Differential Neural Encoding of Sound under Visual Attention is Shaped by Audiovisual Precision and Unimodal Uncertainty Priors

Cervantes Constantino, F.[1,*], Sánchez-Costa, T.[1], Cipriani, G.A.[1], Carboni, A.[1]
[1]Universidad de la República, Montevideo, Uruguay
[*]Correspondence: fcervantes@psico.edu.uy
Centro de Investigación Básica en Psicología
Universidad de la República
Dr. Tristán Narvaja 1674
Montevideo 11200, Uruguay

Abstract

Surroundings continually propagate audiovisual (AV) signals, and by attending we make clear and precise sense of those that matter at any given time. In such cases, parallel visual and auditory contributions may jointly serve as a basis for selection. It is unclear what hierarchical effects arise when initial selection criteria are unimodal, or involve uncertainty. Uncertainty in sensory information is a factor considered in computational models of attention proposing precision weighting as a primary mechanism for selection. The effects of visuospatial selection on auditory processing were investigated here with electroencephalography (EEG). We examined the encoding of random tone pips probabilistically associated to spatially-attended visual changes, via a temporal response function model (TRF) of the auditory EEG timeseries. AV precision, or temporal uncertainty, was manipulated across stimuli while participants sustained endogenous visuospatial attention. TRF data showed that cross-modal modulations were dominated by AV precision between auditory and visual onset times. The roles of unimodal (visuospatial and auditory) uncertainties, each a consequence of non-synchronous AV presentations, were further investigated. The TRF data demonstrated that visuospatial uncertainty in attended sector size determines transfer effects by enabling the visual priming of tones when relevant for auditory segregation, in line with top-down processing timescales. Auditory uncertainty in distractor proportion, on the other hand, determined susceptibility of early tone encoding to automatic change by incoming visual update processing. The findings provide a hierarchical account of the role of uni- and cross-modal sources of uncertainty on the neural encoding of sound dynamics in a multimodal attention task.

Keywords: audiovisual attention; uncertainty; multimodal integration; precision

Introduction

Goal-directed visuospatial attention enables selective processing of specific regions at busy scenes, so as to parse a detailed picture of relevant events (Carrasco, 2014). This mechanism allows for efficient coding by prioritizing input within specific retinotopic domains at the expense of others (O'Connor et al., 2002). Naturalistic scenarios habitually comprise multimodal sensory information in tandem with visual events, yet how visuospatial selective effects translate over concurring multimodal streams is not well understood (Koelewijn et al., 2010; Santangelo & Macaluso, 2012). For instance, unimodal vision incorporates the mechanisms for spreading spatial selection over the extended domain of foreground visual objects (Egly et al., 1994; Hollingworth et al., 2012; Wannig et al., 2011). It is not clear whether similar "spillover" effects hold during processing of cross-modal stimuli associated by temporal proximity instead (Busse et al., 2008; Degerman et al., 2007). Proximity is in such cases indexed by temporal correlation, a variable that underscores the integration of audiovisual (AV) human communication signals at perceptual and neural levels (Spence & Squire, 2003; Holmes & Spence, 2005; Roseboom et al., 2009; Park et al., 2016). For example, unimodal processing in joint visual face and speech sounds can be routinely biased according to coherent, unified percepts that emerge to disambiguate single unimodal contexts and disappear as sensory streams become desynchronyzed (Andersen et al., 2009; McGurk & MacDonald, 1976; Miller & D'Esposito, 2005; Ullas et al., 2020; van Atteveldt et al., 2007).

Spatial retinotopy is a basic selection criterion in visual attention, involving relatively early and lower-level processing than visual object-based selective mechanisms (Martínez et al., 2006; Müller & Kleinschmidt, 2003); and in some cases this form of selection may precede multimodal integration (Talsma & Woldorff, 2005). Intramodal models of visual attention have considered its direct role in the binding of visual features shared by an object, such as its color, orientation and motion, all of which are processed in parallel (Treisman, 1998; Treisman & Gelade, 1980). Whether similar principles apply during selective processing of concurrent multimodal stimuli remains elusive however (Spence & Frings, 2020; Talsma et al., 2010). This may require that the temporal structure of visual input is effective at modulating neural activity at the auditory cortex for instance. Evidence suggests that limiting the temporal coherence of AV objects may constrain how visual spatial attention influences auditory processing (Beauchamp, 2019; King et al., 2019; Van der Burg et al., 2008). Yet, it is currently unknown how are auditory populations hierarchically and functionally organized to interact with visual signals and effect such changes. In hearing, as in vision, similar cortical mechanisms operate upon the representation of competing streams of sensory data, by enhancing activity that underlies processing of attended auditory input, and attenuating that of distractors (Hillyard et al., 1973; Desimone & Duncan, 1995; Herrmann & Knight, 2001; Shinn-Cunningham, 2008). Enhancement and attenuation effects reflect the modulation to neural auditory encoding dynamics sustained during active selection. For AV encoding, these effects potentially offer a means to probe attentional 'transfer' when selection is instead given on visual terms.

In the light of a recent study by Wilsch et al., (2020), fundamental differences exist in the way that visual and auditory systems learn to expect sensory events, supportive of the view that human vision is optimized to attend to information on a spatial basis while hearing is suited to predict unfolding events. We do not currently understand what does the differential treatment of sensory expectations entail at the level of cross-modal interactions, nor its consequences on the neural code. Here, we examined the role that AV temporal uncertainty has in modulating the neural dynamics of sound processing, according to demands imposed by visual spatial attention. A visual dartboard stimulus was created, with randomly localized and occasional visual contrast changes that occurred within the disc. Tone pip presentations were paired to the contrast flips by controlled degrees of temporal uncertainty (AV precision), while participants attended to specified regions of the dartboard. AV precision was manipulated so that tone onsets preceded or succeeded visual onsets by narrow or wide margins, in a probabilistic manner. Participants were explicitly asked to estimate these margins during stimuli presentations, in order to perform a comparison task. At the same time we used electroencephalography (EEG), to assess the dynamics of sound processing under visuospatial attention by means of the temporal response function model (TRF) of cortical encoding (Crosse et al., 2016; David et al., 2007). The auditory TRF represents the linear mapping between the temporal dynamics of relevant sound input and that of the unfolding neural response timeseries, and is conceptually related to the receptive field in neural populations (Gaucher et al., 2012; Gourévitch et al., 2009). In attentional manipulations, relative changes to TRFs, each associated to competing auditory sources, are employed to estimate enhancement/attenuation effects (Ding & Simon, 2012b; O'Sullivan et al., 2015). Thus the neural encoding of random tone pip sets was examined for evidence that they can be neurally differentiated by their association to competing visual streams, according to AV precision constraints.

Temporal precision is among contributing variables to computational models of attention which propose that uncertainty parameters underlie perceptual gain control mechanisms (Dayan & Zemel, 1999; Yu & Dayan, 2005a; Whiteley & Sahani, 2008; Feldman & Friston, 2010a). These models suggest selective attention may be effected by weighing each competing neural representation of sensory input by estimates of its inherent perceptual noise (Dayan et al., 2000; Dayan & Yu, 2003; Hohwy, 2012). Relative differences in the present, or inferred, reliability of competing data sources can then be harnessed by neural computation, so as to refine and drive perceptual selection (Whiteley & Sahani, 2012). There is evidence of involvement of such type of strategies in 'fused' audiovisual perception (Rohe et al., 2019; Rohe & Noppeney, 2015; Meijer et al., 2019). Here, AV precision acts as an external uncertainty parameter, and we addressed its effects on the auditory representations of tones by visuospatial selective criteria. If transfer effects are observed, would they reflect facilitated binding only, or 'precision-weighting' as well? The former scenario explains potential transfer effects in terms of more likely synchronizations when under high AV precision, and higher likelihood of anticipating timing of relevant events, both which may regulate unimodal auditory activity (Bauer et al., 2020; Nobre & Rohenkohl, 2014). The latter scenario, in addition, does not preclude the involvement of other task-relevant uncertainty sources on attentional effects, to the extent that the temporal structure of AV events can be inferred.

To clarify this issue, in the present design participants were asked to selectively attend over different domain sizes of the dartboard (e.g. half, quarter), which allowed to examine the impact of *visuospatial uncertainty* on potential cross-modal transfer effects. If AV events are not perfectly synchronous, sometimes visual events lead auditory ones, and positional uncertainty is associated with the initial pop-up visualization expected within the domain the participant is asked to attend to. Else, when pips lead visual events, *auditory uncertainty* refers to the likelihood that a given sound is to be visually matched eventually within the attended domain. In line with precision-based proposals of attention, we hypothesized that such unimodal sources of uncertainty may additionally contribute to the emergence of cross-modal transfer effects in the task. Moreover, the TRF analyses served to identify the timeline in which these factors may come into effect. In doing so one may determine the hierarchical processing boundaries within which visual attention adjusts auditory encoding suited to cross-modal integration.

Experimental Methods

Subjects

Thirty volunteers (21 female; mean age 23.9 ± 4.0 SD) with no history of neurological or psychiatric disorders voluntarily participated in the study. All subjects provided formal written informed consent. They reported normal hearing and normal or corrected to normal visual acuity. All experiments were performed in accordance with WMA Declaration of Helsinki (World Medical Association, 2009) guidelines. The Ethics in Research Committee of the Faculty of Psychology at Universidad de la República approved the experimental procedures.

Experimental setup

Visual presentation and response time logging were performed with custom PsychoPy (Peirce, 2007) software. Visual displays were delivered over a CRT monitor (E. Systems, Inc., CA) with 40 cm size, 83 dpi resolution, and 60 Hz refresh rate. Continuous EEG recordings were performed using a BioSemi ActiveTwo 64-channel system (BioSemi, The Netherlands) with 10/20 layout at 2048 Hz digitization rate with CMS/DRL (ground). A 5th order cascaded integrator-comb low-pass filter with -3 dB at 410 Hz was applied online, after which signals were decimated to 1024 samples per second. Online high-pass response was fully DC coupled. External recordings, including electrooculographic data were recorded supra- and infra-orbitally as well as from the left versus right orbital rim and the nasal tip. Data were acquired at 24 bit resolution and analysis was performed offline. Full experimental sessions lasted ~2 h.

Experimental protocol

Visual stimuli. A dartboard of 20 cm radius, divided into 20 concentric rings of 48 angular sectors each, was displayed in the centre of a black background (Capilla et al., 2016, Figure 1A, top). Each sector (60 in total) was comprised of a 4 by 4 black and white checkerboard, scaled by corresponding eccentricity and polar angles. After a 1.1 s initial static display, checks at a pseudo-randomly selected sector reversed polarity (here termed as a visual 'flip') with the rest of sectors remaining constant. For each quadrant, flip locations defined by eccentricity (n=5) and angle (m=3), were uniformly sampled and balanced across eccentricities and quadrants, resulting in a total 15 flips per visual sequence for 'Attend-Full' (AF) conditions, 30 for 'Attend-Half' (AH), and 60 for 'Attend-Quarter' (AQ, see *Audiovisual stimuli*, below). Changed checkerboards remained so unless stimulated later again. Asynchrony between subsequent visual flips was nominally defined as a 101 ms plus an exponentially distributed random delay (μ = 99 ms), but to ensure correspondence with the video frame rate (10 fps), *effective* asynchrony values δ_i were rounded in step increments of 100 ms, ranging from 100 to 900 ms across stimuli. This guaranteed non-simultaneous, spatially distributed visual transients that could be assessed individually for temporal alignment with corresponding tone pips (Van der Burg et al., 2010). The order of visual sector flips was pseudo-randomized, and the final configuration was kept fixed for an additional 0.5 s. Visual stimuli were constructed with the MATLAB® software package (Natick, United States), stored as .avi files.

Auditory stimuli. Accompanying sound stimuli sequences were constructed with MATLAB® at a sampling rate of 22.05 KHz, consisting of pseudo-randomly presented tones of 100 ms duration (Figure

1A, bottom). Tone frequencies f_i were taken from a pool of 15 fixed values (range: 100-4846 Hz), separated by 2 equivalent rectangular bandwidth steps (Glasberg & Moore, 1990) specified by $f_i = f_{i-1} + (24.7\ (1 + 4.37 f_{i-1} / 1000))$. Tones were modulated with 5 ms raised cosine on- and off-ramps. Tone level values were calibrated according to the 60-phon normal equal-loudness-level contour (ISO 226:2003) to adjust for perceived relative loudness differences across frequencies. Individual tone onset times were based on the nominal asynchrony calculated for corresponding visual flips, plus a uniformly distributed random shift (positive or negative, Figure 1A,B) which was determined by its AV precision condition (see below). Auditory sequences consisted of 15, 30 or 60 tones mapping to every flip on video. In order to correspond spatial and spectral domains of maximal visual and auditory sensitivity respectively, tone frequency values were spatially arranged by visual eccentricity as follows: highest three frequencies (i.e. 3.2, 3.9, 4.8 KHz) reserved for all central sectors, and lowest three (100, 171, 257 Hz) used only for the most peripheral ones. Intermediate frequencies were similarly selected, in an inversely proportional order to eccentricity level. Unlike visual flips, tone presentations were allowed to overlap in time. Separately, a single “auditory stimulus probe” was constructed, which consisted of a 60-tone sequence built anew under the same rules described above. This auditory-only stimulus was presented with the purpose to find reproducible auditory activity (de Cheveigné & Parra, 2014; de Cheveigné & Simon, 2008b) in the EEG (see below). A basic sequence of 15 tones (every frequency) was quadruplicated and concatenated in time, lasting about 50 s. All auditory stimuli were saved as .wav files.

Audiovisual stimuli. To represent the degree of temporal uncertainty between audio and video streams in the sequence, AV precision conditions were defined by the distribution limits of temporal shifts between tone and flip onsets (Figure 1C). Ten AV precision levels were defined by uniformly distributed lags: in the ±33 ms interval, ±66 ms, and so forth, up to ±330 ms. Three attentional conditions were created, requiring participants to visually attend either the entire dartboard (Attend-Full ‘AF’, 15 tones-flips), or to attend one hemifield and ignore the other (Attend-Half ‘AH’, 30 tones-flips), or to attend one quadrant ignoring the rest of the dartboard (Attend-Quarter ‘AQ’, 60 tones-flips; Figure 2). The present analysis is concerned only with AH and AQ attentional conditions. In these, different AV precision levels were used for the target versus the background sectors. This was done to avoid confounding of differential foreground/background processing from potential binding effects due to high AV precision in the background stream (Jensen et al., 2019). For each stimulus in AH and AQ, attended/unattended AV precision levels were randomized and balanced such that any precision level in one attended stream was paired and equally distributed with one of five other possible values for the unattended stream. For example, a high AV precision level (±66 ms) could be paired with ±33, ±99, ±165, ±231, or ±297 ms levels, across the range. Original video and sound files were merged together into a multimedia video file with Simulink® and transcoded into .m4v format with Video Converter (V3TApps / The HandBrake Team).

Stimuli and task presentation. Subjects were instructed to attend to AV stimuli in pairs at indicated visual sectors, and to compare between first and second presentations for the relative synchronization between flips and tones, within target sectors only (Figure 2). Each trial began with a fixation cross presented during 0.5 s, after which an instruction cue was overlaid. The cue indicated the field to be attended as follows: a filled circle required participants to attend to the entire dartboard (AF); an upwards[downwards] triangle indicated for visual spatial attention to be deployed to the upper[lower] dartboard hemifield (AH); a smaller triangle tilted diagonally pointing to a specific quadrant (i.e. upper right, upper left, lower left, lower right) for AQ conditions (Figure 2 inset). After fixation and cue

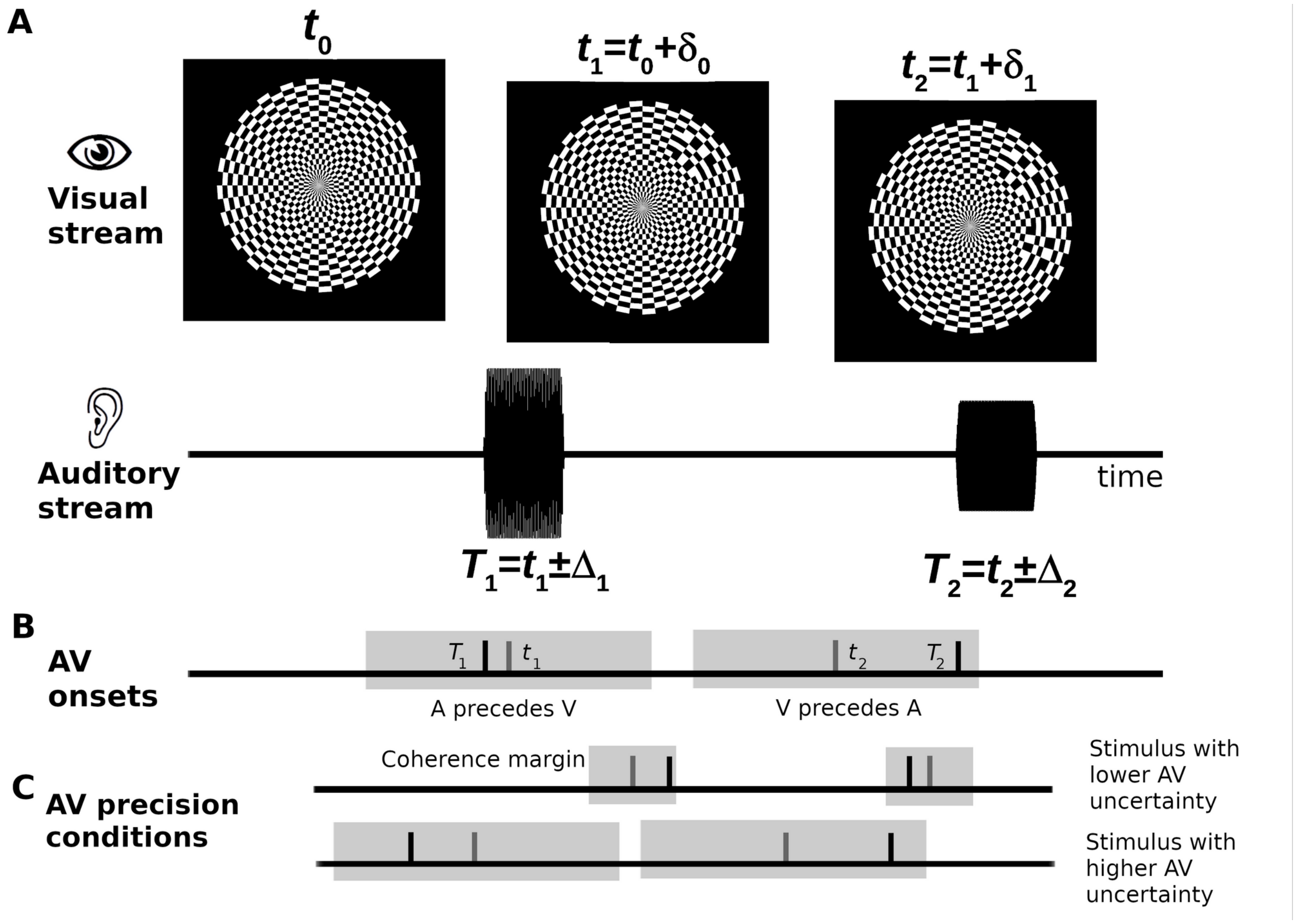

Figure 1. Basic audiovisual sequences and AV precision conditions. (A) The stimulus consists of a dartboard disk where a series of localized contrast changes (‘flips’) occurs dynamically, in parallel to a series of tone pip sounds. In this example, the first flip happens at t_1 in the upper right quadrant, followed at t_2 by another in the lower right (top). Visual events are separated by δ_i delays, ranging 100 to 900 ms. The auditory stream (bottom) consists of a presentation of accompanying tones per flip, with onset times T_i positively or negatively delayed re visual onset. (B) The temporal onset distribution from stimulus in (A) includes both cases where tones occur before or after flips. All Δ_i shifts in the sequence are uniformly distributed with limits according to the AV precision condition for that stimulus (grey area). (C) Distribution limits define the AV precision condition for a stimulus. More precise stimuli (alternatively, lower uncertainty) entail narrower time intervals of auditory lags referenced to flips. Less precise stimuli imply lower probability of tone and flip coordination (bottom).

presentation, the first AV sequence was presented, after which the fixation cross was shown once again, followed by a second sequence of identical attentional but different AV precision levels. After stimuli presentations, a question was displayed: “Which of the two audiovisual sequences was more[less] synchronized?” (here translated) with the alternatives shown (Figure 2). Subjects pressed either the left or right arrow button to highlight their choice, and hit enter to validate and receive feedback. Sequence order, stimuli combinations, question keyword (i.e. more/less), and left/right response positions were all randomized per trial, per subject. A practice session (14 trials) was carried out first, with feedback and with the experimenter present, in order to familiarize participants with every cue type and attentional condition. After this, a brief (~50 s) “auditory stimulus probe” (see *Auditory stimuli*, above) was presented for which participants were asked to listen to, attentively, with their eyes closed. During the main session participants were shown with an optional pause every 25 trials, and the experiment was finalized after 100 trials. After this, the auditory stimulus probe procedure was repeated. In appreciation for their time, participants received a chocolate bar or a cinema ticket.

Data analysis

EEG pre-processing

Data analysis was implemented in MATLAB 2018b. EEG subject data were common average-referenced to the 64 scalp channels, after which DC offset was removed. Signals were bandpass-filtered 1 – 40 Hz with a 20-order elliptical filter of 60 dB attenuation and 1 dB passband ripple. Trials were epoched according to attentional condition between -1 s and 8 s (AF), 11 s (AH) or 17 s (AQ) of AV sequence onset, then downsampled to 256 Hz . Single channel data were subsequently rejected in a blind manner according to a variance-based criterion (Junghöfer et al., 2000) applied over the 2-D matrix of concatenated trials by 64 channels, with confidence coefficient $\lambda_P=4$, and the procedure was repeated separately for the external reference channels. To reduce the effect of general movement artifacts, EEG data were decomposed into independent components using FastICA (Hyvärinen, 1999). Two independent components were automatically selected for their maximal proportion of broadband power in the 10-40 Hz region, and projected out of the raw data. In a further procedure to reduce the effect of ocular artefacts, a time-shifted principal component analysis (de Cheveigné & Simon, 2007) was applied to discard environmental signals recorded on the oculogram reference sensors (shift: ±4 ms). A sensor noise suppression algorithm (de Cheveigné & Simon, 2008a) was applied in order to attenuate artefactual components specific to any single channel (63 neighbors). The blind variance-based rejection procedure was repeated across channel by trial timeseries (64 x 200 = 12800) per participant, resulting in less than 1% rejected single-channel trial timeseries on average (subject range 0.09% - 1.52%). For the “auditory stimulus probe” subject datasets, the pre- and post-experiment recordings were pre-processed as indicated but without downsampling, and epoched into 8 trials.

Spatial filtering

To reduce dimensionality of the data and to improve SNR, a spatial filter was estimated under the criterion that it emphasizes reproducible auditory activity across subjects. This data-driven joint decorrelation or denoising by spatial filtering (DSS) (de Cheveigné & Parra, 2014; de Cheveigné & Simon, 2008b) procedure was simultaneously trained on all 30 listeners' recordings of the “auditory stimulus probe”. First, all subjects’ data were bandpass-filtered in the 1 – 8 Hz region (corresponding to delta and theta band rhythms) with a second-order Butterworth filter. To address variability across participants which may affect spatial filter estimation, the same blind rejection procedure described above was repeated over the 2-D matrix of timeseries including all participants’ data, with a more restrictive confidence coefficient $\lambda_P=2.25$. The single EEG component with the highest evoked activity ratio resulting from this process was used as a single spatial filter for all subjects (Figure 3A). Effectively analogous to sensor selection in standard ERP studies, this DSS timeseries represents the most reproducible component of the evoked data, and was selected as a single virtual sensor in analyses henceforth. To minimize computational cost of TRF estimation in individual datasets, subject DSS timeseries were downsampled to 32 Hz.

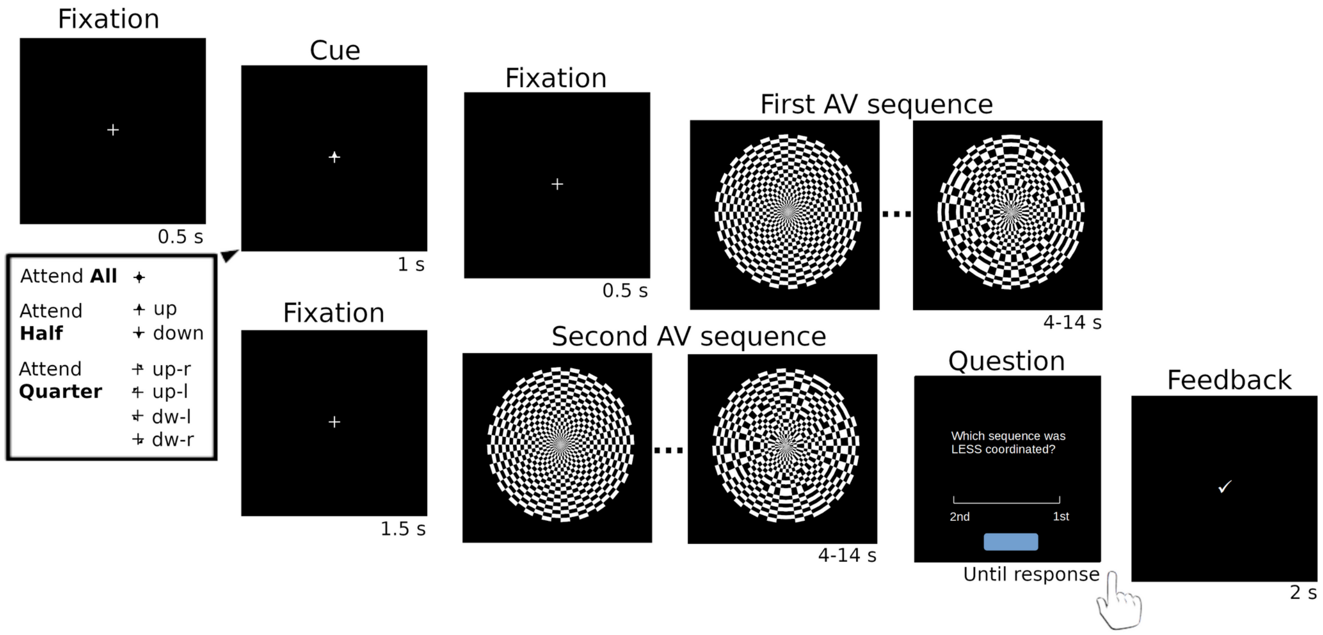

Figure 2. AV precision comparison task design. After a brief fixation period, a cue indicates the visual sector attentional strategy required for the trial. Participants are then serially shown two AV sequence presentations, attending in both as indicated by the cue (inset): over the entire disc (full circle cue, AF), or to the upper/lower half (up/down triangle cues, AH), or to one of the available quadrants (tilted triangle cues, AQ). The participant is required to compare for AV precision levels (“coordination”) of attended sectors between both sequences. After being prompted, responses are entered by a button press and trial feedback is then shown.

Temporal response function (TRF) estimation

All single-trial data were concatenated in time by attentional and AV condition as follows: a Low AV precision tier consisting of trials across the bottom five AV precision levels (±330 to ±198 ms), and a High AV precision tier for the remainder (±165 to ±33 ms). EEG data and corresponding concatenated auditory stimuli were used to estimate the TRF, per attentional and AV precision tier condition. From each EEG response recording, two TRFs were estimated: one relating the neural response to auditory events associated to the attended visual stream (Attended, ‘Att’) and, separately, another relating the same response to all other auditory events (Unattended-simultaneous, ‘Uat-sim’). This allowed to probe modulations specific to sounds from simultaneous attended/unattended presentations. To address general modulations on auditory processing regardless of a given presentation we re-ordered EEG trials by AV precision level in the unattended stream (Unattended-matched, ‘Uat-mch’).

For each subject and set (‘Att’, ‘Uat-sim’, ‘Uat-mch’), we estimated the linear TRF, a mapping between an auditory stimulus representation $S(t)$ and the evoked neural response $r(t)$ it elicits. This linear model is formulated as $r'(t)=\sum_{\tau} TRF(\tau)S(t-\tau)+\varepsilon(t)$, where $\varepsilon(t)$ is the residual contribution to the evoked response not explained by the linear model and $TRF(\tau)$ represents the TRF over the 0 to 600 ms window post auditory onset. The auditory representation for the stimulus was the temporal onset edge timeseries $S(t)$ consisting of a vector equal to 1 at times of relevant tone presentation (where ‘relevant’ depends on whether the tone belongs to an “Att”, “Uat-sim” or “Uat-mch” set) and zero elsewhere (Cervantes Constantino et al., 2017). TRFs were estimated by boosting (David et al., 2007) between stimulus and neural response timeseries, scaled to *z*-units, with 10-fold cross-validation. The average across trained TRFs represented the final TRF estimate for that subject and condition. Grand average TRFs were obtained by averaging individual TRFs across subjects.

Statistical analysis

Temporal cluster analysis. Time intervals indicating differential auditory processing in the linear TRF were obtained by estimation of the *t*-value map from condition contrasts (i.e. Att vs Uat-sim, and Att vs Uat-mch). An a priori threshold corresponding to the *t*-distribution 95th percentile was used for non-parametric randomization-based statistical testing (Maris, 2012; Maris & Oostenveld, 2007). A cluster was deemed significant if its associated *t*-statistic (sum of *t*-values within the cluster) exceeded those within the randomization distribution at an $\alpha=0.05$ level of significance. The distribution was generated by $N=2^{17}$ resamplings where conditions involved in the contrast (e.g. ‘Att’ vs ‘Uat-sim’) were randomly shuffled per participant prior to grand-averaging.

Results

AV precision facilitates visuospatial attention modulations on auditory processing

Participants’ ability to perform the AV precision comparison task was reflected in global correct response average rates of 67% (SD 9%) that were significantly above chance, shown by a Student’s t-test ($t(29)=4.3$; $p=1.8\text{x}10^{-4}$). Compared across attentional conditions, performance was on average reduced for AH and AQ conditions, however these differences were non-significant as shown by a repeated-measures ANOVA on attentional condition ($F(2)=2.69$; $p=0.076$).

To probe how visual spatial attention in the task modulated the auditory EEG during tone processing, a data-driven spatial filtering approach was used, extracting reproducible auditory activity across all participants (Figure 3A). The estimated filter was associated to a central scalp activity topography consistent with auditory sources (Stropahl et al., 2018), to which all individual participant data were projected. For each subject, the resulting auditory timeseries was reverse correlated with the auditory stimulus timeseries corresponding to visually attended (and separately, unattended) sectors in the task (Figure 3B). In addition, as we investigated relative temporal order effects, tones in the auditory stimulus timeseries were also selected according to their relative timing re visual events, i.e. tones preceding or succeeding corresponding flips. For Low AV precision conditions, this resulted in onset asynchrony distributions with median 128 ms, interquartile range 128 ms, and range 0 - 330 ms; for High AV precision, with corresponding median 39 ms, IQR 56 ms, range 0 - 165 ms.

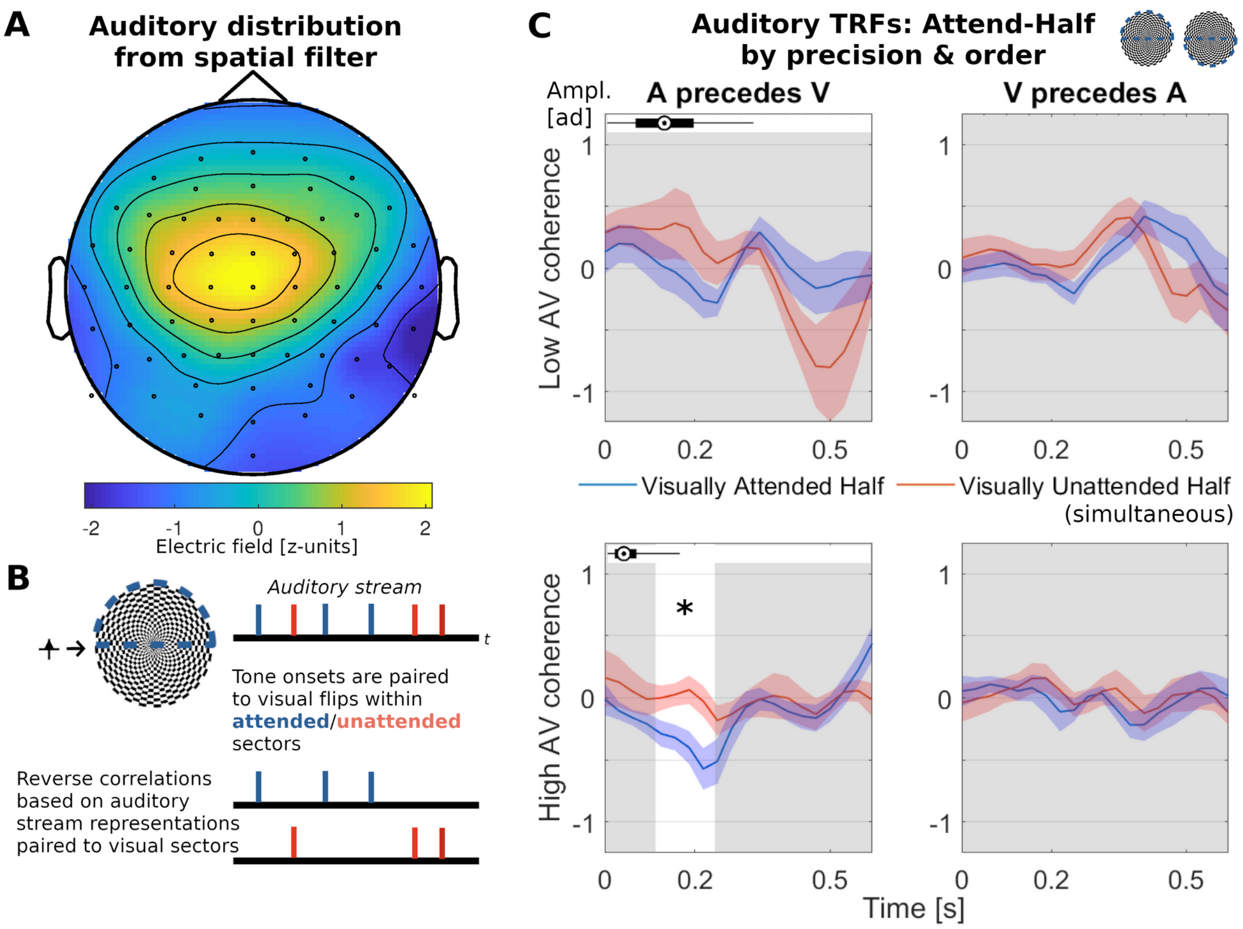

Figure 3. Auditory encoding changes by visuospatial selective attention. (A) Topography of reproducible auditory component across all participants extracted from the EEG recordings to a sample random tone pip stream with eyes closed. Individual subject data were projected onto this component, via spatial filtering, as a single virtual EEG sensor representing reproducible auditory activity. (B) Each AV sequence in AH and AQ conditions contain both attended and unattended visual domains, each associated to distinct tones in the continuous auditory stream. For TRF estimation, the auditory stream is partitioned accordingly, and simultaneous pip sequences are then represented by their corresponding tone onset times. (C) Subject results for AH show no evidence of visual spatial attention-induced changes to auditory processing under high AV uncertainty conditions (top, onset asynchrony distribution shown in inset). Under high AV precision (median shift 39 ms, bottom), visual processing modulates associated, ongoing tone processing. AV precision refers here to the condition of the attended sector only.

Temporal response function (TRF) estimates of subject data showed that, in AH conditions, high AV precision may facilitate changes to auditory processing by visual spatial attention (Figure 3C). Non-parametric testing showed a significant effect in the processing of tones paired to attended sectors versus the remainder of the auditory scene (‘Att’ vs ‘Uat-sim’; 113 to 244 ms; p=0.035; Figure 3C bottom, left), for cases where sound preceded visual onset time. No significant effect was observed however when tones instead succeeded visual onset time (p=0.674; bottom, right). In a secondary analysis, we probed whether differential processing still holds across tone presentations where unattended sectors are matched for AV precision and are not simultaneous (‘Att’ vs ‘Uat-mch’, see Methods). Such comparison served to probe whether transfer effects are specific to the ongoing trial dynamics, or if they involve change to neural tone representations by visual association – regardless of the current auditory foreground/background problem. The previous results were mirrored in this analysis, showing again a significant modulation to processing of tones paired to attended versus unattended visual sectors, for tones preceding (25 to 252 ms; p=0.005) but not succeeding (p=0.093) visual onsets.

In contrast, for low AV precision conditions the data did not reveal that either tones preceding (p=0.750; Figure 3C top, left) or tones succeeding visual onsets (p=0.214; Figure 3C top, right) are differentially processed when associated with the attended half, as expected. Similarly, in the secondary analysis controlling for the AV precision of the unattended stream in non-simultaneous presentations (‘Att’ vs ‘Uat-mch’), non-parametric testing showed that neither processing of preceding nor succeeding tones re visual onsets lead to statistically significant changes in associated TRFs (p=0.735 and p=0.794, respectively).

Cross-modal modulations by visuospatial attention on hearing are shaped by unimodal uncertainty

AQ trials similarly showed that low AV precision led to no significant modulations of tone processing, in attended versus opposite quadrant comparisons, for either preceding (p=0.737; Figure 4A, top left), nor succeeding (p=0.260; top right) tones re visual onset. These negative results were again observed in a secondary analysis (‘Att’ vs ‘Uat-mch’, Figure 5) probing for differences in auditory processing between tones associated to attended versus non-simultaneous unattended sectors controlled for AV precision, at neither preceding (p=0.895) or succeeding (p=0.722) visual onsets. The results so far showed that high temporal uncertainty between auditory and visual input may adversely impact the ability for visuospatial attention to modulate hearing. This parameter appears relevant in both multimodal integration processing and in precision-based accounts of attention. Hence it is not clear whether the observed effects so far are a direct consequence of binding (and lack thereof), or of weighted selection. Crucially, as the latter framework formalizes the involvement of uncertainty factors in attentional effects, we addressed the remaining unimodal uncertainties involved in the AV precision estimation task.

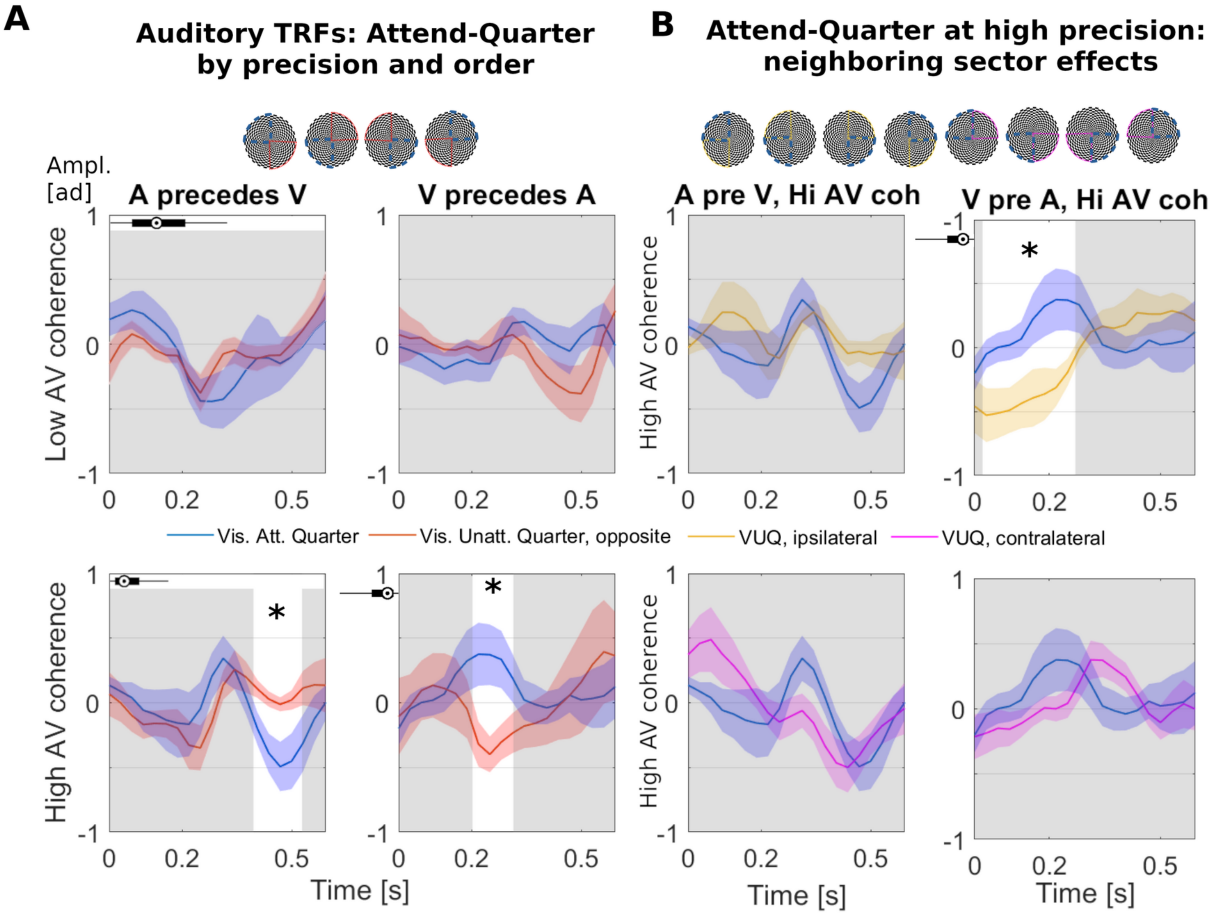

Figure 4. Positional precision required to visually prime tone processing by visual spatial attention. (A) In AQ conditions and under high AV precision, TRF estimates show differential encoding of tones associated to attended versus opposite quadrants by visuospatial attention (bottom). In contrast to AH, this includes cases where visual onsets lead tone processing: flip visualization here also induce changes to incoming tone processing by association to target versus opposite sectors (bottom right). Insets indicate temporal distributions of visual onsets in the attended sector, where significant effects were found, indicated within/outside tone processing window accordingly. (B) TRF estimate comparisons between attended and neighboring quadrants to address selectivity effects. Left: For tones preceding corresponding visual onsets, no relative change by visuospatial attention was found at either ipsilateral or contralateral quadrant comparisons. Right: For tones that follow corresponding visual onsets, relative change was observed between flips from a given quadrant and its hemifield unattended neighbor. As in Figure 3, only attended sector TRF sets are ordered by AV precision (see Figure 5 for ordering of unattended quarters).

At high AV precision, the AQ data did show facilitated cross-modal modulations to auditory processing by visual spatial attention. This was observed both when tones preceded and succeeded visual events. Significant effects were observed on tone encoding by subsequent visual updates as associated to attended versus opposite quadrants, albeit at relatively later latency (396 to 528 ms; *p*=0.041; Figure 4A bottom left) than AH conditions. There was additional evidence of modulations on tone encoding, when primed by visual flips differentially associated to attended versus opposite quadrants (202 to 314 ms; p=0.041; bottom right), an effect that had not been observed for AH conditions. A secondary analysis controlling for AV precision at the unattended stream in non-simultaneous presentations (‘Att’ vs ‘Uat-mch’, Figure 5A) again revealed significant differences to tone processing by visuospatial attention

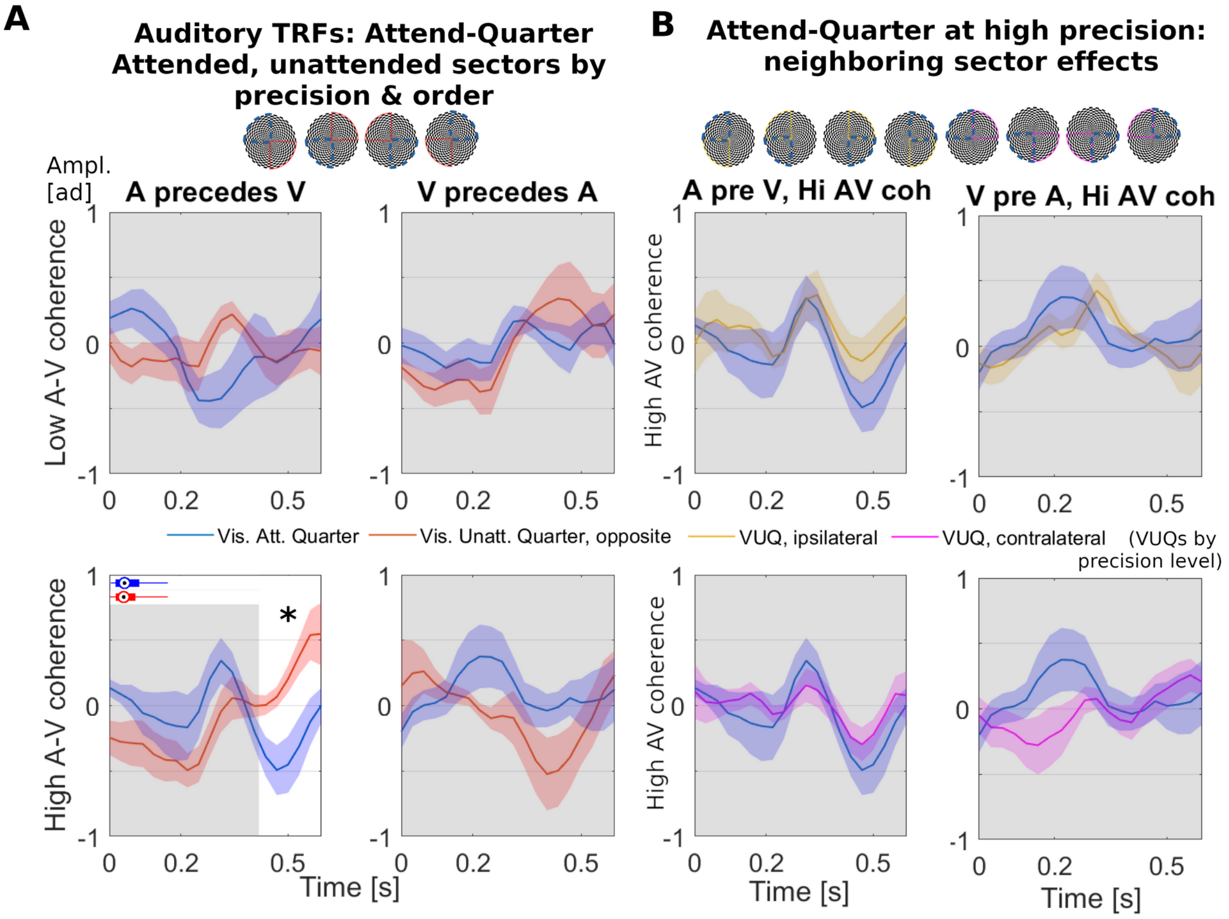

Figure 5. Differential encoding of tone targets by visuospatial primers is not independent from AV dynamics at unattended sectors. (A) For AQ conditions, reordering unattended quadrants by AV precision shows, as in AH, that visually attended flip updating modulates auditory processing regardless of AV dynamics at unattended sectors (bottom left). This however is not the case for visually attended flip priming of tones, where Visually Unattended Quarter (VUQ) TRF estimates are now ordered by AV precision regardless of whether they were presented simultaneously with the Visually Attended Quarter (bottom right). Blue and red onsets show temporal distributions of visual onsets re tones, at attended and unattended sectors respectively. (B) Likewise, no modulatory effects were observed between neighboring quadrants.

when updated by related visual onsets (419 to 594 ms; *p*=0.023). However, a similar effect was no longer observed when tones were primed by visual onsets (*p*=0.182). Finally, we assessed how are observed AQ effects specific to the opposite quadrant, which is maximally distant from the focus of attention on spatial terms. We analyzed auditory encoding of tones associated with flips at neighboring quadrants. These sectors were classified as unattended-ipsilateral and -contralateral, where the former[latter] relates to horizontally[vertically] mirroring quadrants (see Figure 4B inset). When tones preceded visual flips (Figure 4A, bottom left), TRF estimates revealed no significant differences in the processing of tones associated to the attended versus ipsilateral (*p*=0.883) nor versus contralateral quadrants (*p*=0.105; Figure 4B, left). This pattern of results was found again in secondary ‘Att’ vs ‘Uat-mch’ analyses, where unattended sectors are ordered by AV precision in non-simultaneous presentations (*p*=0.783 ipsilateral; *p*=0.731 contralateral; Figure 5B). If tones succeed visual flips, TRF estimates did reveal a significant main effect in processing of tones associated to the attended versus ipsilateral (21 to 271 ms; *p*=0.015 ; Figure 4B, top right) but not contralateral quadrants (*p*=0.799; bottom right). In ‘Att’ vs ‘Uat-mch’ analyses, the above effect with respect to unattended-ipsilateral quadrants was no longer significant however (*p*=0.769; Figure 5B, top right), and neither for unattended-contralateral sectors (*p*=0.235).

Discussion

Computational models of attention have proposed that uncertainty factors intervene in perceptual inference, for enhanced reliability of selected neural representations, therefore leading to attentional effects (P. Dayan & Zemel, 1999; P. Dayan et al., 2000; Feldman & Friston, 2010b; Whiteley & Sahani, 2012; Hohwy, 2012; Yu, 2014). In the present study, participants were asked to estimate the degree of temporal uncertainty between visual and auditory elements of an AV display, while sustaining goal-directed attention over selected visual sectors. The EEG auditory encoding results demonstrate that high AV precision facilitates cross-modal influence by visual spatial attention over auditory processing. Temporal response function (TRF) estimates also showed differential encoding of tone streams by means of visual updating mechanisms which associate individual tones to target areas. Additional visual mechanisms that bias auditory representations include the unimodal visual priming of tone streams, but this only occurs in a context-specific manner. The results, first, indicate that temporally precise margins between auditory and visual sources enable transfer of selective biases from visual attention onto hearing, as predicted in multimodally integrated objects (Donohue et al., 2011; Spence & Frings, 2020; Talsma et al., 2010). Second, that unimodal uncertainty sources arising in non-synchronous presentations shape cross-modal interactions, a finding considered in regards to precision accounts of attention. The data provide empirical support for prior auditory and, separately, visual stream information to constrain, by their precedence order, the hierarchical processing of attended stimuli with multimodal features (Rohe et al., 2019).

Visual updating effects on auditory processing: bottom-up cross-modal interactions

Earliest cross-modal effects were observed for tone streams preceding corresponding visual onsets with high AV precision. TRF analyses from Attend-Half (AH) conditions revealed auditory encoding before 250 ms to be susceptible to cross-modal modulation, consistent with observations of relatively early AV integration during visual spatial attention (Koelewijn et al., 2010). In unimodal vision, goal-directed visuospatial attention effects are typically observed from around 100 ms (Luck & Kappenman, 2012; Talsma & Woldorff, 2005), which here represents a (median) 140 ms latency after tone onset. The data provide evidence of cross-modal interactions already in effect over auditory processing. As expected, these effects were spatially organized according to the focus of visual attention, which was also observed in Attend-Quarter (AQ) conditions. For AQ we ran additional comparisons between attended and neighboring (i.e. not opposite) quadrants, which underscored the arrangement of spatial selectivity in these effects. Nevertheless, in AQ the window of effective modulation was substantially delayed (> 400 ms), which we did not initially expect. While in AH conditions, any given tone had equal prior probability to become paired with target flips, for AQ trials every standalone tone now implies a lesser chance (25%) of relating to the target sector, suggesting this prior auditory uncertainty may impact processing of auditory events within target (Roseboom et al., 2009). A limited basis for pips to anticipate target visual onsets reflects the greater pool of auditory distractors available for bottom-up capture of early visual attention by irrelevant events, thus reflecting a lack of early multimodal integration (Koelewijn et al., 2009). However, evidence of delayed differential auditory processing between target and distractor streams, for additional 280 ms than in AH, is consistent with later stage multimodal integration in these competitive conditions (Cao et al., 2019; Koelewijn et al., 2010).

After additional controlling of AV precision levels by rearrangement of unattended streams, regardless of what occurred simultaneously at the attended stream (Jensen et al., 2019), we observed an identical pattern of results. This suggests that differential encoding of tones, when induced by visual updating, implies changes to their neural representation specified by relevance but not context. Context independence in hierarchical processing assumes that the particular details of the auditory figure/ground stream segregation problem posed by any given trial are not taken into account, nor load demands in the task (Rahne et al., 2007). This appears consistent with automatic representational changes to tone processing by visual input, which in AH conditions were sufficiently early to suggest cross-modal interactions within the first bottom-up sweep of AV integration (Giard & Peronnet, 1999; Kayser et al., 2007; Noesselt et al., 2007; Stein & Meredith, 1993). Due to its temporal processing advantages, hearing is suited for rapid warning to engage or divert visuospatial attention (e.g. Spence & Driver, 1997; Fujisaki et al., 2004). Sound stimuli may 'tag' time instances for attentional deployment, e.g. by preparing cortical excitability at windows that coincide with bottom-up visual processing (Romei et al., 2007), and where subsequent visual recognition feedback serves for multimodal integration (Noesselt et al., 2008; Van der Burg et al., 2008, 2011).

Visual priming effects on auditory processing: top-down cross-modal interactions

When tones followed visual onsets, the results showed cross-modal modulations only at increased visuospatial precision conditions in the task. These AQ findings suggest that when tone onsets effectively close the AV temporal gap initiated by visual flips, positional uncertainties may shape transfer effects over auditory encoding (Andersen et al., 2009). TRF changes, observed from 200 ms, were thus present when visuospatial integration was limited to these smaller regions. As events at AH entail effectively faster presentation rates, null findings therein may reflect a limited ability for individual flips to prompt preparatory auditory processing (e.g. (Olivers & Van der Burg, 2008; Quak et al., 2015; Thorne et al., 2011). Unlike tones, flips are transients distributed across varying spatial channel sizes which impacts serial spatial attention processing (Fujisaki et al., 2006), and the AH results indicated no evidence of subsequent change to auditory encoding nor baseline levels (e.g. Feldman & Friston, 2010a). These results may suggest visuospatial domain size constrains the efficiency with which visual input influences auditory expectations about sound presentations in corresponding auditory areas (Macaluso et al., 2016). They may also relate to proposals that top-down cross-modal influences from visual to auditory cortical regions may be directly exerted by delta/theta rhythm signaling (Keil & Senkowski, 2018) in natural AV speech processing, which corresponds to activity bands upon which auditory TRF analyses of the EEG timeseries are often based (Crosse et al., 2016; Di Liberto et al., 2015; Ding & Simon, 2012a).

Observed transfer effects were no longer evident after reordering presentations by AV precision in the unattended stream. A possible explanation for this result is that transfer effects mainly reflect the encoding of relative differences across attended/unattended precision levels within the same AV sequence, implying some form of explicit processing of events within unattended sectors. However, the present AQ effects occurred at mid-latency (240 ms relative to median visual presentation), while subjects were instructed to compare across AV stimuli for input from attended sectors only. The null results thus appear more consistent with non-local visual 'prime' cueing (Maier et al., 2011) due to high AV precision, i.e., increased synchronization-facilitated priming regardless of spatial location over the disc. In AQ, global priming by high AV precision could then be exploited supramodally according to auditory segregation needs, i.e., the active process of separating between tones in the auditory stream of a given trial by their visual association (Spagna et al., 2015). This process would be compromised in AH due to bottom-up ambient competition countering effects for attended targets (Huang & Elhilali, 2020). Given both the latency and context-specificity of these effects, we suggest that such fine-tuned selectivity would be consistent with top-down modulations from visual to auditory processing. The potential relation of these results is next discussed in relation to in hierarchical models.

Prior uncertainties shape cross-modal visual spatial attention effects on hearing

By addressing the computational goals of selective attention, ideal-observer frameworks have emphasized the role of noise in weighing sensory data for selection (Rohe & Noppeney, 2018; Yu, 2014). In these models, representations about sensory content are considered along with the inferred uncertainty or noise that their sensory causes may have in the environment (Yu & Dayan, 2002, 2005b). Such factors include external uncertainty variables but also internal noise estimation in routine perceptual inference (Feldman & Friston, 2010b; Hohwy, 2012), and it is not well understood how are such parameters implemented into selection (e.g. as division weights, Reynolds & Heeger, 2009) in the case of audiovisual elements. Temporal precision is an external factor that drives binding in multimodal integration, and participants here assessed and learned AV precision of stimuli in relation to a visuospatial target area, in order to perform a task.

A potential interpretation is that facilitated cross-modal transfer effects at high AV precision reflect the modulation of auditory representations of tones as attended elements of a multimodal object, once formed at multisensory areas (Beauchamp et al., 2004; Eimer et al., 2004; Noesselt et al., 2007; Talsma et al., 2010). In the precision-based account of selective attention, it is not necessarily assumed that selective effects precede, correspond to or reflect a multimodal object formation stage. For instance, some evidence suggests that differential weighing may be allocated to individual tones in advance of robust perceptual integration (Ernst & Bülthoff, 2004; Rohe et al., 2019; Spence & Ngo, 2012). Evidence is mixed regarding whether multimodal binding requires attention beforehand or not (Koelewijn et al., 2010), nevertheless. Different task-dependent effects reported across studies may indeed reflect that integration can be effected at subcortical, early unimodal and/or late heteromodal areas (Calvert & Thesen, 2004; Koelewijn et al., 2010), as indicated by across-hierarchy integration sites including the superior colliculus, primary sensory areas, and associative areas like the intraparietal sulcus, involved in visuospatial attention (Koelewijn et al., 2010). For hearing, auditory cross-modal integration and attentional mechanisms may operate on partially overlapping circuitry along the depth axis of the auditory cortex (Gau et al., 2020).

Nevertheless, the precision framework may predict attentional effects to arise also when other sources of uncertainty serve the perceptual inference system in gauging reliability. Indeed, the results showed that unimodal uncertainty sources shape crossmodal interactions, as visual and auditory streams were not perfectly synchronous. This last evidence further indicates that distinct hierarchical relationships emerge in cross-modal interactions depending on precedence orders between visual and auditory streams. First, sounds presented in advance of visual events carry intrinsic uncertainty on whether they may pair with the visually attended sector. AH conditions have such lower auditory uncertainty, which led to transfer effects observed relatively earlier than at AQ. Differential auditory processing was maintained regardless of simultaneous content in unattended streams, which suggested in this case an automatic mode of changes to the auditory representation of tones by visuospatial association. In addition, the timeline of transfer effects for AH also appears more consistent with that of detection mechanisms (Parise & Ernst, 2016) instantiated from the initial bottom-up visual analysis (Foxe & Schroeder, 2005). Second, sounds presented after visual events were subject to the initial uncertainty determined by visual domain size, which affected the expression of cross-modal interactions. Here, AQ conditions of higher positional precision were the only instance where modulatory effects on tone processing by visual priming were observed in the task. This piece of evidence suggests that anticipatory mechanisms triggered by a visual prime may reshape auditory expectations (Nobre & Rohenkohl, 2014). This was not observed across non-simultaneous presentations, however, suggesting the mechanisms are part of active foreground segregation processing of the auditory scene. The results of context-dependent modulations on auditory processing thus appear specifically adapted to the particular dynamics of a given trial, as visual constraints effectively define AV foreground boundaries for effective auditory segregation. The findings overall underscore auditory neural segregation, by top-down visual priming at high spatial precision – when vision leads temporal order-, or possibly by bottom-up visual updating according to spatial selection rules if applied soon after tone processing began to unfold. They both support a hierarchical account of attentional heteromodal transfer mechanisms (Talsma et al., 2010) where AV temporal and visuospatial organization jointly condition the process of neural auditory stream segregation. They may do so by harnessing temporal predictions that audiovisual constructs inherently drive, which in turn influence the order in which cross-modal interactions synchronize the visual and auditory processing streams.

Acknowledgments

The research that produced the results presented in this manuscript received funding from the Agencia Nacional de Investigación e Innovación under the code PD_NAC_2018_1_150365.

Data availability statement

The data that support the findings of this study are openly available in Open Science Framework at http://doi.org/10.17605/OSF.IO/8V9SD

A preliminary version of this work is posted on the bioRxiv preprint server.

References

Andersen, T. S., Tiippana, K., Laarni, J., Kojo, I., & Sams, M. (2009). The role of visual spatial attention in audiovisual speech perception. *Speech Communication*, *51*(2), 184–193. https://doi.org/10.1016/j.specom.2008.07.004

Bauer, A.-K. R., Debener, S., & Nobre, A. C. (2020). Synchronisation of Neural Oscillations and Cross-modal Influences. *Trends in Cognitive Sciences*, *24*(6), 481–495. https://doi.org/10.1016/j.tics.2020.03.003

Beauchamp, M. S. (2019). Using Multisensory Integration to Understand the Human Auditory Cortex. In A. K. C. Lee, M. T. Wallace, A. B. Coffin, A. N. Popper, & R. R. Fay (Eds.), *Multisensory Processes: The Auditory Perspective* (pp. 161–176). Springer International Publishing. https://doi.org/10.1007/978-3-030-10461-0_8

Beauchamp, M. S., Lee, K. E., Argall, B. D., & Martin, A. (2004). Integration of Auditory and Visual Information about Objects in Superior Temporal Sulcus. *Neuron*, *41*(5), 809–823. https://doi.org/10.1016/S0896-6273(04)00070-4

Busse, L., Katzner, S., & Treue, S. (2008). Temporal dynamics of neuronal modulation during exogenous and endogenous shifts of visual attention in macaque area MT. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(42), 16380–16385. https://doi.org/10.1073/pnas.0707369105

Calvert, G. A., & Thesen, T. (2004). Multisensory integration: Methodological approaches and emerging principles in the human brain. *Journal of Physiology, Paris*, *98*(1–3), 191–205. https://doi.org/10.1016/j.jphysparis.2004.03.018

Cao, Y., Summerfield, C., Park, H., Giordano, B. L., & Kayser, C. (2019). Causal Inference in the Multisensory Brain. *Neuron*, *102*(5), 1076-1087.e8. https://doi.org/10.1016/j.neuron.2019.03.043

Capilla, A., Melcón, M., Kessel, D., Calderón, R., Pazo-Álvarez, P., & Carretié, L. (2016). Retinotopic mapping of visual event-related potentials. *Biological Psychology*, *118*, 114–125. https://doi.org/10.1016/j.biopsycho.2016.05.009

Carrasco, M. (2014). Spatial Covert Attention. *The Oxford Handbook of Attention*. https://doi.org/10.1093/oxfordhb/9780199675111.013.004

Cervantes Constantino, F., Villafañe-Delgado, M., Camenga, E., Dombrowski, K., Walsh, B., & Simon, J. Z. (2017). Functional significance of spectrotemporal response functions obtained using magnetoencephalography. *BioRxiv*, 168997. https://doi.org/10.1101/168997

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00604

David, S. V., Mesgarani, N., & Shamma, S. A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network: Computation in Neural Systems*, *18*(3), 191–212. https://doi.org/10.1080/09548980701609235

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, *3 Suppl*, 1218–1223. https://doi.org/10.1038/81504

Dayan, P., & Zemel, R. S. (1999). Statistical models and sensory attention. *1999 Ninth International Conference on Artificial Neural Networks ICANN 99. (Conf. Publ. No. 470)*, *2*, 1017–1022 vol.2. https://doi.org/10.1049/cp:19991246

Dayan, Peter, & Yu, A. J. (2003). Uncertainty and Learning. *IETE Journal of Research*, *49*(2–3), 171–181. https://doi.org/10.1080/03772063.2003.11416335

de Cheveigné, A., & Parra, L. C. (2014). Joint decorrelation, a versatile tool for multichannel data analysis. *NeuroImage*, *98*, 487–505. https://doi.org/10.1016/j.neuroimage.2014.05.068

de Cheveigné, A., & Simon, J. Z. (2007). Denoising based on time-shift PCA. *Journal of Neuroscience Methods*, *165*(2), 297–305. https://doi.org/10.1016/j.jneumeth.2007.06.003

de Cheveigné, A., & Simon, J. Z. (2008a). Sensor noise suppression. *Journal of Neuroscience Methods*, *168*(1), 195–202. https://doi.org/10.1016/j.jneumeth.2007.09.012

de Cheveigné, A., & Simon, J. Z. (2008b). Denoising based on spatial filtering. *Journal of Neuroscience Methods*, *171*(2), 331–339. https://doi.org/10.1016/j.jneumeth.2008.03.015

Degerman, A., Rinne, T., Pekkola, J., Autti, T., Jääskeläinen, I. P., Sams, M., & Alho, K. (2007). Human brain activity associated with audiovisual perception and attention. *NeuroImage*, *34*(4), 1683–1691. https://doi.org/10.1016/j.neuroimage.2006.11.019

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Current Biology*, *25*(19), 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030

Ding, N., & Simon, J. Z. (2012a). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology*, *107*(1), 78–89. https://doi.org/10.1152/jn.00297.2011

Ding, N., & Simon, J. Z. (2012b). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854–11859. https://doi.org/10.1073/pnas.1205381109

Donohue, S. E., Roberts, K. C., Grent-'t-Jong, T., & Woldorff, M. G. (2011). The Cross-Modal Spread of Attention Reveals Differential Constraints for the Temporal and Spatial Linking of Visual and Auditory Stimulus Events. *Journal of Neuroscience*, *31*(22), 7982–7990. https://doi.org/10.1523/JNEUROSCI.5298-10.2011

Egly, R., Driver, J., & Rafal, R. (1994). Shifting Visual-Attention Between Objects and Locations—Evidence from Normal and Parietal Lesion Subjects. *Journal of Experimental Psychology-General*, *123*(2), 161–177. https://doi.org/10.1037//0096-3445.123.2.161

Eimer, M., Velzen, J. van, & Driver, J. (2004). ERP Evidence for Cross-Modal Audiovisual Effects of Endogenous Spatial Attention within Hemifields. *Journal of Cognitive Neuroscience*, *16*(2), 272–288. https://doi.org/10.1162/089892904322984562

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*(4), 162–169. https://doi.org/10.1016/j.tics.2004.02.002

Feldman, H., & Friston, K. (2010a). Attention, Uncertainty, and Free-Energy. *Frontiers in Human Neuroscience*, *4*. https://doi.org/10.3389/fnhum.2010.00215

Feldman, H., & Friston, K. J. (2010b). Attention, Uncertainty, and Free-Energy. *Frontiers in Human*

Neuroscience, *4*. https://doi.org/10.3389/fnhum.2010.00215

Foxe, J. J., & Schroeder, C. E. (2005). The case for feedforward multisensory convergence during early cortical processing. *Neuroreport*, *16*(5), 419–423. https://doi.org/10.1097/00001756-200504040-00001

Fujisaki, W., Koene, A., Arnold, D., Johnston, A., & Nishida, S. (2006). Visual search for a target changing in synchrony with an auditory signal. *Proceedings. Biological Sciences*, *273*(1588), 865–874. https://doi.org/10.1098/rspb.2005.3327

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nature Neuroscience*, *7*(7), 773–778. https://doi.org/10.1038/nn1268

Gau, R., Bazin, P.-L., Trampel, R., Turner, R., & Noppeney, U. (2020). Resolving multisensory and attentional influences across cortical depth in sensory cortices. *ELife*, *9*, e46856. https://doi.org/10.7554/eLife.46856

Gaucher, Q., Edeline, J.-M., & Gourévitch, B. (2012). How different are the local field potentials and spiking activities? Insights from multi-electrodes arrays. *Journal of Physiology-Paris*, *106*(3–4), 93–103. https://doi.org/10.1016/j.jphysparis.2011.09.006

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

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research*, *47*(1–2), 103–138. https://doi.org/10.1016/0378-5955(90)90170-T

Gourévitch, B., Noreña, A., Shaw, G., & Eggermont, J. J. (2009). Spectrotemporal receptive fields in anesthetized cat primary auditory cortex are context dependent. *Cerebral Cortex (New York, N.Y.: 1991)*, *19*(6), 1448–1461. https://doi.org/10.1093/cercor/bhn184

Herrmann, C. S., & Knight, R. T. (2001). Mechanisms of human attention: Event-related potentials and oscillations. *Neuroscience & Biobehavioral Reviews*, *25*(6), 465–476. https://doi.org/10.1016/S0149-7634(01)00027-6

Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical signs of selective attention in the human brain. *Science (New York, N.Y.)*, *182*(4108), 177–180.

Hohwy, J. (2012). Attention and Conscious Perception in the Hypothesis Testing Brain. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00096

Hollingworth, A., Maxcey-Richard, A. M., & Vecera, S. P. (2012). The spatial distribution of attention within and across objects. *Journal of Experimental Psychology. Human Perception and Performance*, *38*(1), 135–151. https://doi.org/10.1037/a0024463

Holmes, N. P., & Spence, C. (2005). Multisensory integration: Space, time, & superadditivity. *Current Biology : CB*, *15*(18), R762–R764. https://doi.org/10.1016/j.cub.2005.08.058

Huang, N., & Elhilali, M. (2020). Push-pull competition between bottom-up and top-down auditory attention to natural soundscapes. *ELife*, *9*, e52984. https://doi.org/10.7554/eLife.52984

Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, *10*(3), 626–634. https://doi.org/10.1109/72.761722

Jensen, A., Merz, S., Spence, C., & Frings, C. (2019). Overt spatial attention modulates multisensory selection. *Journal of Experimental Psychology. Human Perception and Performance*, *45*(2), 174–188. https://doi.org/10.1037/xhp0000595

Junghöfer, M., Elbert, T., Tucker, D. M., & Rockstroh, B. (2000). Statistical control of artifacts in dense array EEG/MEG studies. *Psychophysiology*, *37*(04), 523–532. https://doi.org/null

Kayser, C., Petkov, C. I., Augath, M., & Logothetis, N. K. (2007). Functional imaging reveals visual modulation of specific fields in auditory cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *27*(8), 1824–1835. https://doi.org/10.1523/JNEUROSCI.4737-06.2007

Keil, J., & Senkowski, D. (2018). Neural Oscillations Orchestrate Multisensory Processing. *The Neuroscientist*, *24*(6), 609–626. https://doi.org/10.1177/1073858418755352

King, A. J., Hammond-Kenny, A., & Nodal, F. R. (2019). Multisensory Processing in the Auditory Cortex. In A. K. C. Lee, M. T. Wallace, A. B. Coffin, A. N. Popper, & R. R. Fay (Eds.), *Multisensory Processes: The Auditory Perspective* (pp. 105–133). Springer International Publishing. https://doi.org/10.1007/978-3-030-10461-0_6

Koelewijn, T., Bronkhorst, A., & Theeuwes, J. (2009). Auditory and visual capture during focused visual attention. *Journal of Experimental Psychology. Human Perception and Performance*, *35*(5), 1303–1315. https://doi.org/10.1037/a0013901

Koelewijn, T., Bronkhorst, A., & Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audiovisual studies. *Acta Psychologica*, *134*(3), 372–384. https://doi.org/10.1016/j.actpsy.2010.03.010

Luck, S. J., & Kappenman, E. S. (2012). ERP components and selective attention. In *The Oxford handbook of event-related potential components* (pp. 295–327). Oxford University Press.

Macaluso, E., Noppeney, U., Talsma, D., Vercillo, T., Hartcher-O'Brien, J., & Adam, R. (2016). The Curious Incident of Attention in Multisensory Integration: Bottom-up vs. Top-down. *Multisensory Research*, *29*(6–7), 557–583. https://doi.org/10.1163/22134808-00002528

Maier, J. X., Di Luca, M., & Noppeney, U. (2011). Audiovisual asynchrony detection in human speech. *Journal of Experimental Psychology. Human Perception and Performance*, *37*(1), 245–256. https://doi.org/10.1037/a0019952

Maris, E. (2012). Statistical testing in electrophysiological studies. *Psychophysiology*, *49*(4), 549–565. https://doi.org/10.1111/j.1469-8986.2011.01320.x

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Martínez, A., Teder-Sälejärvi, W., Vazquez, M., Molholm, S., Foxe, J. J., Javitt, D. C., Di Russo, F., Worden, M. S., & Hillyard, S. A. (2006). Objects Are Highlighted by Spatial Attention. *Journal of Cognitive Neuroscience*, *18*(2), 298–310. https://doi.org/10.1162/jocn.2006.18.2.298

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

Meijer, D., Veselič, S., Calafiore, C., & Noppeney, U. (2019). Integration of audiovisual spatial signals is not consistent with maximum likelihood estimation. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *119*, 74–88. https://doi.org/10.1016/j.cortex.2019.03.026

Miller, L. M., & D'Esposito, M. (2005). Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *25*(25), 5884–5893. https://doi.org/10.1523/JNEUROSCI.0896-05.2005

Müller, N. G., & Kleinschmidt, A. (2003). Dynamic Interaction of Object- and Space-Based Attention in Retinotopic Visual Areas. *Journal of Neuroscience*, *23*(30), 9812–9816. https://doi.org/10.1523/JNEUROSCI.23-30-09812.2003

Nobre, A. C., & Rohenkohl, G. (2014). Time for the fourth dimension in attention. In *The Oxford handbook of attention* (pp. 676–721). Oxford University Press.

Noesselt, T., Bergmann, D., Hake, M., Heinze, H.-J., & Fendrich, R. (2008). Sound increases the saliency of visual events. *Brain Research*, *1220*, 157–163. https://doi.org/10.1016/j.brainres.2007.12.060

Noesselt, T., Rieger, J. W., Schoenfeld, M. A., Kanowski, M., Hinrichs, H., Heinze, H.-J., & Driver, J. (2007). Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *27*(42), 11431–11441. https://doi.org/10.1523/JNEUROSCI.2252-07.2007

O'Connor, D. H., Fukui, M. M., Pinsk, M. A., & Kastner, S. (2002). Attention modulates responses in the human lateral geniculate nucleus. *Nature Neuroscience*, *5*(11), 1203–1209. https://doi.org/10.1038/nn957

Olivers, C. N. L., & Van der Burg, E. (2008). Bleeping you out of the blink: Sound saves vision from oblivion. *Brain Research*, *1242*, 191–199. https://doi.org/10.1016/j.brainres.2008.01.070

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., Slaney, M., Shamma, S. A., & Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706. https://doi.org/10.1093/cercor/bht355

Parise, C. V., & Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. *Nature Communications*, *7*, 11543. https://doi.org/10.1038/ncomms11543

Park, H., Kayser, C., Thut, G., & Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *ELife*, *5*. https://doi.org/10.7554/eLife.14521

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Quak, M., London, R. E., & Talsma, D. (2015). A multisensory perspective of working memory. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00197

Rahne, T., Böckmann, M., von Specht, H., & Sussman, E. S. (2007). Visual cues can modulate integration and segregation of objects in auditory scene analysis. *Brain Research*, *1144*, 127–135. https://doi.org/10.1016/j.brainres.2007.01.074

Reynolds, J. H., & Heeger, D. J. (2009). The Normalization Model of Attention. *Neuron*, *61*(2), 168–185. https://doi.org/10.1016/j.neuron.2009.01.002

Rohe, T., Ehlis, A.-C., & Noppeney, U. (2019). The neural dynamics of hierarchical Bayesian causal inference in multisensory perception. *Nature Communications*, *10*(1), 1–17. https://doi.org/10.1038/s41467-019-09664-2

Rohe, T., & Noppeney, U. (2015). Cortical Hierarchies Perform Bayesian Causal Inference in Multisensory Perception. *PLOS Biology*, *13*(2), e1002073. https://doi.org/10.1371/journal.pbio.1002073

Rohe, T., & Noppeney, U. (2018). Reliability-Weighted Integration of Audiovisual Signals Can Be Modulated by Top-down Attention. *ENeuro*, *5*(1). https://doi.org/10.1523/ENEURO.0315-17.2018

Romei, V., Murray, M. M., Merabet, L. B., & Thut, G. (2007). Occipital Transcranial Magnetic Stimulation Has Opposing Effects on Visual and Auditory Stimulus Detection: Implications for Multisensory Interactions. *Journal of Neuroscience*, *27*(43), 11465–11472. https://doi.org/10.1523/JNEUROSCI.2827-07.2007

Roseboom, W., Nishida, S., & Arnold, D. H. (2009). The sliding window of audio-visual simultaneity. *Journal of Vision*, *9*(12), 4.1-8. https://doi.org/10.1167/9.12.4

Santangelo, V., & Macaluso, E. (2012). Spatial attention and audiovisual processing. In *The New*

Handbook of Multisensory Processing. MIT Press. https://www.academia.edu/16745204/Spatial_attention_and_audiovisual_processing

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

Spagna, A., Mackie, M.-A., & Fan, J. (2015). Supramodal executive control of attention. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00065

Spence, C., & Driver, J. (1997). Audiovisual links in exogenous covert spatial orienting. *Perception & Psychophysics*, *59*(1), 1–22. https://doi.org/10.3758/bf03206843

Spence, C., & Ngo, M. K. (2012). Does Attention or Multisensory Integration Explain the Cross-Modal Facilitation of Masked Visual Target Identification? In Stein Barry E. (Ed.), *The New Handbook of Multisensory Processing* (1st edition). The MIT Press. http://cognet.mit.edu/erefs/new-handbook-of-multisensory-processing

Spence, C., & Frings, C. (2020). Multisensory feature integration in (and out) of the focus of spatial attention. *Attention, Perception, & Psychophysics*, *82*(1), 363–376. https://doi.org/10.3758/s13414-019-01813-5

Spence, C., & Squire, S. (2003). Multisensory Integration: Maintaining the Perception of Synchrony. *Current Biology*, *13*(13), R519–R521. https://doi.org/10.1016/S0960-9822(03)00445-7

Stein, B. E., & Meredith, M. A. (1993). *The merging of the senses* (pp. xv, 211). The MIT Press.

Stropahl, M., Bauer, A.-K. R., Debener, S., & Bleichner, M. G. (2018). Source-Modeling Auditory Processes of EEG Data Using EEGLAB and Brainstorm. *Frontiers in Neuroscience*, *12*. https://doi.org/10.3389/fnins.2018.00309

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Sciences*, *14*(9), 400–410. https://doi.org/10.1016/j.tics.2010.06.008

Talsma, D., & Woldorff, M. G. (2005). Selective attention and multisensory integration: Multiple phases of effects on the evoked brain activity. *Journal of Cognitive Neuroscience*, *17*(7), 1098–1114. https://doi.org/10.1162/0898929054475172

Thorne, J. D., De Vos, M., Viola, F. C., & Debener, S. (2011). Cross-modal phase reset predicts auditory task performance in humans. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(10), 3853–3861. https://doi.org/10.1523/JNEUROSCI.6176-10.2011

Treisman, A. (1998). Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *353*(1373), 1295–1306.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Ullas, S., Hausfeld, L., Cutler, A., Eisner, F., & Formisano, E. (2020). Neural Correlates of Phonetic Adaptation as Induced by Lexical and Audiovisual Context. *Journal of Cognitive Neuroscience*, 1–14. https://doi.org/10.1162/jocn_a_01608

van Atteveldt, N. M., Formisano, E., Blomert, L., & Goebel, R. (2007). The effect of temporal asynchrony on the multisensory integration of letters and speech sounds. *Cerebral Cortex (New York, N.Y.: 1991)*, *17*(4), 962–974. https://doi.org/10.1093/cercor/bhl007

Van der Burg, E., Cass, J., Olivers, C. N. L., Theeuwes, J., & Alais, D. (2010). Efficient Visual Search from Synchronized Auditory Signals Requires Transient Audiovisual Events. *PLOS ONE*, *5*(5), e10664. https://doi.org/10.1371/journal.pone.0010664

Van der Burg, E., Olivers, C. N. L., Bronkhorst, A. W., & Theeuwes, J. (2008). Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology. Human Perception and Performance*, *34*(5), 1053–1065. https://doi.org/10.1037/0096-1523.34.5.1053

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

Wannig, A., Stanisor, L., & Roelfsema, P. R. (2011). Automatic spread of attentional response modulation along Gestalt criteria in primary visual cortex. *Nature Neuroscience*, *14*(10), 1243–1244. https://doi.org/10.1038/nn.2910

Whiteley, L., & Sahani, M. (2008). Implicit knowledge of visual uncertainty guides decisions with asymmetric outcomes. *Journal of Vision*, *8*(3), 2.1-215. https://doi.org/10.1167/8.3.2

Whiteley, L., & Sahani, M. (2012). Attention in a Bayesian Framework. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00100

Wilsch, A., Mercier, M. R., Obleser, J., Schroeder, C. E., & Haegens, S. (2020). Spatial Attention and Temporal Expectation Exert Differential Effects on Visual and Auditory Discrimination. *Journal of Cognitive Neuroscience*, *32*(8), 1562–1576. https://doi.org/10.1162/jocn_a_01567

World Medical Association (WMA). (2009). Declaration of Helsinki. Ethical Principles for Medical Research Involving Human Subjects. *Jahrbuch Für Wissenschaft Und Ethik*, *14*(1), 233–238. https://doi.org/10.1515/9783110208856.233

Yu, A. J. (2014). Bayesian Models of Attention. In A. C. Nobre & S. Kastner (Eds.), *The Oxford Handbook of Attention*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199675111.013.025

Yu, A. J., & Dayan, P. (2002). Acetylcholine in cortical inference. *Neural Networks: The Official Journal of the International Neural Network Society*, *15*(4–6), 719–730. https://doi.org/10.1016/s0893-6080(02)00058-8

Yu, A. J., & Dayan, P. (2005a). Inference, Attention, and Decision in a Bayesian Neural Architecture. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17* (pp. 1577–1584). MIT Press. http://papers.nips.cc/paper/2548-inference-attention-and-decision-in-a-bayesian-neural-architecture.pdf

Yu, A. J., & Dayan, P. (2005b). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026