Visuospatial attention updates the cortical representation of sound under reduced sensory uncertainty

Cervantes Constantino, F., Sánchez-Costa, T., Cipriani, G. A.[†], Carboni, A.

Centro de Investigación Básica en Psicología; Universidad de la República; Montevideo, Uruguay
[†]Current address: Facultad de Psicología, Universidad Autónoma de Madrid; Madrid, Spain

Corresponding author:
Francisco Cervantes Constantino, Ph.D.
Centro de Investigación Básica en Psicología
Facultad de Psicología
Universidad de la República
Dr. Tristán Narvaja 1674
Montevideo 11200 Uruguay
Email: fcervantes@psico.edu.uy

Word count: 7184

Abstract

Our surroundings continually propagate audiovisual (AV) signals, and attention assists us in making a clear sense of them at any given time. Visual and auditory streams may serve jointly or separately as basis for selection. Mechanisms of attentional weighing have been suggested to be based on uncertainty estimation processes at the neural level. We examine how do reduced temporal and spatial uncertainty conditions facilitate the cross-modal transfer of selective biases by visuospatial attention on auditory processing. Auditory encoding of random tone pips, devoid of spatial information, was investigated as pips were associated to spatially informative visual contrast reversals (‘flips’). In a two-interval forced choice task, participants compared between AV sequences of different temporal uncertainty, while sustaining endogenous visuospatial attention over fixed foregrounds (full, half or quarter of the disc). Neural encoding of pips was addressed via a temporal response function model (TRF) of the participants’ auditory electroencephalogram (EEG) timeseries. Cross-modal modulations from visual input on auditory processing was evidenced at low temporal uncertainty trials, for quarter but not half visual foreground sizes. Evidence of the relatively late effect (~300 ms) was observed once pips were closely followed by flips, and transfer effect sizes were dependent on proximity to the visual target. The results suggest the presence of updating mechanisms over the ongoing neural representations of sounds, visually incorporating relevant spatial attributes for auditory stream segregation. The findings also illustrate the factoring of both bimodal and unimodal sources of uncertainty on the modulation of neural encoding dynamics by AV attention.

Keywords: audiovisual attention; uncertainty; multimodal integration; precision

1. Introduction

To parse an image carefully we have to shift focus on places that matter at any given time. Goal-directed spatial attention provides selective and efficient processing of visual regions at busy scenes. One key neural mechanism is prioritizing neural coding of specific retinotopic domains at the expense of others (Carrasco, 2014; O'Connor et al., 2002). As naturalistic scenarios often comprise multimodal information, it has been addressed how visuospatial selective effects may also manifest cross-modally (Koelewijn et al., 2010; Santangelo & Macaluso, 2012), however the details remain unclear. It is well established that in the production and perception of audiovisual (AV) human communication signals, the temporal proximity of auditory and visual streams determines multimodal integration (Spence & Squire, 2003; Holmes & Spence, 2005; Roseboom et al., 2009). For example, the unimodal processing of speech sounds can become systematically biased for coherence with joint visual face displays. These biases include the apparent spatial location shifts of the ventriloquist illusion, and inferred speech utterances in the McGurk effect (Andersen et al., 2009; McGurk & MacDonald, 1976; Miller & D'Esposito, 2005; van Atteveldt et al., 2007).

Mechanisms of visual attention include the spreading of a selection bias across diverse locations which belong to a foreground object (Egly et al., 1994; Hollingworth et al., 2012; Wannig et al., 2011). It has been addressed whether, for multimodal objects, similar "spillover" effects analogously prioritize input across object features now associated by temporal proximity (e.g. Busse et al., 2008; Degerman et al., 2007). In the unimodal case, object-based selection is considered to involve relatively later higher-level processing than visuospatial selection (Martínez et al., 2006; Müller & Kleinschmidt, 2003). Visuospatial attention has been long regarded as a driver for the binding of visual features shared by an object at a given location and processed in parallel, such as color, orientation and motion (Treisman, 1998; Treisman & Gelade, 1980). Whether visuospatial attention likewise drives binding during selective processing of concurrent multimodal stimuli remains unclear, yet temporal coherence between multimodal streams is considered to also play a significant role (Spence & Frings, 2020; Talsma et al., 2010). A possibility is that, under AV attention, multimodal object formation may be subsequent to visuospatial selection (Talsma & Woldorff, 2005). There is evidence that by limiting cross-modal temporal coherence one reduces the effectiveness of visuospatial attention to influence auditory processing (Beauchamp, 2019; King et al., 2019; Van der Burg et al., 2008). In addition, full temporal coherence with a dynamic visual stimulus has been shown to promote auditory segregation behaviorally (Maddox et al., 2015). It remains unknown how at the neural level do auditory populations organize to exert such changes under visuospatial selection criteria.

In hearing and vision, cortical mechanisms operate to represent foreground and background selections of competing sensory input. These mechanisms include the enhancement of activity representing attended auditory or visual input, and attenuation of that from distractors (Hillyard et al., 1973; Herrmann & Knight, 2001; Shinn-Cunningham, 2008). Modulatory dynamics of cortical encoding therefore offer a means to test for cross-modal 'transfer' effects during active selection of competing sources. Here, modulation of auditory processing was investigated when selection is given on visual terms. Because multimodal binding may still occur for non-synchronous auditory and visual streams, i.e. in temporal windows of effective multimodal integration (van Wassenhove et al., 2007), the role of AV temporal uncertainty was addressed. In contrast to temporally matched streams, these windows

imply that one modality leads or lags another in time. Importantly, this condition enables the examination of order or hierarchical effects that arise during cross-modal transfer, given an initial unimodal selection.

Here, a synthetic visual dartboard stimulus was created to display a dynamic video sequence of local contrast changes within disc sectors at random times. Tone pip presentations were paired to the sequence of visual events, by varying amounts of AV precision. This parameter was manipulated so that auditory onsets led and lagged visual ones by narrow (high AV precision) or wide (low AV precision) margins, in a probabilistic manner. Healthy participants performed a two-interval forced choice task, where they compared between AV precision levels across two stimuli presentations while attending only to indicated visual sectors. Using electroencephalography (EEG), the dynamics of sound processing under visuospatial attention was addressed, by means of the temporal response function model (TRF) of cortical encoding (Crosse et al., 2016; David et al., 2007). The TRF represents the linear mapping between the temporal dynamics of auditory 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 associated to one of competing auditory sources have been employed to estimate rapid enhancement and attenuation effects on encoding dynamics (Ding & Simon, 2012; O'Sullivan et al., 2015). Neural encoding of random tone pip sets (Cervantes Constantino et al., 2017) was thus examined, for evidence of and conditions leading to neural segregation by association to competing visual streams.

Uncertainty and multimodal attention

Rather than proximity or asynchrony parameters, we addressed uncertainty (i.e. the variance of a distribution of asynchronies) considering it represents an experimental approach to computational models of attention proposing gain mechanisms by uncertainty estimation (Dayan & Zemel, 1999; Yu & Dayan, 2005a; Whiteley & Sahani, 2008; Feldman & Friston, 2010a; Yu, 2014). These models propose that selective attentional biases are an outcome of weighing each competing neural representation of sensory input by estimates of its inherent perceptual noise, with attention serving the purpose of increasing perceptual reliability (Dayan et al., 2000; Dayan & Yu, 2003). Reliability estimation is considered a mechanism harnessed by neural computation to refine and drive perceptual selection (Whiteley & Sahani, 2012), and there is evidence of involvement of such strategies in fused audiovisual perception (Rohe et al., 2019; Rohe & Noppeney, 2015; Meijer et al., 2019). Nevertheless, evidence of 'precision-weighting' on sensory representations under multimodal attention is still unclear.

Here, TRFs were addressed in relation to two key uncertainty parameters: AV precision for its role in promoting binding, and visuospatial uncertainty since it is associated to unimodal selection. Upon AV integration, unimodal processing may be subject to considerable changes, in order to maintain consistency with a multimodal object representation (French & DeAngelis, 2020; Rohe et al., 2019). Therefore we hypothesized that reductions to both uncertainty parameters jointly facilitate changes to neural representations of sound observable in the TRF, according to visual input. In the present design, participants were asked to selectively attend over different positional uncertainty conditions, as they attended to either full, half, or quarter domain sizes of the dartboard. The TRF analyses then served to identify the conditions and stages at which the hypothesized cross-modal transfer effects emerge neurally from vision to hearing.

2. Methods

2.1 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 School of Psychology Research Ethics Committee at Universidad de la República approved the experimental procedures.

2.2 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 analysis was performed offline. Full experimental sessions lasted ~2 h.

2.3 Stimuli

Visual stream. Visual sequences consisted of a black and white checkerboard disc of 23 dva radius, uniformly divided into 20 concentric rings by 48 angular divisions, that was displayed in the centre of a black background. The resulting dartboard was grouped into sixty 4x4 checkerboard sectors distributed over 5 eccentricity levels and 12 angular divisions (e.g. Capilla et al., 2016). From 1 s onwards, individual checkerboard sectors underwent a series of polarity reversals, here termed ‘flips’ (Figure 1A), with nominal asynchronies of 101 ms plus an exponentially distributed random delay (μ = 99 ms). The asynchronies were subsequently discretized for correspondence with the video file frame rate (10 fps), resulting in an effective distribution ranging between 100 and 900 ms in step increments of 100 ms. Flip locations were uniformly sampled and balanced across eccentricities and disc quadrants; the number of flips in a given sequence varied by attentional condition (see *Audiovisual stimuli*, below). Checkerboards with contrast reversals remained static unless stimulated again later, resulting in non-simultaneous, spatially distributed visual transients (Van der Burg et al., 2010). Across a given sequence, flip location order was pseudo-randomized, and the final configuration was kept fixed for 0.5 s. Visual stimuli were constructed with the MATLAB® software package (Natick, United States), and stored in .avi format.

Auditory stream. Auditory sequences consisted of a series of pure tone sounds of 100 ms duration each

(Figure 1A) with selected frequencies from a fixed pool ranging 100 – 4846 Hz, separated by 2 equivalent rectangular bandwidth steps (Glasberg & Moore, 1990). The pips were modulated with 5 ms raised cosine on- and off-ramps, and amplitudes calibrated according to the 60-phon normal equal-loudness-level contour (ISO 226:2003) to adjust for perceived relative loudness differences across frequencies. The number of tones in the auditory sequence matched that of flips in the visual sequence, and each tone onset time was based on the nominal asynchrony calculated earlier for its corresponding visual flip, plus a uniformly distributed random shift (Figure 1A). The bounds of the zero-mean shift distribution depended on audiovisual precision condition for that sequence (see *Audiovisual stimuli*, below). Pip presentations could therefore lead flips (auditory precedes visual, 'ApV' conditions) or succeed them (visual precedes auditory, 'VpA' conditions) with equal probability. Unlike transient visual flips, pips were allowed to overlap in time. In addition, a frequency-eccentricity association was introduced, so that the lowest three pip frequencies (100, 171, 257 Hz) were reserved for flips located at the most peripheral eccentricity. Likewise, the highest three frequencies (3.2, 3.9, 4.8 KHz) were used only for flips at the most central eccentricity; intermediate frequencies were similarly arranged. Frequency values were always balanced across the auditory sequence.

Finally, an "auditory stimulus probe" was constructed separately. This stimulus consisted of a 15-tone sequence built under the same specifications as described above, but presented without an accompanying visual stream. This auditory-only stimulus was used to find reproducible auditory activity in the EEG (see *Spatial filtering*, below). The 15-tone sequence was quadruplicated and concatenated in time, lasting about 50 s. All auditory stimuli were constructed with MATLAB® at a sampling rate of 22.05 KHz and saved in .wav format.

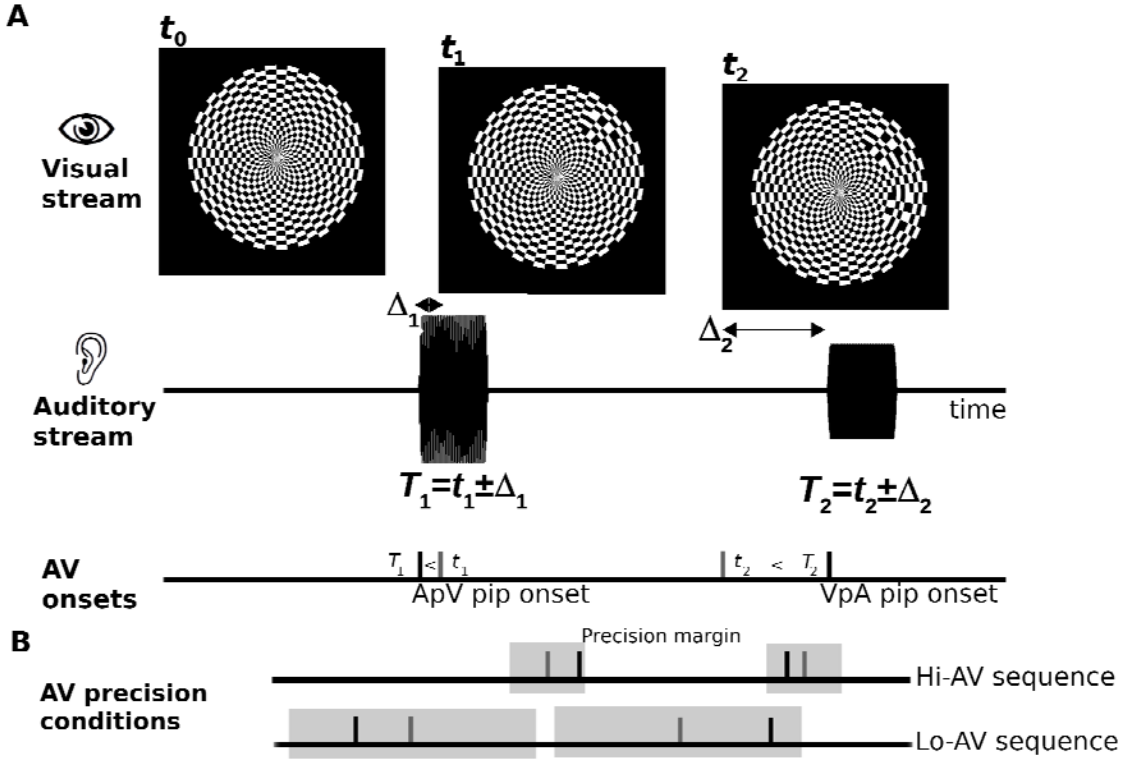

Figure 1. Basic audiovisual sequence and AV precision conditions. (A) Each audiovisual stimulus consists of a dartboard disk displaying a series of localized dynamic contrast changes ('flips'), in parallel to a stream of tone pip sounds. In this example, the first flip occurs in the upper right quadrant at t_1, and in the lower right at t_2 (top). For each flip, a pip is presented in the auditory stream (middle), with onset times T_i positively or negatively delayed at zero-mean uniformly distributed shifts Δ_i. Pips' auditory onsets precede the corresponding flip ('ApV') or viceversa ('VpA', bottom). (B) Delay distribution bounds are determined by the AV precision condition for a given stimulus sequence. The grey areas show the interval range of possible Δ_i values, which is fixed per stimulus sequence. High AV precision sequences ('Hi-AV') have narrower AV precision margins (top) than low AV precision sequences ('Lo-AV', bottom).

Audiovisual stimuli. Distribution of temporal shifts Δ_i between pip and flip onsets within a sequence depended on its AV precision condition between relevant audio and video streams across the stimulus sequence (Figure 1B). Ten AV precision margins were defined in the ±33 ms, ±66 ms, ±99 ms intervals and so forth up to ±330 ms. The AV precision conditions were balanced across each of the three attentional conditions required from participants tin the task. In the "Attend-All" (AA) condition, participants were asked to attend to the entire dartboard as a full foreground. Each AA stimulus consisted of a sequence of 15 pip-flip pairs, of variable duration lasting approximately 4 – 5 s in total. In the "Attend-Half" (AH) condition, participants were asked to attend to either the upper or lower dartboard hemifield as foreground, ignoring visual events from the opposite hemifield. Each AH stimulus consisted of a sequence of 30 pip-flip pairs, lasting approximately 7 – 8 s in total. In the "Attend-Quarter" (AQ) condition, participants were asked to attend to either the upper/lower, left/right dartboard quadrant as foreground, ignoring visual events from all other sectors. Each AQ stimulus consisted of a sequence of 60 pip-flip pairs, lasting approximately 12 – 15 s in total. This ensured that the number of to-be-attended events was constant across conditions (15 flips), while the number of events in the background varied by condition (AA: 0, AH: 15, AQ: 45). In addition, for AH and AQ

stimuli the AV precision level of the target sector was never matched to the AV precision level of the remaining disc surface, which prevented participant deployment of a generic “Attend-All” strategy. A fixed pool of 40 AA, 80 AH and 80 AQ stimuli was constructed for all participants. 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).

2.4 Task

At the start of a trial, subjects were indicated the sector they were required to attend. The spatial selection strategy required at a given trial was cued during initial fixation by a full circle (AA stimuli trials), an upwards[downwards] triangle (AH), or smaller triangle tilted diagonally towards a specific quadrant (AQ, Figure 2 inset). Two AV stimuli were subsequently presented in series, as part of a two-interval forced choice task (Figure 2, Video 1). Subjects were asked to compare AV precision between the first and second stimuli, and instructed to base their response on events from within the targeted visual sector only, and for both presentations. After presentations, subjects were prompted: “Which sequence was MORE [LESS] coordinated?” (here translated) with the alternatives shown (Figure 2). Subjects pressed a left/right arrow button to highlight their choice, and hit enter to validate. A feedback cue was then presented.

The pair combination of AV stimuli, question keyword (‘1st’ or ‘2nd’ stimulus, Figure 2) and response positions were randomized per trial, per subject. For any trial, AV precision across both presentations’ target sectors was never the same. A practice session consisting of 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, participants were asked to listen attentively and with their eyes closed to a brief (~50 s) auditory stimulus probe (see *Auditory stimuli*, above). During the main session, participants had an optional pause every 25 trials, and the experiment was finalized after 100 trials. Because responses were entered “until press”, participants had further opportunity to pause. After the main session, the auditory stimulus probe procedure was repeated. In appreciation for their time, participants received a chocolate bar or a cinema ticket.

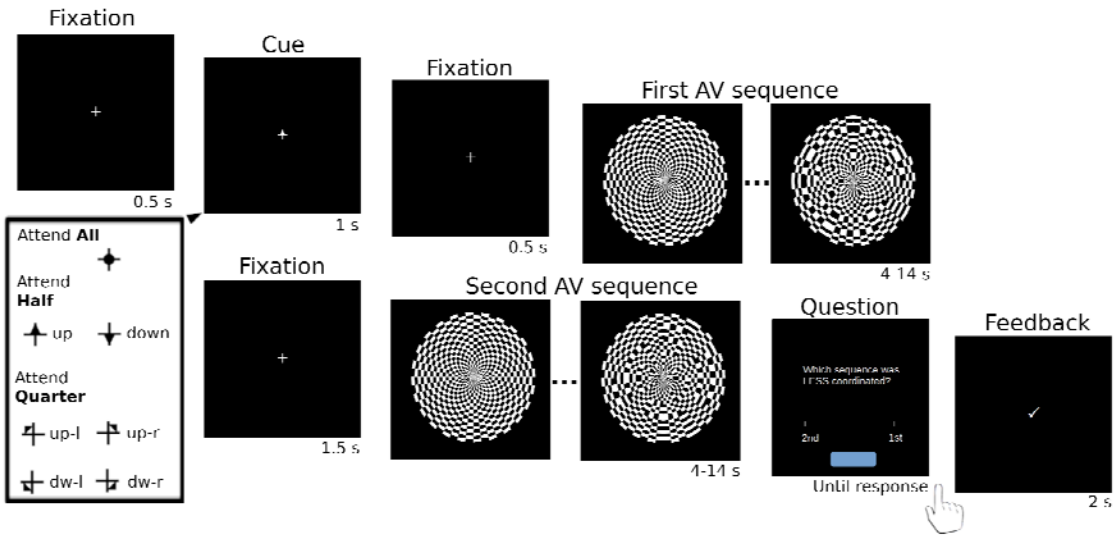

Figure 2. Audiovisual task. In a two-interval forced choice design, participants were asked which of two stimuli sequences was more coordinated audiovisually. An initial cue (inset) indicated the visual spatial attention strategy required for the trial. Participants were then shown both AV displays, and deployed visual spatial attention over the required sector for both sequences. In Attend-All they were asked to do so over the entire disc (full circle cue); for Attend-Half, over the upper/lower half (up/down triangle cues); for Attend-Quarter, only over one quadrant (tilted triangle cues). Participants had to compare between AV precision levels of attended sectors across both sequences. Once prompted, they were asked for a response, with trial feedback shown thereafter.

2.5 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 between -1 s and, 8 s (AA stimuli), 11 s (AH) or 17 s (AQ) of AV stimulus 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 across scalp sensor 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 (Hyvarinen, 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 and trial timeseries (64 x 200 = 12800), 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 equally pre-processed but without downsampling, and epoched into 8 trials.

Spatial filtering. To reflect auditory activity, and improve data SNR while reducing dimensionality, a spatial filter was constructed emphasizing signals that reflect reproducible activity across subjects during listening. The denoising by spatial filtering (DSS) (de Cheveigné & Parra, 2014; de Cheveigné & Simon, 2008b) data-driven procedure was applied on all 30 listeners' recordings during the "auditory stimulus probe" (see *Auditory stream*, above). First, all subjects' data were bandpass-filtered in the 1 – 8 Hz region 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 channel and trial timeseries now including all participants' data, and 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). In analogy to sensor selection in standard ERP studies, the DSS timeseries represents the most reproducible component of the evoked data, and was selected as a single virtual sensor in analyses henceforth. Subject DSS timeseries were downsampled to 32 Hz in order to reduce computational cost of TRF estimation, and potential temporal variability effects due to inter-subject and visual contribution factors into the final estimates.

Temporal response function (TRF) estimation. For AA conditions all single-trial data were concatenated in time. Separately, for all three attentional conditions, single-trial were concatenated according to AV precision as follows: a low AV precision group ('Lo-AV') that consisted of data from the bottom 5 AV precision levels; and a high AV precision group ('Hi-AV') corresponding to the top 5 levels. EEG concatenated recordings were paired to their corresponding concatenated stimuli timeseries. The auditory stimuli timeseries corresponded to tone pip onset mappings, namely, a sparse representation of the stimulus consisting of a vector equal to 1 at the onset time T_i of any given pip, and zero elsewhere (Cervantes Constantino et al., 2017):

$$S(t) = \begin{cases} 1, t = T_i \\ 0, t \neq T_i \end{cases}$$

To examine for hierarchical effects due to leading auditory/visual input on auditory encoding, auditory stimuli timeseries were subsequently divided by the sign of the effective asynchrony between auditory onsets T_i versus visual onsets t_i:

$$S_{VpA}(T_i) = \begin{cases} 0, T_i < t_i \\ S(T_i), T_i > t_i \end{cases} \qquad S_{ApV}(T_i) = \begin{cases} 0, T_i > t_i \\ S(T_i), T_i < t_i \end{cases}$$

so that $S(t) = S_{VpA}(t) + S_{ApV}(t)$. Except for AA conditions, to reflect attentional selection each $S_{VpA}(t)$ and $S_{ApV}(t)$ timeseries were similarly sub-divided further according to whether the associated flips were located in attended versus unattended visual halves (for AH), or opposing quadrants (AQ). For AQ, additional auditory stimuli timeseries were obtained from the two visual flanking quadrants that were also part of the AQ background.

For each attentional, AV precision, precedence order, and foreground/background condition, EEG subject data and corresponding auditory stimulus representations were used to estimate the temporal response function (TRF). The linear TRF is a mapping between an auditory stimulus representation such as $S(t)$ and the evoked neural response $r(t)$ it elicits. In this example, the 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 under the conditions set by a chosen stimulus representation $S(t)$ for the same period of neural data. TRFs were estimated by boosting (David et al., 2007) between stimulus and EEG timeseries scaled to *z*-units, with 10-fold cross-validation, and the kernel spanned the -70 to 530 ms window post pip onset approximately. Grand average TRFs were obtained by averaging individual TRFs across subjects.

2.6 Statistical analysis

Experimental conditions, i.e. attentional foreground (AH, AQ), AV precision (high, low), and precedence order (VpA, ApV), were addressed for evidence of cross-modal effects after visuospatial selection across attended versus unattended domains. For this, time intervals indicating differential auditory encoding were obtained by non-parametric testing (Maris, 2012; Maris & Oostenveld, 2007) corrected from multiple comparisons. *t*-value timeseries were estimated from the 8 attentional contrasts, and an a priori threshold corresponding to the *t*-distribution's 99.375th percentile was used for randomization-based testing. Any cluster exceeding the threshold was deemed significant if its associated *t*-statistic (sum of *t*-values within the cluster) exceeded those within the randomization distribution at α=0.05/8 level of significance. The distribution was generated by $N=2^{17}$ resamplings of attentional conditions randomly shuffled per participant prior to grand-averaging. Post-hoc analyses to address enhancement effects by attention on TRF peak processing were performed with one-sided Wilcoxon signed rank tests at α=0.05 level of significance, with corrections for multiple comparisons where indicated.

3. Results

The global correct response rates of participants in the task were significantly above chance level, as shown by a Student's *t*-test (mean ± SD 67±9%; $t(29)=4.3$; $p=1.8\text{x}10^{-4}$). Performance differences between Attend-All, Attend-Half and Attend-Quarter conditions were non-significant (mean ± SD 70±14% AA ; 67±11% AH; 65±9% AQ), as shown by a repeated-measures ANOVA of subjects' performance scores by attentional condition ($F(2)=2.69$; $p=0.076$).

To probe the effect of visuospatial selection on auditory processing in the task, first, a data-driven spatial filter emphasizing reproducible auditory activity across all participants was applied. Its resulting central scalp activity topography (Figure 3A) appears consistent with auditory sources (Stropahl et al., 2018). First, auditory encoding during joint auditory and visual presentations in AA trials was

addressed via the temporal response function method. A generic biphasic response profile was observed with a positive and a negative peak in the 150-300 ms interval post pip onset (Figure 1B). This biphasic TRF was then examined in relation to temporal AV precision and unimodal precedence order. For this, TRF estimates were obtained based on stimulus representation timeseries containing only pips preceding (ApV) or succeeding (VpA) visual flips, separately for high AV ('Hi-AV') versus low AV ('Lo-AV') precision sequences, in the same AA trials. For Lo-AV, the onset delay distributions have a median 128 ms, interquartile range (IQR) 128 ms, and range 0 - 330 ms. For Hi-AV, they have a median 39 ms, IQR 56 ms, range 0 - 165 ms fo Hi-AV. ApV and VpA onset delay distributions are accordingly similar and of opposite sign (Figure 1C). Qualitatively, the AA TRF estimates similarly show an the early positive peak in the 150-200 ms interval again, and a later negative peak in the 250-300 ms is most salient for tones closely followed by visual flips in the Hi-AV, ApV condition.

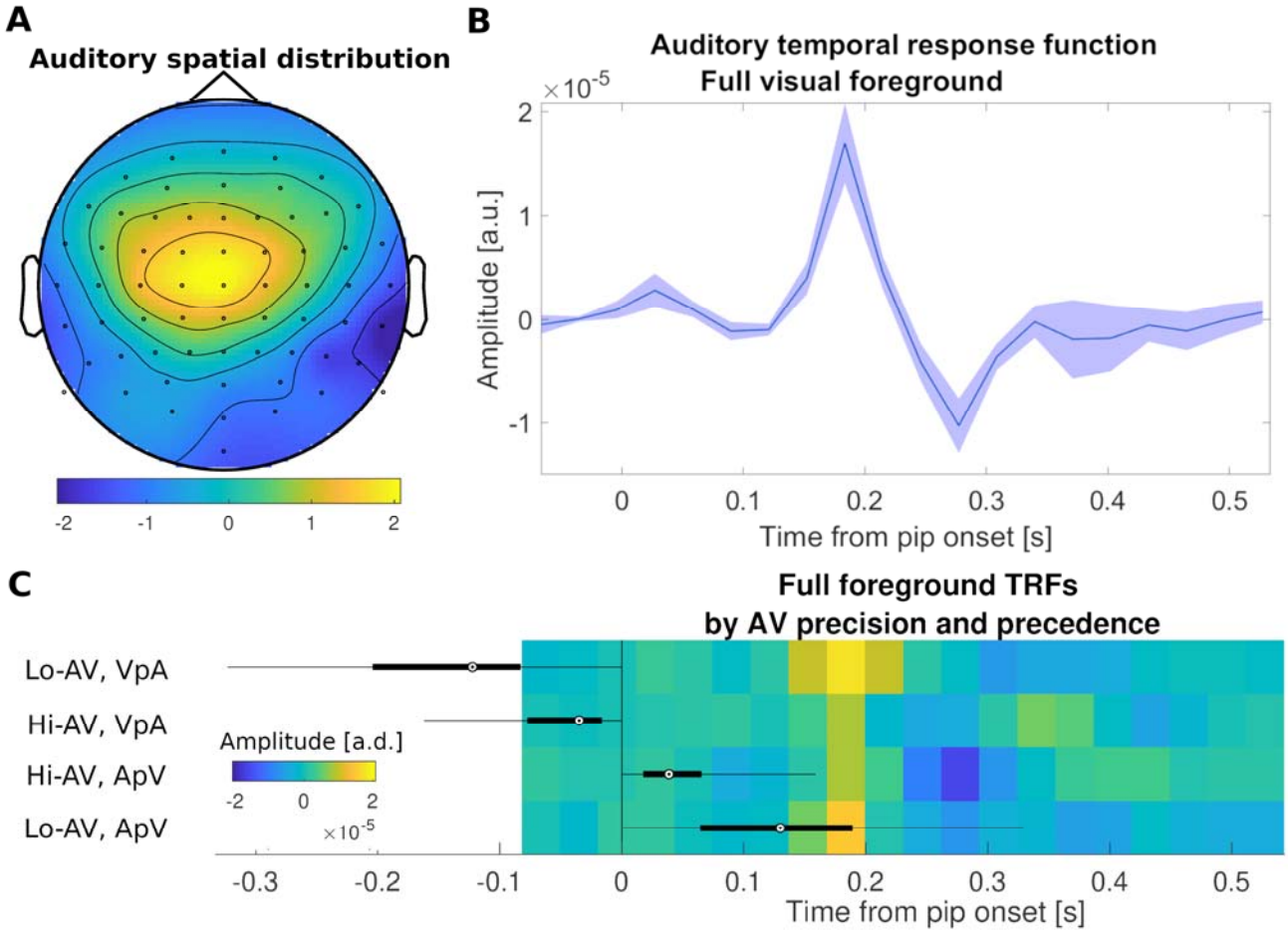

Figure 3. Pip encoding within concurrently attended visual presentations. (A) EEG topography of top reproducible auditory activity across all participants. All sensor recordings were projected onto this auditory component via spatial filtering, which served for TRF estimation. (B) The EEG TRF grand-average estimated across all pip presentations in Attend-All trials (full foreground) displays a biphasic profile. Surface area corresponds to ±1 standard error of the mean. (C) TRF estimation in Attend-All by AV precision and precedence order. Four grand average TRF estimates, with amplitude now shown by depth, after High- versus Low-AV precision trials where individual pips are also separated by precedence order. Boxplots indicate corresponding

flip times. The earlier TRF peak, locked to the pip onset, can be observed across conditions. The late TRF peak appears most clear as flips occur early on during ongoing pip processing (Hi-AV, ApV).

Since a goal of analyses is to address AV competition leading to visuospatial selection on auditory processing, TRFs of pips paired to the visual foreground versus background from AH and AQ trials were estimated. The auditory stream timeseries was then separated according to the correspondence of pips to attended/unattended visual sectors and, similarly, by precedence order between sound and visual events, and AV precision level corresponding to the attended visual sector (Figure 4A). For pips corresponding to the ongoing background, i.e. opposite hemifield or opposite quadrant (Figure 4B,C), TRF estimates were separately obtained. Contrasts were used between TRF estimates from pips associated to 'Attended' versus 'Unattended' visual sectors, to address differential auditory encoding. Non-parametric permutation testing corrected for multiple comparisons revealed a significant difference by visuospatial selection on the auditory TRF in the 262-324 ms post pip onset range (p=0.002). A greater peak amplitude was observed for pips associated to 'Attended' than to 'Unattended' visual sectors. The effect was observed only at AQ trials and for pips closely followed by flips (Hi-AV, ApV), and corresponded to the second TRF peak stage (Figure 4A).

To exclude the alternative that, rather than visuospatial selection, the observed effect is potentially driven by differences in AV precision between foreground and background sectors, we additionally estimated the background TRF with matching Hi-AV precision conditions as in the foreground. For this, all AQ trial ApV data where Unattended visual sectors were of high AV precision were used for TRF estimation, regardless of whether these presentations were concurrent to those used for estimation in Attended sector presentations. TRF differences therein indicate that changes to auditory representations of pips by visuospatial association may generally hold for pips across opposite sectors, and thus independent of the particular foreground-background dynamics context in a trial. A one-sided Wilcoxon signed-rank test of the mean TRF amplitude within the temporal cluster detected earlier showed a significant greater TRF amplitude related to the 'Attended' visual sector estimated earlier (Figure 4A and Figure 4D, blue) versus the 'Unattended' sector now matched for Hi-AV precision (Z= -3.06, p=0.001) (Figure 4D, orange). The latter appears similar to the concurrent 'Unattended' TRF estimated earlier (Figure 4A and Figure 4D, red).

Since, by design, both attended and unattended auditory streams in AH and AQ trials were allowed to overlap, temporal associations other than the nominal AV pairings may occur, particularly at low AV precision. To address this, the TRFs were estimated again on pip-flip pairings based on the same temporal rank order (first pip with first flip, and so on), regardless of associations such as tone frequency and visual eccentricity. This analysis similarly showed a significant difference in TRFs at Hi-AV, ApV for AQ trials, in the 262 – 324 ms post pip onset range (Z= -2.74, p=0.006) without similar effects in the rest of conditions. To discard the possibility that the observed differences in AQ but not in AH were mainly driven by attentional drift rates over shorter AH versus longer AQ presentations (i.e. 30 vs 60 pip-flip events), only data from the first half of each AQ trial were used for TRF estimation anew. A one-sided Wilcoxon signed-rank test of the mean TRF amplitude of this shorter dataset similarly showed a significant modulation to the second TRF peak amplitude, which increased in 'Attended' than 'Unattended' sectors at the 293 – 324 ms window (Z= -2.47, p=0.007).

Are pip representations subject to visuospatial selective biases gradually, as they associate closer or away from a target visual sector? To address this, pip encoding at AQ conditions was examined in relation to the intermediate Unattended sectors that are not the target nor its opposite quadrant (i.e. same abscissae and ordinates, Figure 4C). TRF estimates from these quadrants illustrate that the second TRF peak that was earlier found subject to visuospatial attentional modulation, also encodes pips at intermediate amplitudes for these flanking quadrants (Figure 4E). A one-sided Wilcoxon signed-rank test of the mean TRF amplitude between both flanking quadrants showed a significant greater TRF peak amplitude for pips on target (Figure 4E,D, blue) versus pips on flanking quadrants (Figure 4E, yellow and purple) (Z=-2.03, adjusted p=0.042, Bonferroni corrected), in the 262 – 293 ms window. In addition, the average TRF peak amplitude for pips on flanking quadrants was significantly greater than that of the opposite quadrant (Figure 4E,D, red) (Z=-2.13, adjusted p=0.034), for the same window.

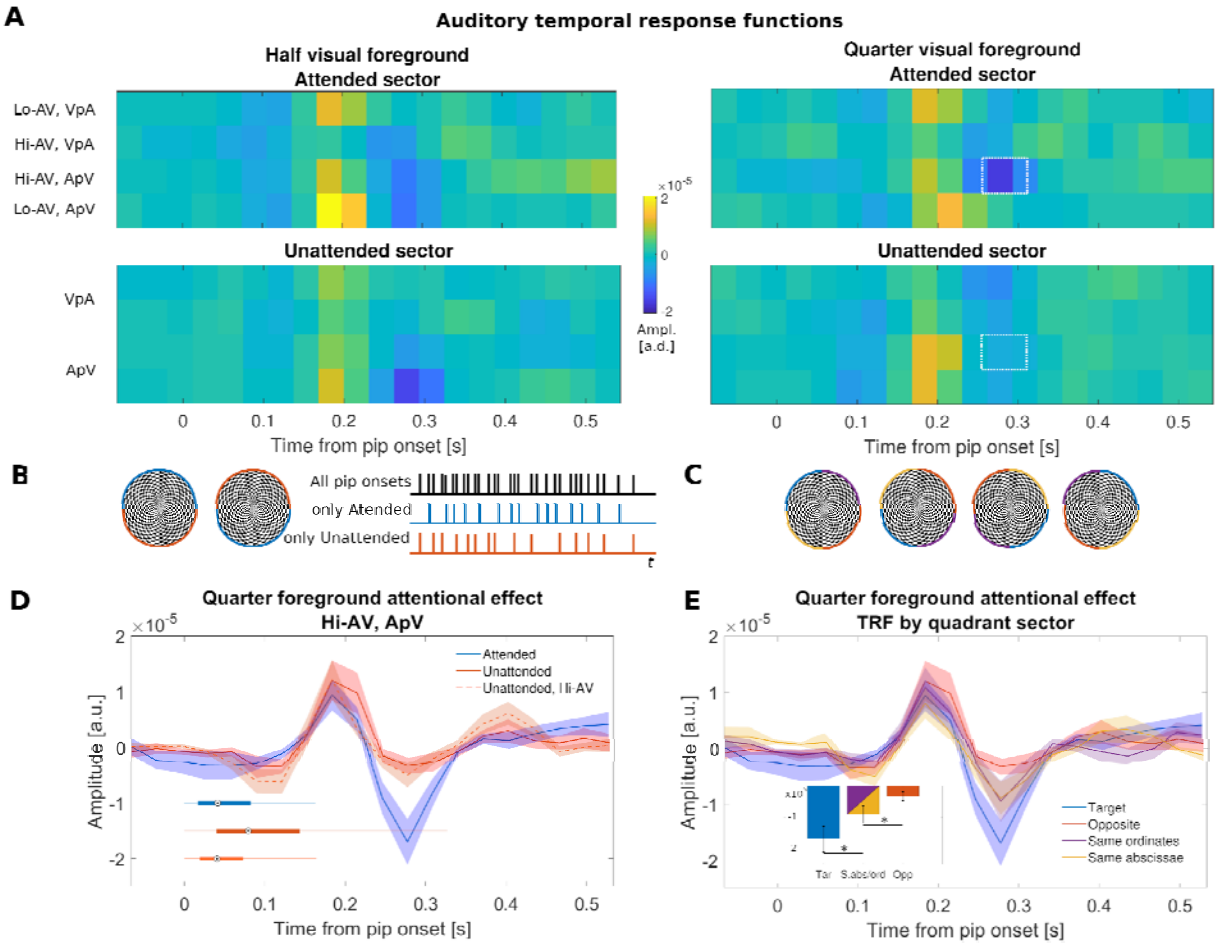

Figure 4. Differential processing of competing auditory stimuli by visuospatial criteria. (A) Grand average TRFs from Attend-Half (left) and Attend-Quarter trials (right). When pips are associated to a high AV precision and quarter foreground, a crossmodal effect of selective attention is observed in the 250 – 310 ms interval as tones precede visual input. (B) For illustration, in AH pips belong to one of two hemifields and TRF estimates are obtained from onset representations of the auditory stream, containing only pips within the attended or unattended hemifield. (C) In quarter foregrounds, the unattended opposite quadrant ('Unattended' in A,D) is also flanked by two other unattended quadrants at any given presentation, here indicated by color code. (D) TRFs where cross-modal transfer was observed, displayed again along with accompanying temporal shift distributions of visual onsets. Unattended pips entail a comparatively lower AV precision, however controlling for this across non-simultaneous presentations reveals a similar TRF (Unattended Hi-AV, orange). Blue and red are as in A. (E)

Comparison of target, opposite, and flanking quadrants (target/opposite as in D) shows an intermediate degree of representation consistent with spatial selectivity (inset). Blue and red as in A,D.

4. Discussion

In the present study, participants judged AV displays to determine temporal uncertainty differences between visual and spatially uninformative auditory elements, while restricting attention over selected locations of the visual scene. The resulting emergence of a cross-modal bias from visual spatial attention over auditory neural encoding was identified at around 300 ms post audio onset. This transfer effect was found under reduced bimodal temporal and unimodal visuospatial uncertainty conditions. Since it was observed with pips' lead, but not lag, the involvement of visual updating over ongoing auditory processing is indicated. Results also suggest that changes to representations of sound may occur according to their likely relevance for auditory segregation. Since relevance was determined visuospatially, transfer is attributed by the temporal and inferred spatial proximity of a sound to the attended region.

Cross-modal modulations on auditory processing by visual updating

Cross-modal binding and multimodal attentional selection both involve the exchange of information across representations of features shared by a same object (Bizley et al., 2016; Donohue et al., 2011; Spence & Frings, 2020; Talsma et al., 2010). The analyses show neural auditory encoding is susceptible to cross-modal modulation under unimodal visuospatial selection, at a late timeframe which appears posterior to automatic AV integration (Koelewijn et al., 2010; Giard & Peronnet, 1999; Kayser et al., 2007; Noesselt et al., 2007). In unimodal visual scenarios, goal-directed spatial attention effects are typically early observed about 100 ms post visual onset (Luck & Kappenman, 2012). Additional evidence from simultaneous AV presentations indicate crossmodal interactions from auditory input on visual areas may occur by 70 ms (Rohe et al., 2019; Shams et al., 2005). These two time frames then suggest a 70 – 100 ms window where, in a synchronous AV presentation, early cross-modal effects would be exerted from auditory to visual regions prior to visual attentional deployment. It is nevertheless unclear how early preattentive cross-modal effects similarly manifest in non-simultaneous AV presentations. Here, at high AV precision, ApV presentations carry a typical 40 ms delay so that unimodal visuospatial attentional effects may be expected, at visual areas, approximately around 140 ms with respect to the leading pip onset time. Thus the present ApV conditions may typically convey sufficient time for early cross-modal effects to be induced on visual areas by the leading sound. In contrast, VpA presentations even at high AV precision represent a substantially limited window for analog interactions at visual regions since flips are offset by a 40 ms delay until pips arrive. The results show that only for ApV conditions was auditory processing effectively modulated by visual input. Changes to the late but not early cortical representation of the leading auditory event appears consistent with setting of the auditory event as a one feature of a multimodal object. The effect latency (280 to 300 ms), even accounting for the typical bimodal asynchrony here, is posterior to observations of attentional spread in object-based vision, around 180 ms (Martínez et al., 2006), suggesting the involvement of high-order non-modality specific associative processes, e.g. the P3b. Because selection in the task stems unimodally from vision, the finding also supports the role for visual updating mechanisms in auditory scene analysis under competitive AV selection by visuospatial criteria (Koelewijn et al., 2009; Roseboom et al., 2009). In other words, for stream segregation to be

instantiated neurally, the status of spatially uninformative pips may be updated online in order for pip representations to become spatially organized, as targets according to the focus of visual attention.

Due to its temporal processing advantages, hearing is well suited for rapid warning to engage or redirect visuospatial attention (e.g. Spence & Driver, 1997; Fujisaki et al., 2004). A leading sound stimulus may serve to 'tag' time instances to enhance attentional deployment at specified visual sectors, while subsequent visual recognition feedback prepares auditory processing for perceptual binding (Noesselt et al., 2008; Shams et al., 2005; Van der Burg et al., 2008, 2011). Future studies may conversely address the role of auditory updating on visual representations, and of judging AV precision under auditory selective constraints. Hierarchically, it is predicted that cross-modal updating may take place on the neural representations of those unisensory features that do not initially carry relevant selection information.

Emergence of cross-modal attentional transfer with reduced uncertainty

For ApV conditions the relevance of an individual pip is initially uncertain, and cross-modal effects were observed for AQ but not AH foregrounds. An interpretation of this that in AH pips have individually less efficient roles in tagging times to prompt visual updating (e.g. Olivers & Van der Burg, 2008; Quak et al., 2015). AH visual flips entail a higher chance to pair on target than those in AQ a priori, an advantage that may also be accentuated by the effectively faster presentation rates at AQ, i.e. 30 in AH versus 15 relevant flips over presentations that are twice as long in AQ. Flips are stimuli transients that are distributed across spatial channels. Channel size is known to impact serial spatial attention processing (Fujisaki et al., 2006), while narrow spans may facilitate detection given their lower perceptual load (Lavie, 1995). The results then suggest a facilitation of attentional transfer by reduced unimodal positional uncertainty (Andersen et al., 2009), rather than reliance on prior auditory expectations in AH.

In the task design, to ensure a correct attentional strategy, AV precision levels of the unattended sector were different from those in the attended sector. A potential explanation of the results may then be that the effects correspond differential bimodal integration across the sectors rather than attentional transfer. There is evidence however that, outside of the focus of visuospatial attention, synchronized auditory and visual streams of distractors simultaneous to an overtly attended target will be less likely to integrate bimodally (Jensen et al., 2019). This would suggest that pip/flip pairs are unlikely to interact cross-modally at unattended regions regardless of AV precision level. Indeed, in analyses to control for AV precision at background sectors, it was observed that even when unattended events were not concurrent to those in the foreground, the neural representation of tone pips was still changed by AV association. This observation suggests that auditory stream segregation was visually driven in a top-down manner, and irrespective of the particular dynamics or contingencies of context in a total auditory scene at a given trial (Rahne et al., 2007).

Precision-weighting and multimodal integration

In naturalistic settings, multimodal stimuli do not have to be simultaneous in order to be inferred as part of a single source (van Wassenhove et al., 2007). Rather, perceptual organization of AV objects appears to admit some degree of temporal misalignment in a relatively flexible manner. The present work extends upon evidence of cross-modal binding and influence from visual dynamics onto auditory stream segregation by means of perfect temporal matching (Bizley et al., 2016; Maddox et al., 2015)

(Maddox et al., 2015). At the neural coding level, our results identify an attentional transfer process that is compatible with binding and which assists auditory segregation in competitive AV scenarios. Cross-modal attentional transfer was facilitated under reduced physical stimulus uncertainties involved at both the unimodal selection and bimodal interaction stages.

Ideal-observer computational frameworks of attention propose that noise is actively used by perceptual systems to weigh sensory data for selection (Hesselmann et al., 2010; Macaluso et al., 2016; Rohe & Noppeney, 2018; Yu, 2014). In these models, the brain holds representations of sensory input and pairs (or normalizes) them with representations of their uncertainty, as inferred from sensory causes in the environment (Yu & Dayan, 2002, 2005b). Attentional selection therein emerges by inversely factoring sensory representations by their relative uncertainty. Therefore, enhancement to a sensory representation may emerge by changing input itself externally, or by reducing its associated uncertainty. Uncertainty is concerned with the reliability of external sensory variables for the brain's own internal perceptual inference. While the present results do not demonstrate weighing to be computed by uncertainty estimation directly, judgment of temporal precision was actively required from participants in the task. For top-down attentional transfer from vision onto auditory representations, the results suggest attentional weighting involves factoring of at least these two relevant uncertainty parameters. The findings suggest that visuospatial transfer was facilitated by jointly reducing both terms so that auditory encoding of tones that were originally devoid of spatial selectivity became updated as elements of an attended multimodal object, indicating re-weighing for auditory neural segregation (Beauchamp et al., 2004; Noesselt et al., 2007; Talsma et al., 2010).

Where does cross-modal binding stand in relation to attentional transfer? As discussed earlier, simultaneous AV presentations could allow for early, possibly preattentive, cross-modal influences from auditory onto visual processing (Shams et al., 2005). Arguably, such effects may still occur in ApV scenarios if integrative processes are triggered automatically and with sufficiently low temporal asynchrony. However such early interactions would not necessarily constitute 'binding' to support AV object formation (Bizley et al., 2016). In the case of unimodal vision, it has been long proposed that spatial selection is a building block for the binding of object features linked by a common location (Spence & Frings, 2020; Treisman, 1998), i.e. a precise position in space. With cross-modal features, an analog role for time is considered; thus it is of interest whether both spatial and temporal terms can be put in a common statistical framework in multimodal feature binding. The results support the notion that such framework may be based on distributional parameter such as variance. It is of further interest whether uncertainty estimation is followed by cross-modal attentional transfer and binding processes in parallel, or whether the latter two occur serially. For instance, some evidence suggests that attentional weighing may precede forms of cross-modal integration, as differential weighing on the representation of tones may directly assist robust perceptual integration later on (Ernst & Bülthoff, 2004; Rohe et al., 2019; Spence & Ngo, 2012). Additional evidence suggests that both auditory cross-modal integration and attentional mechanisms operate on partially overlapping circuitry along the depth axis of the auditory cortex (Gau et al., 2020).

5. Conclusion

Optimal attentional selection relies on adopting those statistical properties of natural signals that are relevant for a task. Ideal observer models of attention consider weighing by the neural variability

associated to external state variables, both of which may be encoded in tandem by populations in sensory cortex (Walker et al., 2020). One key uncertainty parameter in multisensory integration is in the temporal relation between unimodal streams. For an audiovisual task based on unimodal visuospatial selection, it is shown that neural representations of sound may attain spatial attributes under increased temporal, but also visuospatial, precision conditions. The effects are observable asymmetrically at a late stage of sensory processing suggesting that, in auditory responses, visual updating rather than visual priming mechanisms weigh in AV object formation and auditory stream segregation.

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.

Competing interests statement

Declarations of interest: none

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

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

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

Bizley, J. K., Maddox, R. K., & Lee, A. K. C. (2016). Defining Auditory-Visual Objects: Behavioral Tests and Physiological Mechanisms. *Trends in Neurosciences*, *39*(2), 74–85. https://doi.org/10.1016/j.tins.2015.12.007

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

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

Ding, N., & Simon, J. Z. (2012). 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

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

French, R. L., & DeAngelis, G. C. (2020). Multisensory neural processing: From cue integration to

causal inference. *Current Opinion in Physiology*, *16*, 8–13. https://doi.org/10.1016/j.cophys.2020.04.004

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

Hesselmann, G., Sadaghiani, S., Friston, K. J., & Kleinschmidt, A. (2010). Predictive coding or evidence accumulation? False inference and neuronal fluctuations. *PloS One*, *5*(3), e9926. https://doi.org/10.1371/journal.pone.0009926

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.

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

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

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

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology. Human Perception and Performance*, *21*(3), 451–468. https://doi.org/10.1037//0096-1523.21.3.451

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

Maddox, R. K., Atilgan, H., Bizley, J. K., & Lee, A. K. C. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. *ELife*, *4*. https://doi.org/10.7554/eLife.04995

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

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
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
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
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
Shams, L., Iwaki, S., Chawla, A., & Bhattacharya, J. (2005). Early modulation of visual cortex by sound: An MEG study. *Neuroscience Letters*, *378*(2), 76–81. https://doi.org/10.1016/j.neulet.2004.12.035
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
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, Charles, & 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, Charles, & 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
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
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
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
van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598–607. https://doi.org/10.1016/j.neuropsychologia.2006.01.001
Walker, E. Y., Cotton, R. J., Ma, W. J., & Tolias, A. S. (2020). A neural basis of probabilistic computation in visual cortex. *Nature Neuroscience*, *23*(1), 122–129. https://doi.org/10.1038/s41593-019-0554-5
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
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

Video legend

Video 1. Sample Attend-All trial in the 2-IFC task. After the initial cue to attend the full foreground, the two AV sequences are presented serially. In this example, the first stimulus is built with lower AV precision than the second. In AA trials, the participant is asked to always consider in both cases all pip-flip pairings across the disc. After the second presentation, participants respond which of the two sequences had less or more temporal coordination. Upon validating the response, feedback is shown. An example selection of the correct response is represented here with a hand icon.