10.1038/s41598-021-90183-w

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

Fetsch C. R., DeAngelis G. C., Angelaki D. E. (2013) **Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons** *Nature Reviews Neuroscience* **14**:429–442 https://doi.org/10.1038/nrn3503

Fort A., Delpuech C., Pernier J., Giard M.-H. (2002) **Dynamics of Cortico-subcortical Cross-modal Operations Involved in Audio-visual Object Detection in Humans** *Cerebral Cortex* **12**:1031–1039 https://doi.org/10.1093/cercor/12.10.1031

Frassinetti F., Bolognini N., Ladavas E. (2002) **Enhancement of visual perception by crossmodal visuo-auditory interaction** *Experimental Brain Research* **147**:332–343 https://doi.org/10.1007/s00221-002-1262-y

Frens M. A., Van Opstal A. J. (1998) **Visual-auditory interactions modulate saccade-related activity in monkey superior colliculus** *Brain Research Bulletin* **46**:211–224 https://doi.org/10.1016/S0361-9230(98)00007-0

Giard M. H., Peronnet F. (1999) **Auditory-Visual Integration during Multimodal Object Recognition in Humans: A Behavioral and Electrophysiological Study** *Journal of Cognitive Neuroscience* **11**:473–490 https://doi.org/10.1162/089892999563544

Harrison W. J., Bays P. M., Rideaux R. (2023) **Neural tuning instantiates prior expectations in the human visual system** *Nature Communications* **14** https://doi.org/10.1038/s41467-023-41027-w

James T. W., Stevenson R. A., Kim S., Stein B. E. (2012) **Inverse Effectiveness and BOLD fMRI** *The New Handbook of Multisensory Processing* https://doi.org/10.7551/mitpress/8466.003.0020

Joassin F., Maurage P., Campanella S. (2011) **The neural network sustaining the crossmodal processing of human gender from faces and voices: An fMRI study** *Neuroimage* **54**:1654–1661 https://doi.org/10.1016/j.neuroimage.2010.08.073

Keil J., Senkowski D. (2018) **Neural Oscillations Orchestrate Multisensory Processing** *The Neuroscientist* **24**:609–626 https://doi.org/10.1177/1073858418755352

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

Laurienti P. J., Perrault T. J., Stanford T. R., Wallace M. T., Stein B. E. (2005) **On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies** *Experimental Brain Research* **166**:289–297 https://doi.org/10.1007/s00221-005-2370-2

Leone L. M., McCourt M. E. (2015) **Dissociation of perception and action in audiovisual multisensory integration** *European Journal of Neuroscience* **42**:2915–2922 https://doi.org/10.1111/ejn.13087

Lovelace C. T., Stein B. E., Wallace M. T. (2003) **An irrelevant light enhances auditory detection in humans: a psychophysical analysis of multisensory integration in stimulus detection** *Cognitive Brain Research* **17**:447–453 https://doi.org/10.1016/s0926-6410(03)00160-5

Meredith M. A., Stein B. E. (1983) **Interactions Among Converging Sensory Inputs in the Superior Colliculus** *Science* **221**:389–391 https://doi.org/10.1126/science.6867718

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** *Cognitive Brain Research* **14**:115–128 https://doi.org/10.1016/S0926-6410(02)00066-6

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

Nikbakht N., Tafreshiha A., Zoccolan D., Diamond M. E. (2018) **Supralinear and Supramodal Integration of Visual and Tactile Signals in Rats: Psychophysics and Neuronal Mechanisms** *Neuron* **97**:626–639 https://doi.org/10.1016/j.neuron.2018.01.003

Ohshiro T., Angelaki D. E., DeAngelis G. C. (2017) **A Neural Signature of Divisive Normalization at the Level of Multisensory Integration in Primate Cortex** *Neuron* **95**:399–411 https://doi.org/10.1016/j.neuron.2017.06.043

Pernet C. R., Latinus M., Nichols T. E., Rousselet G. A. (2015) **Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study** *Journal of Neuroscience Methods* **250**:85–93 https://doi.org/10.1016/j.jneumeth.2014.08.003

Porada D. K., Regenbogen C., Freiherr J., Seubert J., Lundström J. N. (2021) **Trimodal processing of complex stimuli in inferior parietal cortex is modality-independent** *Cortex* **139**:198–210 https://doi.org/10.1016/j.cortex.2021.03.008

Puce A., Epling J. A., Thompson J. C., Carrick O. K. (2007) **Neural responses elicited to face motion and vocalization pairings** *Neuropsychologia* **45**:93–106 https://doi.org/10.1016/j.neuropsychologia.2006.04.017

Rach S., Diederich A. (2006) **Visual-tactile integration: does stimulus duration influence the relative amount of response enhancement?** *Experimental Brain Research* **173**:514–520 https://doi.org/10.1007/s00221-006-0452-4

Rach S., Diederich A., Colonius H. (2010) **On quantifying multisensory interaction effects in reaction time and detection rate** *Psychological Research* **75**:77–94 https://doi.org/10.1007/s00426-010-0289-0

Rideaux R. (2024) **Task-related modulation of event-related potentials does not reflect changes to sensory representations** *bioRxiv* https://doi.org/10.1101/2024.01.20.576485

Rideaux R., Storrs K. R., Maiello G., Welchman A. E. (2021) **How multisensory neurons solve causal inference** *Proceedings of the National Academy of Sciences* **118** https://doi.org/10.1073/pnas.2106235118

Rideaux R., Welchman A. E. (2018) **Proscription supports robust perceptual integration by suppression in human visual cortex** *Nature Communications* **9** https://doi.org/10.1038/s41467-018-03400-y

Rideaux R., West R. K., Rangelov D., Mattingley J. B. (2023) **Distinct early and late neural mechanisms regulate feature-specific sensory adaptation in the human visual system** *Proceedings of the National Academy of Sciences* **120** https://doi.org/10.1073/pnas.2216192120

Robinson A. K., Grootswagers T., Shatek S. M., Gerboni J., Holcombe A., Carlson T. A. (2021) **Overlapping neural representations for the position of visible and imagined objects** *Neurons, behavior, data analysis, and theory* **4** https://doi.org/10.51628/001c.19129

Rohe T., Noppeney U. (2018) **Reliability-Weighted Integration of Audiovisual Signals Can Be Modulated by Top-down Attention** *eneuro* **5** https://doi.org/10.1523/ENEURO.0315-17.2018

Ross L. A., Molholm S., Butler J. S., Bene V. A. D., Foxe J. J. (2022) **Neural correlates of multisensory enhancement in audiovisual narrative speech perception: A fMRI investigation** *Neuroimage* **263** https://doi.org/10.1016/j.neuroimage.2022.119598

Scheliga S., Kellermann T., Lampert A., Rolke R., Spehr M., Habel U. (2023) **Neural correlates of multisensory integration in the human brain: an ALE meta-analysis** *Reviews in the Neurosciences* **34**:223–245 https://doi.org/10.1515/revneuro-2022-0065

Senkowski D., Saint-Amour D., Hofle M., Foxe J. J. (2011) **Multisensory interactions in early evoked brain activity follow the principle of inverse effectiveness** *Neuroimage* **56**:2200–2208 https://doi.org/10.1016/j.neuroimage.2011.03.075

Stanislaw H., Todorov N. (1999) **Calculation of signal detection theory measures** *Behavior Research Methods, Instruments, & Computers* **31**:137–149 https://doi.org/10.3758/BF03207704

Stein B. E., Stanford T. R. (2008) **Multisensory integration: current issues from the perspective of the single neuron** *Nature Reviews Neuroscience* **9**:255–266 https://doi.org/10.1038/nrn2331

Stekelenburg J. J., Vroomen J. (2007) **Neural Correlates of Multisensory Integration of Ecologically Valid Audiovisual Events** *Journal of Cognitive Neuroscience* **19**:1964–1973 https://doi.org/10.1162/jocn.2007.19.12.1964

Stevenson R. A., Geoghegan M. L., James T. W. (2007) **Superadditive BOLD activation in superior temporal sulcus with threshold non-speech objects** *Experimental Brain Research* **179**:85–95 https://doi.org/10.1007/s00221-006-0770-6

Stevenson R. A., Ghose D., Fister J. K., Sarko D. K., Altieri N. A., Nidiffer A. R., Kurela L. R., Siemann J. K., James T. W., Wallace M. T. (2014) **Identifying and quantifying multisensory integration: a tutorial review** *Brain Topography* **27**:707–730 https://doi.org/10.1007/s10548-014-0365-7

Stevenson R. A., James T. W. (2009) **Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition** *Neuroimage* **44**:1210–1223 https://doi.org/10.1016/j.neuroimage.2008.09.034

Sumby W. H., Pollack I. (1954) **Visual Contribution to Speech Intelligibility in Noise** *The Journal of the Acoustical Society of America* **26**:212–215 https://doi.org/10.1121/1.1907309

Talsma D., Doty T. J., Woldorff M. G. (2007) **Selective Attention and Audiovisual Integration: Is Attending to Both Modalities a Prerequisite for Early Integration?** *Cerebral Cortex* **17**:679–690 https://doi.org/10.1093/cercor/bhk016

Teder-Sälejärvi W. A., McDonald J. J., Di Russo F., Hillyard S. A. (2002) **An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings** *Cognitive Brain Research* **14**:106–114 https://doi.org/10.1016/S0926-6410(02)00065-4

The MathWorks Inc (2021) **MATLAB version: 9.11.0 (R2021b), Natick, Massachusetts** *The MathWorks Inc*

Van Wanrooij M. M., Bell A. H., Munoz D. P., Van Opstal A. J. (2009) **The effect of spatial-temporal audiovisual disparities on saccades in a complex scene** *Experimental Brain Research* **198**:425–437 https://doi.org/10.1007/s00221-009-1815-4

Venezia J. H., Matchin W., Hickok G., Toga A. W. (2015) **Multisensory Integration and Audiovisual Speech Perception** *Brain Mapping* Academic Press :565–572 https://doi.org/10.1016/B978-0-12-397025-1.00047-6

Vroomen J., Stekelenburg J. J. (2010) **Visual Anticipatory Information Modulates Multisensory Interactions of Artificial Audiovisual Stimuli** *Journal of Cognitive Neuroscience* **22**:1583–1596 https://doi.org/10.1162/jocn.2009.21308

Wang Y., Celebrini S., Trotter Y., Barone P. (2008) **Visuo-auditory interactions in the primary visual cortex of the behaving monkey: Electrophysiological evidence** *BMC neuroscience* **9** https://doi.org/10.1186/1471-2202-9-79

Werner S., Noppeney U. (2010) **Superadditive Responses in Superior Temporal Sulcus Predict Audiovisual Benefits in Object Categorization** *Cerebral Cortex* **20**:1829–1842 https://doi.org/10.1093/cercor/bhp248

Werner S., Noppeney U. (2011) **The Contributions of Transient and Sustained Response Codes to Audiovisual Integration** *Cerebral Cortex* **21**:920–931 https://doi.org/10.1093/cercor/bhq161

Whitworth R. H., Jeffress L. A. (1961) **Time vs Intensity in the Localization of Tones** *The Journal of the Acoustical Society of America* **33**:925–929 https://doi.org/10.1121/1.1908849

Wightman F. L., Kistler D. J. (1992) **The dominant role of low-frequency interaural time differences in sound localization** *The Journal of the Acoustical Society of America* **91**:1648–1661 https://doi.org/10.1121/1.402445

Wilcox R.R. (2016) **Comparing dependent robust correlations** *British Journal of Mathematical & Statistical Psychology* **69**:215–224 https://doi.org/10.1111/bmsp.12069

Author information

Zak Buhmann
Queensland Brain Institute, The University of Queensland, Brisbane, Australia
ORCID iD: 0009-0002-4249-462X

For correspondence: z.buhmann@uq.net.au

Amanda K Robinson
Queensland Brain Institute, The University of Queensland, Brisbane, Australia
ORCID iD: 0000-0002-7378-2803

Jason B Mattingley
Queensland Brain Institute, The University of Queensland, Brisbane, Australia, School of Psychology, The University of Queensland, Brisbane, Australia
ORCID iD: 0000-0003-0929-9216

Reuben Rideaux
Queensland Brain Institute, The University of Queensland, Brisbane, Australia, School of Psychology, University of Sydney, Sydney, Australia
ORCID iD: 0000-0001-8416-005X

Editors

Reviewing Editor
Maria Chait
University College London, London, United Kingdom

Senior Editor
Barbara Shinn-Cunningham
Carnegie Mellon University, Pittsburgh, United States of America

Reviewer #1 (Public review):

This study presents a novel application of inverted encoding (i.e., decoding) to detect non-linear correlates of crossmodal integration in human neural activity, using EEG

(electroencephalography). The method is successfully applied to data from a group of 41 participants, performing a spatial localization task on auditory, visual, and audio-visual events. The analyses clearly show a behavioural superiority for audio-visual localization. Like previous studies, the results when using traditional univariate ERP analyses were inconclusive, showing once more the need for alternative, more sophisticated approaches. The inverted encoding approach of this study, harnessing on the multivariate nature of the signal, captured clear signs of super-additive responses, considered by many as the hallmark of multisensory integration. Despite the removal of eye-movement artefacts from the signal eliminated the significant decoding effect, the author's control analyses showed that decoding is more effective from parietal, compared to frontal electrodes, thereby ruling out ocular contamination as the sole origin of the relevant signal.

This significant finding can bear important advances in the many fields where multisensory integration has been shown to play an important role, by providing a way to bring much needed coherence across levels of analysis, from behaviour to single-cell electrophysiology. To achieve this, it would be ideal to contrast whether the pattern of super-additive effects in other scenarios where clear behavioural signs of multisensory integration are also observed. One could also try to further support the posterior origin of the super-additive effects by source localization.

Comments on revised version:

All my previous concerns have been addressed. I congratulate the authors on a very nice paper.

https://doi.org/10.7554/eLife.97230.2.sa2

Reviewer #2 (Public review):

Summary:

This manuscript seeks to reconcile observations in multisensory perception - from behavior and from neural responses. It is intuitively obvious that perceiving a stimulus via two senses results in better performance than one alone. However, the nature of this interaction is complicated and relating different measures (behavioural, neural) is challenging.

It is not uncommon to observe that for a perceptual task the percentage of correct responses seen with two senses is higher than the sum of the percentage correct obtained with each modality individually. i.e. the gains are "superadditive". The gains of adding a second sense are typically larger when the performance with the first sense is relatively poor - this effect is often called the principle inverse effectiveness. More generally, what this tells us is that performance in a multisensory perceptual task is a non-linear sum of performance for each sensory modality alone. In invasive recordings from single neurons, a wide range of non-linear interactions is observed - some superadditive, and some sub-additive.

Despite this abundance evidence of non-linearity in some measures of multisensory integration, evoked responses (EEG) to such sensory stimuli often show little evidence of it - and this is the problem this manuscript tackles. The key assertion made is that a univariate analysis of the EEG signal is likely to average out non-linear effects of integration. This is a reasonable assertion, and their analysis does indeed provide evidence that a multivariate approach can reveal non-linear interactions in the evoked responses.

Strengths:

It is of great value to understand how the process of multisensory integration occurs, and despite a wealth of observations of the benefits of perceiving the world with multiple senses,

we still lack a reasonable understanding of how the brain integrates information. For example - what underlies the large individual differences in the benefits of two senses over one? One way to tackle this is via brain imaging, but this is problematic if important features of the processing - such as non-linear interactions are obscured by the lack of specificity of the measurements. The approach they take to analysis of the EEG data allows the authors to look in more detail at the variation in activity across EEG electrodes, which averaging across electrodes cannot.

This version of the manuscript is well written and for the most part clear and the report of non-linear summation of neural responses is convincing. A particular strength of the paper is their use of a statistical model of multisensory integration as their "null" model of neural responses, and the "inverted-encoder" which infers an internal representation of the stimulus which can explain the EEG responses. This encoder generates a prediction of decoding performance, which can be used to generate predictions of multisensory decoding from unisensory decoding, or from a sum of the unisensory internal representations.

In behavioural performance, it is frequently observed that the performance increase from two senses is close to what is expected from the optimal integration of information across the senses, in a statistical sense. It can be plausibly explained by assuming that people are able to weight sensory inputs according to their reliability - and somewhat optimally. Critically the apparent "superadditive" effect on performance described above does not require any non-linearity in the sum of information across the senses, but can arise from correctly weighting the information according to reliability.

The authors apply a similar model to predict the neural responses expected to audiovisual stimuli from the neural responses to audio and visual stimuli alone, assuming optimal statistical integration of information. The neural responses to audiovisual stimuli exceed the predictions of this model and this is the main evidence supporting their conclusion, and it is convincing.

Weaknesses:

The main weakness of the manuscript is that their behavioural data show no evidence of performance that exceeds the predictions of these statistical models. In fact, the models predict multisensory performance from unisensory performance pretty well. This makes it hard to interpret their results, as surely if these nonlinear neural interactions underlie the behaviour, then we should be able to see evidence of it in the behaviour. I cannot offer an easy explanation for this.

Overall, therefore, I applaud the motivation and the sophistication of the analysis method and think it shows great promise for tackling these problems.

https://doi.org/10.7554/eLife.97230.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***Reviewer 1:***

We thank Reviewer 1 for their helpful comments and hope that the changes made to the revised manuscript have addressed their points.

> *This study presents a novel application of the inverted encoding (i.e., decoding) approach to detect the correlates of crossmodal integration in the human EEG*

> *(electrophysiological) signal. The method is successfully applied to data from a group of 41 participants, performing a spatial localization task on auditory, visual, and audiovisual events. The analyses clearly show a behavioural superiority for audio-visual localization. Like previous studies, the results when using traditional univariate ERP analyses were inconclusive, showing once more the need for alternative, more sophisticated approaches. Instead, the principal approach of this study, harnessing the multivariate nature of the signal, captured clear signs of super-additive responses, considered by many as the hallmark of multisensory integration. Unfortunately, the manuscript lacks many important details in the descriptions of the methodology and analytical pipeline. Although some of these details can eventually be retrieved from the scripts that accompany this paper, the main text should be self-contained and sufficient to gain a clear understanding of what was done. (A list of some of these is included in the comments to the authors). Nevertheless, I believe the main weakness of this work is that the positive results obtained and reported in the results section are conditioned upon eye movements. When artifacts due to eye movements are removed, then the outcomes are no longer significant.*
>
> *Therefore, whether the authors finally achieved the aims and showed that this method of analysis is truly a reliable way to assess crossmodal integration, does not stand on firm ground. The worst-case scenario is that the results are entirely accounted for by patterns of eye movements in the different conditions. In the best-case scenario, the method might truly work, but further experiments (and/or analyses) would be required to confirm the claims in a conclusive fashion.*
>
> *One first step toward this goal would be, perhaps, to facilitate the understanding of results in context by reporting both the uncorrected and corrected analyses in the main results section. Second, one could try to support the argument given in the discussion, pointing out the origin of the super-additive effects in posterior electrode sites, by also modelling frontal electrode clusters and showing they aren't informative as to the effect of interest.*

We performed several additional analyses to address concerns that our main result was caused by different eye movement patterns between conditions. We re-ran our key analyses using activity exclusively from frontal electrodes, which revealed poorer decoding performance than that from posterior electrodes. If eye movements were driving the non-linear enhancement in the audiovisual condition, we would expect stronger decoding using sensors closer to the source, i.e., the extraocular muscles. We also computed the correlations between average eye position and stimulus position for each condition to evaluate whether participants made larger eye movements in the audiovisual condition, which might have contributed to better decoding results. Though we did find evidence for eye movements toward stimuli, the degree of movement did not significantly differ between conditions.

Furthermore, we note that the analysis using a stricter eye movement criterion, acknowledged in the Discussion section of the original manuscript, resulted in very similar results to the original analysis. There was significantly better decoding in the AV condition (as measured by d') than the MLE prediction, but this difference did not survive cluster correction. The most likely explanation for this is that the strict eye movement criterion combined with our conservative measure of (mass-based) cluster correction led to reduced power to detect true differences between conditions. Taken together with the additional analyses described in the revised manuscript and supplementary materials, the results show that eye movements are unlikely to account for differences between the multisensory and unisensory conditions. Instead, our decoding results likely reflect nonlinear neural integration between audio and visual sensory information.

"Any experimental design that varies stimulus location needs to consider the potential contribution of eye movements. We computed correlations between participants' average eye position and each stimulus position between the three sensory conditions (auditory, visual and audiovisual; Figure S1) and found evidence that participants made eye movements toward stimuli. A re-analysis of the data with a very strict eye-movement criterion (i.e., removing trials with eye movements >1.875º) revealed that the super-additive enhancement in decoding accuracy no longer survived cluster correction, suggesting that our results may be impacted by the consistent motor activity of saccades towards presented stimuli. Further investigation, however, suggests this is unlikely. Though the correlations were significantly different from 0, they were not significantly different from each other. If consistent saccades to audiovisual stimuli were responsible for the nonlinear multisensory benefit we observed, we would expect to find a higher positive correlation between horizontal eye position and stimulus location in the audiovisual condition than in the auditory or visual conditions. Interestingly, eye movements corresponded more to stimulus location in the auditory and audiovisual conditions than in the visual condition, indicating that it was the presence of a sound, rather than a visual stimulus, that drove small eye movements. This could indicate that participants inadvertently moved their eyes when localising the origin of sounds. We also re-ran our analyses using the activity measured from the frontal electrodes alone (Figure S2). If the source of the nonlinear decoding accuracy in the audiovisual condition was due to muscular activity produced by eye movements, there should be better decoding accuracy from sensors closer to the source. Instead, we found that decoding accuracy of stimulus location from the frontal electrodes (peak d' = 0.08) was less than half that of decoding accuracy from the more posterior electrodes (peak d' = 0.18). These results suggest that the source of neural activity containing information about stimulus position was located over occipito-parietal areas, consistent with our topographical analyses (inset of Figure 3)."

> *The univariate ERP analyses an outdated contrast, AV <> A + V to capture multisensory integration. A number of authors have pointed out the potential problem of double baseline subtraction when using this contrast, and have recommended a number of solutions, experimental and analytical. See for example: [1] and [2].*
>
> *(1) Teder-Salejarvi, W. A., McDonald, J. J., Di Russo, F., & Hillyard, S. A. (2002). Cognitive Brain Research, 14, 106-114.*
>
> *(2) Talsma, D., & Woldorff, M. G. (2005). Journal of cognitive neuroscience, 17(7), 1098-1114.*

We thank the reviewer for raising this point. Comparing ERPs across different sensory conditions requires careful analytic choices to discern genuine sensory interactions within the signal. The AV <> (A +V) contrast has often been used to detect multisensory integration, though any non-signal related activity (i.e. anticipatory waves; Taslma & Woldorff, 2005) or pre-processing manipulation (e.g. baseline subtraction; Teder-Sälejärvi et al., 2002) will be doubled in (A + V) but not in AV. Critically, we did not apply a baseline correction during preprocessing and thus our results are not at risk of double-baseline subtraction in (A + V). Additionally, we temporally jittered the presentation of our stimuli to mitigate the potential influence of consistent overlapping ERP waves (Talsma & Woldorff, 2005).

> *The results section should provide the neurometric curve/s used to extract the slopes of the sensitivity plot (Figure 2B).*

We thank the reviewer for raising this point of clarification. The sensitivity plots for Figures 2B and 2C were extracted from the behavioural performance of the behavioural and EEG tasks, respectively. The sensitivity plot for Figure 2B was extracted from individual

psychometric curves, whereas the d' values for Figure 2C were calculated from the behavioural data for the EEG task. This information has been clarified in the manuscript.

"Figure 1. Behavioural performance is improved for audiovisual stimuli. A) Average accuracy of responses across participants in the behavioural session at each stimulus location for each stimulus condition, fitted to a psychometric curve. Steeper curves indicate greater sensitivity in identifying stimulus location. B) Average sensitivity across participants in the behavioural task, estimated from psychometric curves, for each stimulus condition. The red cross indicates estimated performance assuming optimal (MLE) integration of unisensory cues. C) Average behavioural sensitivity across participants in the EEG session for each stimulus condition. Error bars indicate ±1 SEM."

> *The encoding model was fitted for each electrode individually; I wonder if important information contained as combinations of (individually non-significant) electrodes was then lost in this process and if the authors consider that this is relevant.*

Although the encoding model was fitted for each electrode individually for the topographic maps (Figure 4B), in all other analyses the encoding model was fitted across a selection of electrodes (see final inset of Figure 3). As this electrode set was used for all other neural analyses, our model would allow for the detection of important information contained in the neural patterns across electrodes. This information has been clarified in the manuscript.

"Thus, for all subsequent analyses we only included signals from the central-temporal, parietal-occipital, occipital and inion sensors for computing the inverse model (see final inset of Figure 2). As the model was fitted for multiple electrodes, subtle patterns of neural information contained within combinations of sensors could be detected."

> *Neurobehavioral correlations could benefit from outlier rejection and the use of robust correlation statistics.*

We thank the reviewer for raising this issue. Note, however, that the correlations we report are resistant to the influence of outliers because we used Spearman's rho1 (as opposed to Pearson's). This information has been communicated in the manuscript.

(1) Wilcox, R.R. (2016), Comparing dependent robust correlations. *British Journal of Mathematical & Statistical Psychology, 69*(3), 215-224. https://doi.org/10.1111/bmsp.12069

"Neurobehavioural correlations. As behavioural and neural data violated assumptions of normality, we calculated rank-order correlations (Spearman's rho) between the average decoding sensitivity for each participant from 150-250 ms poststimulus onset and behavioural performance on the EEG task. As Spearman's rho is resistant to outliers (Wilcox, 2016), we did not perform outlier rejection."

"Wilcox, R.R. (2016), Comparing dependent robust correlations. *British Journal of Mathematical & Statistical Psychology, 69*(3), 215-224. https://doi.org/10.1111/bmsp.12069"

> *Many details that are important for the reader to evaluate the evidence and to understand the methods and analyses aren't given; this is a non-exhaustive list:*

We thank the reviewer for highlighting these missing details. We have updated the manuscript where necessary to ensure the methods and analyses are fully detailed and replicable.

> *- specific parameters of the stimuli and performance levels. Just saying "similarly difficult" or "marginally higher volume" is not enough to understand exactly what was done.*

"The perceived source location of auditory stimuli was manipulated via changes to interaural level and timing (Whitworth & Jeffress, 1961; Wightman & Kistler, 1992). The precise timing of when each speaker delivered an auditory stimulus was calculated from the following formula:

$$ITD_{L,R} = \frac{\sqrt{(x \pm r)^2 + z^2}}{s}$$

where *x* and *z* are the horizontal and forward distances in metres between the ears and the source of the sound on the display, respectively, *r* is the head radius, and *s* is the speed of sound. We used a constant approximate head radius of 8 cm for all participants. *r* was added to *x* for the left speaker and subtracted for the right speaker to produce the interaural time difference. For ±15° source locations, interaural timing difference was 1.7 ms. To simulate the decrease in sound intensity as a function of distance, we calculated interaural level differences for the left and right speakers by dividing the sounds by the left and right distance vectors. Finally, we resampled the sound using linear interpolation based on the calculations of the interaural level and timing differences. This process was used to calculate the soundwaves played by the left and right speakers for each of the possible stimulus locations on the display. The maximum interaural level difference between speakers was 0.14 *A* for ±15° auditory locations, and 0.07 *A* for ±7.5°."

> *- where are stimulus parameters adjusted individually or as a group? Which method was followed?*

To clarify, stimulus parameters (frequency, size, luminance, volume, location, etc.) were manipulated throughout pilot testing *only*. Parameters were adjusted to achieve similar pilot behavioural results between the auditory and visual conditions. For the experiment proper, parameters remained constant for both tasks and were the same for all participants.

"During pilot testing, stimulus features (size, luminance, volume, frequency etc.) were manipulated to make visual and auditory stimuli similarly difficult to spatially localize. These values were held constant in the main experiment."

> *- specify which response buttons were used.*

"Participants were presented with two consecutive stimuli and tasked with indicating, via button press, whether the first ('1' number-pad key) or second ('2' number-pad key) interval contained the more leftward stimulus."

"At the end of each sequence, participants were tasked with indicating, via button press, whether more presentations appeared on the right ('right' arrow key) or the left ('left' arrow key) of the display."

> *- no information is given as to how many trials per condition remained on average, for analysis.*

The average number of remaining trials per condition after eye-movement analysis is now included in the Methods section of the revised manuscript.

"We removed trials with substantial eye movements (>3.75 away from fixation) from the analyses. After the removal of eye movements, on average 2365 (*SD* = 56.94), 2346 (*SD* = 152.87) and 2350 (*SD* = 132.47) trials remained for auditory, visual and audiovisual conditions, respectively, from the original 2400 per condition."

> *- no information is given on the specifics of participant exclusion criteria. (even if the attrition rate was surprisingly high, for such an easy task).*

The behavioural session also served as a screening task. Although the task instructions were straightforward, perceptual discrimination was not easy due to the ambiguity of the stimuli. Auditory localization is not very precise, and the visual stimuli were brief, dim, and diffuse. The behavioural results reflect the difficulty of the task. Attrition rate was high as participants who scored below 60% correct in any condition were deemed unable to accurately perform the task, were not invited to complete the subsequent EEG session, and omitted from the analyses. We have included the specific criteria in the manuscript.

“Participants were first required to complete a behavioural session with above 60% accuracy in all conditions to qualify for the EEG session (see *Behavioural session* for details).”

> *- EEG pre-processing: what filter was used? How was artifact rejection done? (no parameters are reported); How were bad channels interpolated?*

We used a 0.25 Hz high-pass filter to remove baseline drifts, but no low-pass filter. In line with recent studies on the undesirable influence of EEG preprocessing on ERPs1, we opted to avoid channel interpolation and artifact rejection. This was erroneously reported in the manuscript and has now been clarified. For the sake of clarity, here we demonstrate that a reanalysis of data using channel interpolation and artifact rejection returned the same pattern of results.

(1) Delorme, A. (2023). EEG is better left alone. *Scientific Reports*, 13, 2372. https://doi.org/10.1038/s41598-023-27528-0

> *- specific electrode locations must be given or shown in a plot (just "primarily represented in posterior electrodes" is not sufficiently informative).*

A diagram of the electrodes used in all analyses is included within Figure 3, and we have drawn readers’ attention to this in the revised manuscript.

“Thus, for all subsequent analyses we only included signals from the central-temporal, parietal-occipital, occipital and inion sensors for computing the inverse model (see final inset of Figure 2).”

> *- ERP analysis: which channels were used? What is the specific cluster correction method?*

We used a conservative mass-based cluster correction from Pernet et al. (2015) - this information has been clarified in the manuscript.

“A conservative mass-based cluster correction was applied to account for spurious differences across time (Pernet et al., 2015).”

“Pernet, C. R., Latinus, M., Nichols, T. E., & Rousselet, G. A. (2015). Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study. *Journal of Neuroscience Methods*, *250*, 85-93. https://doi.org/https://doi.org/10.1016/j.jneumeth.2014.08.003”

> *- results: descriptive stats on performance must be given (instead of saying "participants performed well").*

The mean and standard deviation of participants’ performance for each condition in the behavioural and EEG experiments are now explicitly mentioned in the manuscript.

"A quantification of the behavioural sensitivity (i.e., steepness of the curves) revealed significantly higher sensitivity for the audiovisual stimuli (M = .04, SD = .02) than for the auditory stimuli alone (M = .03, SD = .01; Z = -3.09, p = .002), and than for the visual stimuli alone (M = .02, SD = .01; Z = -5.28, p = 1.288e-7; Figure 1B). Sensitivity for auditory stimuli was also significantly higher than sensitivity for visual stimuli (Z = 2.02, p = .044)."

"We found a similar pattern of results to those in the behavioural session; sensitivity for audiovisual stimuli (M = .85, SD = .33) was significantly higher than for auditory (M = .69, SD = .41; Z = -2.27, p = .023) and visual stimuli alone (M = .61, SD = .29; Z = -3.52, p = 4.345e-4), but not significantly different from the MLE prediction (Z = -1.07, p = .285)."

> *- sensitivity in the behavioural and EEG sessions is said to be different, but no comparison is given. It is not even the same stimulus set across the two tasks...*

This relationship was noted as a potential explanation for the higher sensitivities obtained in the EEG task, and was not intended to stand up to statistical scrutiny. We agree it makes little sense to compare statistically between the EEG and behavioural results as they were obtained from different tasks. We would like to clarify, however, that the stimuli used in the two tasks were the same, with the exception that in the EEG task the stimuli were presented from 5 locations versus 8 in the behavioural task. To avoid potential confusion, we have removed the offending sentence from the manuscript:

> ***Reviewer 2:***
>
> *Their measure of neural responses is derived from the decoder responses, and this takes account of the reliability of the sensory representations - the d' statistics - which is an excellent thing. It also means if I understand their analysis correctly (it could bear clarifying - see below), that they can generate from it a prediction of the performance expected if an optimal decision is made combining the neural signals from the individual modalities. I believe this is the familiar root sum of squares d' calculation (or very similar). Their decoding of the audiovisual responses comfortably exceeds this prediction and forms part of the evidence for their claims.*
>
> *Yet, superadditivity - including that in evidence in the principle of inverse effectiveness more typically quantifies the excess over the sum of proportions correct in each modality. Their MLE d' statistic can already predict this form of superadditivity. Therefore, the superadditivity they report here is not the same form of superadditivity that is usually referred to in behavioural studies. It is in fact a stiffer definition. What their analysis tests is that decoding performance exceeds what would be expected from an optimally weighted linear integration of the unisensory information. As this is not the common definition it is difficult to relate to behavioral superadditivity reported in much literature (of percentage correct). This distinction is not at all clear from the manuscript.*
>
> *But the real puzzle is here: The behavioural data or this task do not exceed the optimal statistical decision predicted by signal detection theory (the MLE d'). Yet, the EEG data would suggest that the neural processing is exceeding it. So why, if the neural processing is there to yield better performance is it not reflected in the behaviour? I cannot explain this, but it strikes me that the behaviour and neural signals are for some reason not reflecting the same processing.*
>
> *Be explicit and discuss this mismatch they observe between behaviour and neural responses.*

Thank you, we agree that it is worth expanding on the observed disconnect between MSI in behaviour and neural signals. We have included an additional paragraph in the Discussion of

the revised manuscript. Despite the mismatch, we believe the behavioural and neural responses still reflect the same underlying processing, but at different levels of sensitivity. The behavioural result likely reflects a coarse down-sampling of the precision in location representation, and thus less likely to reflect subtle MSI enhancements.

“An interesting aspect of our results is the apparent mismatch between the behavioural and neural responses. While the behavioural results meet the optimal statistical threshold predicted by MLE, the decoding analyses suggest that the neural response exceeds it. Though non-linear neural responses and statistically optimal behavioural responses are reliable phenomena in multisensory integration (Alais & Burr, 2004; Ernst & Banks, 2002; Stanford & Stein, 2007), the question remains – if neural super-additivity exists to improve behavioural performance, why is it not reflected in behavioural responses? A possible explanation for this neurobehavioural discrepancy is the large difference in timing between sensory processing and behavioural responses. A motor response would typically occur some time after the neural response to a sensory stimulus (e.g., 70-200 ms), with subsequent neural processes between perception and action that introduce noise (Heekeren et al., 2008) and may obscure super-additive perceptual sensitivity. In the current experiment, participants reported either the distribution of 20 serially presented stimuli (EEG session) or compared the positions of two stimuli (behavioural session), whereas the decoder attempts to recover the location of every presented stimulus. While stimulus location could be represented with higher fidelity in multisensory relative to unisensory conditions, this would not necessarily result in better performance on a binary behavioural task in which multiple temporally separated stimuli are compared. One must also consider the inherent differences in how super-additivity is measured at the neural and behavioural levels. Neural super-additivity should manifest in responses to each individual stimulus. In contrast, behavioural super-additivity is often reported as proportion correct, which can only emerge between conditions after being averaged across multiple trials. The former is a biological phenomenon, while the latter is an analytical construct. In our experiment, we recorded neural responses for every presentation of a stimulus, but behavioural responses were only obtained after multiple stimulus presentations. Thus, the failure to find super-additivity in behavioural responses might be due to their operationalisation, with between-condition comparisons lacking sufficient sensitivity to detect super-additive sensory improvements. Future work should focus on experimental designs that can reveal super-additive responses in behaviour.”

> *Re-work the introduction to explain more clearly the relationship between the behavioural superadditivities they review, the MLE model, and the superadditivity it actually tests.*

We agree it is worth discussing how super-additivity is operationalised across neural and behavioural measures. However, we do not believe the behavioural studies we reviewed claimed super-additive behavioural enhancements. While MLE is often used as a behavioural marker of successful integration, it is not necessarily used as evidence for super-additivity within the behavioural response, as it relies on linear operations.

“It is important to consider the differences in how super-additivity is classified between neural and behavioural measures. At the level of single neurons, superadditivity is defined as a non-linear response enhancement, with the multisensory response exceeding the sum of the unisensory responses. In behaviour, meanwhile, it has been observed that the performance improvement from combining two senses is close to what is expected from optimal integration of information across the senses (Alais & Burr, 2004; Stanford & Stein, 2007). Critically, behavioural enhancement of this kind does not require non-linearity in the neural response, but can arise from a reliability-weighted average of sensory information. In short, behavioural performance that conforms to MLE is not necessarily indicative of neural super-additivity, and the MLE model can be considered a linear baseline for multisensory integration.”

> *Regarding the auditory stimulus, this reviewer notes that interaural time differences are unlikely to survive free field presentation.*

Despite the free field presentation, in both the pilot test and the study proper participants were able to localize auditory stimuli significantly above chance.

> *"However, other studies have found super-additive enhancements to the amplitude of sensory event-related potentials (ERPs) for audiovisual stimuli (Molholm et al., 2002; Talsma et al., 2007), especially when considering the influence of stimulus intensity (Senkowski et al., 2011)." - this makes it obvious that there are some studies which show superadditivity. It would have been good to provide a little more depth here - as to what distinguished those studies that reported positive effects from those that did not.*

We have provided further detail on how super-additivity appears to manifest in neural measures.

“In EEG, meanwhile, the evoked response to an audiovisual stimulus typically conforms to a sub-additive principle (Cappe et al., 2010; Fort et al., 2002; Giard & Peronnet, 1999; Murray et al., 2016; Puce et al., 2007; Stekelenburg & Vroomen, 2007; Teder- Sälejärvi et al., 2002; Vroomen & Stekelenburg, 2010). However, when the principle of inverse effectiveness is considered and relatively weak stimuli are presented together, there has been some evidence for super-additive responses (Senkowski et al., 2011).”

“While behavioural outcomes for multisensory stimuli can be predicted by MLE, and single neuron responses follow the principles of inverse effectiveness and super- additivity, among others (Rideaux et al., 2021), how audiovisual super-additivity manifests within populations of neurons is comparatively unclear given the mixed findings from relevant fMRI and EEG studies. This uncertainty may be due to biophysical limitations of human neuroimaging techniques, but it may also be related to the analytic approaches used to study these recordings. For instance, superadditive responses to audiovisual stimuli in EEG studies are often reported from very small electrode clusters (Molholm et al., 2002; Senkowski et al., 2011; Talsma et al., 2007), suggesting that neural super-additivity in humans may be highly specific. However, information encoded by the brain can be represented as increased activity in some areas, accompanied by decreased activity in others, so simplifying complex neural responses to the average rise and fall of activity in specific sensors may obscure relevant multivariate patterns of activity evoked by a stimulus.”

> *P9. "(25-75 W, 6 Ω)." This is not important, but it is a strange way to cite the power handling of a loudspeaker.*

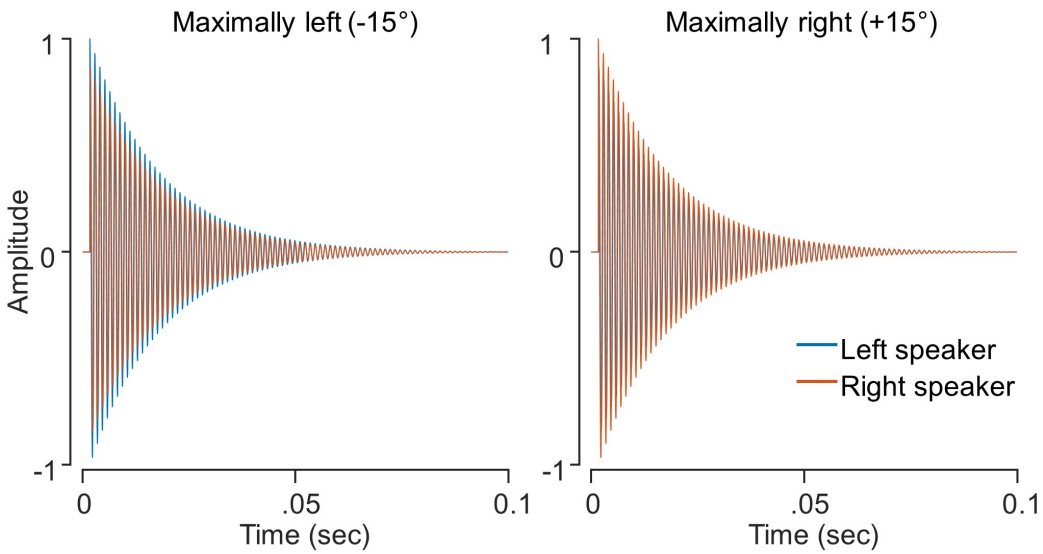
“The loudspeakers had a power handling capacity of 25-75 W and a nominal impedance of 6 Ω.”

> *I am struggling to understand the auditory stimulus:*
>
> *"Auditory stimuli were 100 ms clicks". Is this a 100-ms long train of clicks? A single pulse which is 100ms long would not sound like a click, but two clicks once filtered by the loudspeaker. Perhaps they mean 100us.*
>
> *"..with a flat 850 Hz tone embedded within a decay envelope". Does this mean the tone is gated - i.e. turns on and off slowly? Or is it constant?*

We thank the reviewer for catching this. ‘Click’ may not be the most apt way of defining the auditory stimulus. It was a 100 ms square wave tone with decay, i.e., with an onset at

maximal volume before fading gradually. Given that the length of the stimulus was 100 ms, the decay occurs quickly and provides a more 'click-like' percept than a pure tone. We have provided a representation of the sound below for further clarification. This represents the amplitude from the L and R speakers for maximally-left and maximally-right stimuli. We have added this clarification in the revised manuscript.

Author response image 1.



"Auditory stimuli were 100 ms, 850 Hz tones with a decay function (sample rate = 44, 100 Hz; volume = 60 dBA SPL, as measured at the ears)."

> *P10. "Stimulus modality was either auditory, visual, or audiovisual. Trials were blocked with short (~2 min) breaks between conditions".*
>
> *Presumably the blocks were randomised across participants.*

Condition order was not randomised across participants, but counterbalanced. This has been clarified in the manuscript.

"Stimulus modality was auditory, visual or audiovisual, presented in separate blocks with short breaks (~2 min) between conditions (see Figure 6A for an example trial). The order of conditions was counterbalanced across participants."

> *P15. Feels like there is a step not described here: "The d' of the auditory and visual conditions can be used to estimate the predicted 'optimal' sensitivity of audiovisual signals as calculated through MLE." Do they mean sqrt[(d'A)^2 + (d'V)^2] ? If it is so simple then it may as well be made explicit here. A quick calculation from eyeballing Figures 2B and 2C suggests this is the case.*

We thank the reviewer for raising this point of clarification. Yes, the 'optimal' audiovisual sensitivity was calculated as the hypotenuse of the auditory and visual sensitivities. This calculation has been made explicit in the revised manuscript.

The d' from the auditory and visual conditions can be used to estimate the predicted 'optimal' sensitivity to audiovisual signals as calculated through the following formula:

$$d'AV = \sqrt{(d'A)^2 + (d'V)^2}$$

> *"The perceived source location of auditory stimuli was manipulated via changes to interaural intensity and timing (Whitworth & Jeffress, 1961; Wightman & Kistler, 1992)." The stimuli were delivered by a pair of loudspeakers, and the incident sound at each ear would be a product of both speakers. And - if there were a time delay between the two speakers, then both ears could potentially receive separate pulses one after the other at different delays. Did they record this audio stimulus with manikin? If not, it would be very difficult to know what it was at the ears. I don't doubt that if they altered the relative volume of the loudspeakers then some directionality would be perceived but I cannot see how the interaural level and timing differences could be matched - as if the sound were from a single source. I doubt that this invalidates their results, but to present this as if it provided matched spatial and timing cues is wrong, and I cannot work out how they can attribute an azimuthal location to the sound. For replication purposes, it would be useful to know how far apart the loudspeakers were and what the timing and level differences actually were.*

The behavioural tasks each had evenly distributed 'source locations' on the horizontal azimuth of the computer display (8 for the behavioural session, 5 for the EEG session). We manipulated the perceived location of auditory stimuli through interaural time delays and interaural level differences. By first measuring the forward (*z*) and horizontal (*x*) distance of each source location to each ear, the method worked by calculating what the time-course of a sound wave should be at the location of the ear given the sound wave at the source. Then, for each source location, we can calculate the time delay between speakers given the vectors of *x* and *z*, the speed of sound and the width of the head. As the intensity of sound drops inversely with the square of the distance, we can divide the sound wave by the distance for each source location to provide the interaural level difference. Though we did not record the auditory stimulus with a manikin, our behavioural analyses show that participants were able to detect the directions of auditory stimuli from our manipulations, even to a degree that significantly exceeded the localisation accuracy for visual stimuli (for the behavioural session task). This information has been clarified in the manuscript.

"Auditory stimuli were played through two loudspeakers placed either side of the display (80 cm apart for the behavioural session, 58 cm apart for the EEG session)."

"The perceived source location of auditory stimuli was manipulated via changes to interaural level and timing (Whitworth & Jeffress, 1961; Wightman & Kistler, 1992). The precise timing of when each speaker delivered an auditory stimulus was calculated from the following formula:

$$ITD_{L,R} = \frac{\sqrt{(x \pm r)^2 + z^2}}{s}$$

where *x* and *z* are the horizontal and forward distances in metres between the ears and the source of the sound on the display, respectively, *r* is the head radius, and *s* is the speed of sound. We used a constant approximate head radius of 8 cm for all participants. *r* was added to *x* for the left speaker and subtracted for the right speaker to produce the interaural time difference. For ±15° source locations, interaural timing difference was 1.7 ms. To simulate the decrease in sound intensity as a function of distance, we calculated interaural level differences for the left and right speakers by dividing the sounds by the left and right distance vectors. Finally, we resampled the sound using linear interpolation based on the calculations of the interaural level and timing differences. This process was used to calculate the soundwaves played by the left and right speakers for each of the possible stimulus

locations on the display. The maximum interaural level difference between speakers was 0.14 *A* for ±15° auditory locations, and 0.07 *A* for ±7.5°.

> *I am confused about this statement: "A quantification of the behavioural sensitivity (i.e., steepness of the curves) revealed significantly greater sensitivity for the audiovisual stimuli than for the auditory stimuli alone (Z = -3.09,* p = *.002)," It is not clear from the methods how they attributed sound source angle to the sounds. Conceivably they know the angle of the loudspeakers, and this would provide an outer bound on the perceived location of the sound for extreme interaural level differences (although free field interaural timing cues can create a wider sound field).*

Our analysis of behavioural sensitivity was dependent on the set 'source locations' that were used to calculate the position of auditory and audiovisual stimuli. In the behavioural task, participants judged the position of the target stimulus relative to a central stimulus. Thus, for each source location, we recorded how often participants correctly discriminated between presentations. The quoted analysis acknowledges that participants were more sensitive to audiovisual stimuli than auditory stimuli in the context of this task. A full explanation of how source location was implemented for auditory stimuli has been clarified in the manuscript.

> *It would be very nice to see some of the "channel" activity - to get a feel for the representation used by the decoder.*

We have included responses for the five channels as a Supplemental Figure.

> *Figure 6 appears to show that there is some agreement between behaviour and neural responses - for the audiovisual case alone. The positive correlation of behavioural and decoding sensitivity appears to be driven by one outlier - who could not perform the audiovisual task (and indeed presumably any of them). Furthermore, if we were simply Bonferonni correct for the three comparisons, this would become non-significant. It is also puzzling why the unisensory behaviour and EEG do not correlate - which seems to again suggest a poor correspondence between them. Opposite to the claim made.*

We understand the reviewer's concern here. We would like to note, however, that each correlation used unique data sets – that is, the behavioural and neural data for each separate condition. In this case, we believe a Bonferroni correction for multiple comparisons is too conservative, as no data set was compared more than once. Neither the behavioural nor the neural data were normally distributed, and both contained outliers. Rather than reduce power through outlier rejection, we opted to test correlations using Spearman's rho, which is resistant to outliers1. It is also worth noting that, without outlier rejection, the audiovisual correlation (p = .003) would survive a Bonferroni correction for 3 comparisons. The nonsignificant correlation in the auditory and visual conditions might be due to the weaker responses elicited by unisensory stimuli, with the reduced signal-to-noise ratio obscuring potential correlations. Audiovisual stimuli elicited more precise responses both behaviourally and neurally, increasing the power to detect a correlation.

(1) Wilcox, R.R. (2016), Comparing dependent robust correlations. *British Journal of Mathematical & Statistical Psychology, 69*(3), 215-224. https://doi.org/10.1111/bmsp.12069

"We also found a significant positive correlation between participants' behavioural judgements in the EEG session and decoding sensitivity for audiovisual stimuli. This result suggests that participants who were better at identifying stimulus location also had more reliably distinct patterns of neural activity. The lack of neurobehavioural correlation in the unisensory conditions might suggest a poor correspondence between the different tasks, perhaps indicative of the differences between behavioural and neural measures explained previously. However, multisensory stimuli have consistently been found to elicit stronger

neural responses than unisensory stimuli (Meredith & Stein, 1983; Puce et al., 2007; Senkowski et al., 2011; Vroomen & Stekelenburg, 2010), which has been associated with behavioural performance (Frens & Van Opstal, 1998; Wang et al., 2008). Thus, the weaker signalto-noise ratio in unisensory conditions may prevent correlations from being detected."

Further changes:

(1) To improve clarity, we shifted the Methods section to after the Discussion. This change included updating the figure numbers to match the new order (Figure 1 becomes Figure 6, Figure 2 becomes Figure 1, and so on).

(2) We also resolved an error on Figure 2 (previously Figure 3). The final graph (Difference between AV and A + V) displayed incorrect values on the Y axis.

This has now been remedied.

https://doi.org/10.7554/eLife.97230.2.sa0