Stimulus specific and configurational features of multiple visual stimuli shape multisensory binding

Christoph Kayser [a], Nienke Debats [a], Herbert Heuer [a,b]

a. Department of Cognitive Neuroscience, Universität Bielefeld, Bielefeld, Germany

b. Leibniz Research Centre for Working Environment and Human Factors, Dortmund, Germany

Contact: Christoph Kayser, Universitätsstrasse 25, 33615 Bielefeld, Germany, christoph.kayser@uni-bielefeld.de

Keywords: multisensory, cross-modal, audiovisual, multisensory causal inference, temporal asynchrony

Abstract

Studies on multisensory perception often focus on simplistic conditions where one signal is presented per modality. Yet, in everyday life we encounter multiple signals per modality that may be associated with each other. Using the spatial ventriloquism paradigm as model, we tested how the presence of two visual stimuli shapes sound location judgements. Across three experiments we systematically paired one sound with either one or two visual stimuli that differed in relative timing and spatial offset to the sound. We focused on two biases typically studied in this paradigm, the ventriloquism effect and the trial-wise aftereffect. Our results show that these are not shaped by a single visual stimulus in a winner-takes-all manner but rather by their collective influence. In particular, the contribution of individual visual stimuli to the ventriloquism bias depends on their own relative timing to the sound and the temporal context provided by the other visual signal. Depending on this context, two visual signals influence the collective bias with comparable or asymmetric weights, which we attribute to the temporal evolvement of evidence for sensory causal relations during the individual trials. Concerning the aftereffect, we show that this emerges following both auditory and visual judgements in the preceding AV trials and following asynchronous AV stimulus pairs, confirming the notion that this reflects an automatic adjustment to spatial discrepancies.

Introduction

Multisensory perception involves the binding of stimuli across sensory modalities. In real life, we usually encounter multiple cues in each modality that can belong to the same or distinct objects. For example, while hearing the hiss of a snake we may spot movements at multiple locations on the ground, only one of which gives cues about the snake's location. This raises the question of how the presence of more than one stimulus per modality influences multisensory perception. Previous work suggests that the manner in which multisensory signals are combined in general depends on their spatio-temporal configuration, semantic attributes, attentional focus and task relevance and is generally guided by the evidence about sensory causal relations between individual signals (Talsma et al., 2010; Noppeney, 2021; Shams and Beierholm, 2022). However, still little is known about how multiple signals from one modality are combined with one signal from another, even in hallmark paradigms such as the spatial ventriloquism effect (Howard and Templeton, 1966; Bertelson et al., 2000; Bruns, 2019).

To understand how we combine two visual signals with one sound across space and time in the spatial ventriloquism paradigm, we devised a series of three experiments. In the multisensory trials of these experiments one sound could be accompanied by one or two visual stimuli that differed in their relative timing to the sound (preceding, synchronous or delayed) as well as their spatial offset to the sound, and either modality could be task relevant. One possibility is that in this setting only one visual signal enters (captures) the binding process, for example the one most proximal or synchronous with the sound (Soto-Faraco et al., 2002; Bruns and Roder, 2010; Krause et al., 2012). Spatio-temporal proximity generally indicates a causal relation between sensory stimuli and facilitates the capture of crossmodal attention (Kording et al., 2007; Spence and Santangelo, 2009; Macaluso, 2010; Talsma et al., 2010). This is for example seen in the *pip and pop effect,* where a single sound is bound with one out of many visual stimuli based on its temporal alignment (Van der Burg et al., 2008b, a). Alternatively, perception could combine all stimuli in a weighted manner, as typically reported for the reliability-weighted integration of two isolated signals in each modality (Ernst and Bulthoff, 2004; Angelaki et al., 2009). Thereby, the weights attributed to each visual stimulus could be the same regardless of their temporal separation, or could depend on both their spatial and temporal distances to the sound. Importantly, the relative influence of each visual stimulus may depend only on its own cross-modal relation to the sound, or may also depend on the overall context provided by all visual signals. We here compared these alternatives based on dedicated experimental manipulations and the modelling of participants' response biases.

So far, few studies have investigated multisensory perception in similar settings, that is, when two stimuli in one modality are presented together with one in another. Using an experiment in the visuo-motor domain, we have tested how the presence of two visual cues affects judgements about the proprioceptively-felt hand position (Debats et al., 2021). In that study participants made center-out hand movements and were presented with two spatially- and temporally-offset visual signals that provided discrepant evidence about their movement direction. Participants' judgements about movement endpoints were shaped by both visual signals, pointing to the weighted influence of multiple signals within and across modalities. In the audio-visual domain, previous work showed that the interaction between multiple visual stimuli and one sound is shaped by semantic congruency (Kanaya and Yokosawa, 2011; Shavit-Cohen and Zion Golumbic, 2019), temporal synchronization (Lewkowicz et al., 2021) and prior exposure to these stimuli (Tong et al., 2020). Furthermore, in a spatial ventriloquism experiment with two visual stimuli flanking one sound at equal spatial distances, the more salient visual signal dominated the binding process (Bertelson et al., 2000). However, when both visual signals were equally salient their effects cancelled, providing support for the notion that multiple signals within a modality collectively shape multisensory binding.

To probe how two visual signals shape perceptual judgements of a sound we manipulated the spatio-temporal alignment between visual and auditory signals in two ways. First, in two experiments we placed two visual stimuli that were presented at different onset times relative to the sound at identical distances

but opposite directions from the sound. This allowed us to directly probe whether they exert comparable influences on sound localization judgements by contrasting the condition-wise biases. In a third experiment we manipulated the relative spatial discrepancies of the two visual stimuli independently, so that they could feature discrepancies of same or distinct magnitude and same and distinct directions. This allowed us to directly probe their relative weight by modelling their influence on trial-wise judgements. To provide a comprehensive assessment of multisensory perception we probed two perceptual biases that are typically studied in this experimental paradigm: the ventriloquism effect, describing how visual and acoustic evidence in a given audio-visual trial is combined to judge the location of these stimuli (Bertelson and Radeau, 1981; Rohe and Noppeney, 2015), and the immediate (trial-wise) ventriloquism aftereffect. This reflects a bias in the judgement of unisensory auditory stimuli that are presented following a multisensory trial and is considered an instance of crossmodal adaptation or learning (Radeau and Bertelson, 1974; Recanzone, 1998; Bruns et al., 2011; Park and Kayser, 2019, 2021).

METHODS

Participants

This study was approved by the ethics committee of Bielefeld University. A total of 95 adult participants was recruited from the local student population and submitted informed consent prior to entering the study. Participants had self-reported normal vision and hearing, indicated no history of neurological disorders and were compensated for their time. The data were collected anonymously and it is possible that some individuals participated in more than one of the experiments described below. For each experiment we aimed for a sample size of at least 20, following previous studies using similar experimental protocols (Park and Kayser, 2019; Park et al., 2021; Park and Kayser, 2022) and recommendations for sample sizes in experimental psychology (Simmons et al., 2011).

General experimental setup and paradigm

The paradigms were based on a single-trial audio-visual localization task in which the individual trials and conditions were designed to probe the ventriloquism effect and the immediate (trial-wise) ventriloquism aftereffect (Wozny and Shams, 2011b; Park and Kayser, 2019, 2021). Participants were seated in an echo-free booth (E:Box, Desone) in front of an acoustically transparent screen (Screen International Modigliani, 2x1 m^2, about 1m from the participants head) with their head supported on a chin rest. Each experiment comprised audio-visual (AV) trials, auditory (A) trials and visual (V) trials, and participants were asked to localize either the sound (in AV or A trials) or a visual stimulus (in AV or V trials; for details see individual experiments below). The main effects of interest were probed based on AV trials requiring participants to judge the location of the sound. The pure visual trials or visual judgements in AV trials were mostly included to maintain attention divided across modalities (though for experiment 1 we also compared response biases between visual and auditory judgements). Each trial started with a fixation period (uniform 700 to 1100 ms), followed by the stimulus presentation (Figure 1). The individual stimuli lasted 50 ms but were possibly presented with different stimulus onset asynchronies (SOA). After a post-stimulus interval (uniform 400 to 700 ms) a response cue was presented. This was a horizontal bar along which participants could move a mouse cursor and indicate their judgement by clicking the mouse button. A letter instructed participants about which stimulus they were supposed to localize. Inter-trial intervals varied randomly (uniform 900 to 1200 ms). Participants were asked to maintain fixation during the fixation and stimulus periods but could move their eyes when responding and in inter-trial intervals. Eye movements were not monitored but stimuli had sufficiently short durations to avoid eye movements during stimulus presentation; see (Bosen et al., 2017; Park and Kayser, 2019, 2021) for a discussion of a potential role of eye movements in these paradigms.

Stimulus presentation was controlled using the Psychophysics toolbox (Brainard, 1997) for MATLAB (The MathWorks Inc., Natick, MA) with confirmed temporal synchronization of auditory and visual stimuli. The acoustic stimulus was a 1300 Hz sine wave tone (50 ms duration; sampled at 48 kHz) that was presented at 65 dB SPL through one of 5 speakers (Monacor MKS-26/SW, MONACOR International GmbH & Co. KG). These speakers were located behind the screen at 5 discrete horizontal locations (±22°,±11°, 0°), with the center position always being directly in front of the participant. Sound presentation was controlled via a multi-channel soundcard (Creative Sound Blaster Z) and amplified via an audio amplifier (t.amp E4-130, Thomann). The visual stimuli were projected (LG HF65, LG electronics) onto the acoustically transparent screen. The visual stimulus was a cloud of dots distributed according to a two-dimensional Gaussian distribution (200 dots, SD of vertical and horizontal spread = 1.6°, width of a single dot = 0.12°, duration 50 ms), similar to previous studies (Park and Kayser, 2019, 2021). The color of the dots differed between experiments. For the pilot study these were white (rgb: 250, 250, 250), for experiments 1 to 3 these were orange or blue to allow their differentiation in the task instructions and to perceptually separate two visual stimuli presented in the same trial. The precise values of the blue (rgb 0, 65, 205) and orange hue (rgb 160, 60, 0) were chosen so

they had similar perceptual saliency (luminance) when projected on the screen. These precise values were obtained by letting five individuals adjust the overall luminance of these hues to equate their perceptual similarity. The visual stimuli were either presented at the same discrete locations as the sounds or were positioned relative to the sound locations (see below). The individual locations of auditory and visual stimuli within trials (for AV trials) and between trials were drawn semi-independently resulting in stimulus positions that corresponded to different degrees of spatial (audio-visual) discrepancies. The AV trials (probing the ventriloquism effect) were always followed by A trials (probing the aftereffect), which were either followed by the next AV trial, or less frequently by V trials. Before each experiment participants were given the chance to familiarize themselves with the task and stimuli by performing 20 to 40 practice trials, for which they received feedback on their responses. Before the actual experiment participants were also tested for their spatial hearing abilities. In a screening task, participants were asked to localize noise bursts (50ms duration, 65 dB SPL) presented from the 4 lateralized speakers in a left/right two-choice task, performing 10 repeats per location. Participants performing below an average of 75% correct responses were not invited for the actual experiments (4 out of 95 tested individuals). The average performance of the included participants was 93 % correct (mean; range 78% to 100%).

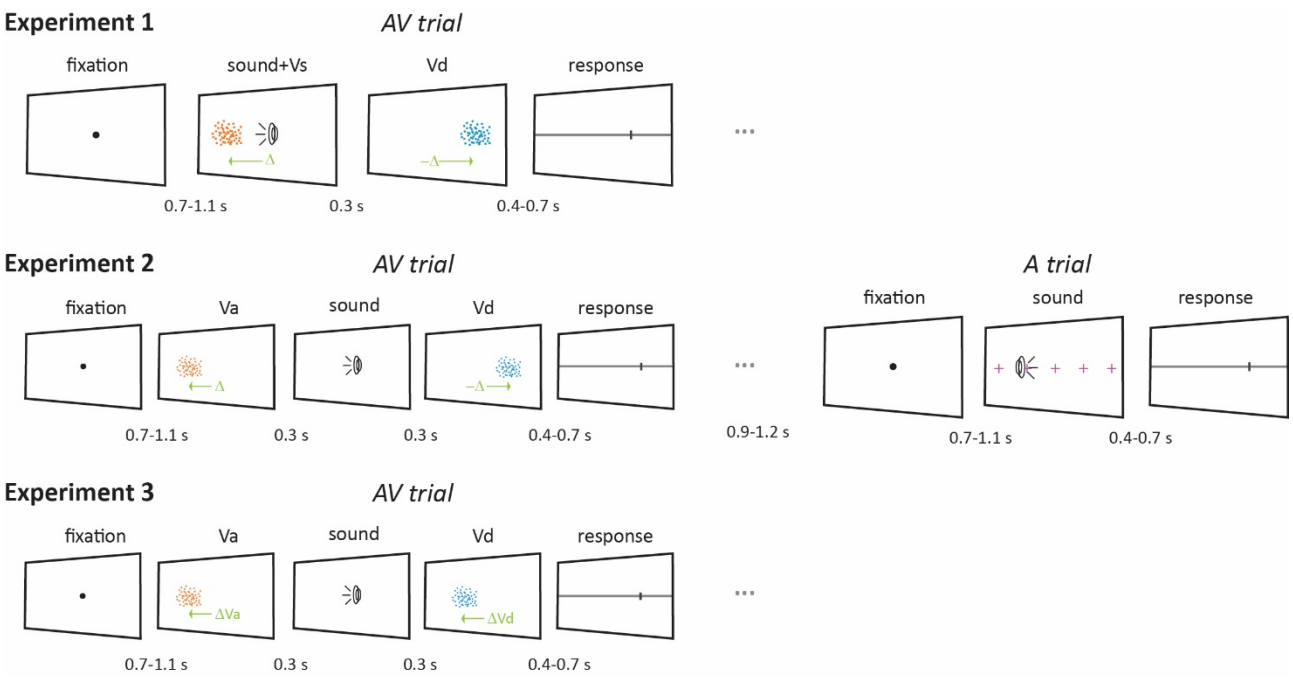

Figure 1. Experimental paradigm. The main paradigm used to probe the ventriloquism bias and aftereffect consisted of a sequence of audio-visual (AV) and auditory trials (A) presented in alternating order (for some experiments these were interspersed with few visual trials). Each trial consisted of a fixation interval, stimulus presentation and a response interval. During the AV trials a sound was presented together with one or two visual stimuli. For AV trials with two visual stimuli in experiment 1 one of them was synchronous with the sound (labelled Vs) the other delayed by 0.3 s (Vd). Both featured spatial offsets (discrepancies, green arrows) relative to the sound of same magnitude (Δ) but opposing directions. In experiment 2 the visual stimuli could either precede the sound (Va), be synchronous (Vs) or delayed (Vd) and again featured discrepancies of same magnitude but opposing directions. In experiment 3 the two visual stimuli featured spatial discrepancies that were independent of each other and could have the same and opposing directions and distinct magnitudes (indicated by ΔVs and ΔVd); as in experiment 2 the visual stimuli could precede or follow the sound or be synchronized. Sounds were presented from one of five discrete locations (indicated by magenta + signs in the A trial) and sound locations were pseudorandomized and counterbalanced across trials and conditions. The visual stimuli were presented with specific spatial discrepancies around each sound location and differed in color, as described in the Methods. Participants localized the sound (in most conditions and trials) or a visual stimulus by adjusting the location of a mouse cursor on the screen.

Pilot study

We conducted a pilot study to select suitable temporal offsets of multiple visual stimuli presented in the same AV trial relative to the sound. This probed the influence of the relative stimulus onset asynchrony (SOA) between auditory and visual stimuli on the ventriloquism effect. To this end we presented the auditory and visual stimuli (in AV trials) at five SOAs (±600ms, ±300ms, and 0ms). The individual stimuli were centered on one of five horizontal locations (±22°,±11°, 0°), with the auditory and visual stimuli always being spatially offset at spatial discrepancies of ±33°, ±22°, and ±11° (defined as visual minus auditory stimulus position). Participants were asked to localize the sound in either AV or A trials, or to localize the visual stimulus in V trials. Each SOA was presented 84 times (14 times per spatial discrepancy) in pseudo-random order. For this pilot we collected data from 23 participants, of which one dataset had to be excluded (see below).

Experiment 1

This experiment was designed to probe how the presence of two (compared to one) visual stimuli in the AV trials affects the two biases. The AV trials implemented a total of eight experimental conditions and either featured one or two visual stimuli. In the case of one visual stimulus this could be synchronous with the sound (condition AVs) or delayed by 300ms (AVd). In the case of two visual stimuli one was synchronous with the sound and one was delayed (AVsVd; Figure 1). The delay of 300ms was determined based on the results from the pilot study. The auditory stimuli could appear at one of five horizontal locations (±22°,±11°, 0°) and the visual stimuli were placed systematically around these five sound locations, implementing six relative spatial discrepancies (±10°,±6°,±2°). Importantly, in condition AVsVd one visual stimulus was displaced to the left of the sound and the other to the right. Hence, the two visual stimuli featured opposing spatial discrepancies of same magnitude. During AV trials participants were asked to either localize the sound or one of the visual stimuli. To facilitate this, the two visual stimuli differed in color (blue or orange, with the assignment of Vs and Vd to each color being randomized across trials). As a result, the individual conditions were defined by the presented stimuli and the judgment type (judging the position of either the sound, the synchronous visual stimulus, Vs, or the delayed visual stimulus, Vd), with each judgement being required equally often. Further, to test whether a potential influence of Vd on judgements in the condition AVsVd is modulated by the position of Vd relative to the sound (rather than just its presence) this experiment also included a condition in which Vd was presented at the same location as the sound (AVsVd0). Participants performed 960 trials, split across 10 blocks. These 960 trials reflect 480 AV trials (60 per each of the eight AV conditions differing by stimulus composition and judgments) and 480 A trials. For each condition the 60 trials were equally divided across the five sound locations and the six spatial discrepancies. Conditions were presented in pseudo-random order. We collected data from 21 participants, of which one had to be excluded.

Experiment 2

This experiment was designed to extend the insights from experiment 1 to conditions in which the two visual stimuli could either precede or follow the sound. As a result, the timing of the two visual stimuli was not predictable. In practice, the visual stimulus could either precede the sound by 300 ms (denoted Va), be synchronous with this (Vs), or delayed relative to this by 300 ms (Vd). As in experiment 1, the visual stimuli were placed around the sounds at six spatial discrepancies (±10°,±6°,±2°). In trials featuring two visual stimuli, these had spatial offsets of same magnitude but opposing directions. Unlike in experiment 1, the vast majority of AV trials required participants to judge the sound, and visual judgements were required only in infrequent trials to maintain balanced attention across modalities. We reduced the fraction of visual judgements in this experiment as the results from experiment 1 revealed a minimal ventriloquism bias following visual judgements. In total this experiment featured three conditions with one visual stimulus (AVs, AVa, AVd) and three with two visual stimuli (AVsVd, AVsVa, AVaVd). Participants performed 936 trials in total, split across 12 blocks. These included 360 AV trials with an auditory judgement, 108 AV trials requiring participants to judge one of the visual stimuli and 468 A trials. The AV trials with an auditory judgement were

equally divided across the five sound locations and the six spatial discrepancies. We collected data from 26 participants, of which two datasets had to be excluded.

Experiment 3

As in experiments 1 and 2 the two visual stimuli had spatial offsets of same magnitude but opposing directions relative to the sound (except condition AVsVd0 in experiment 1), we implemented a further experiment in which we manipulated the relative magnitude of spatial offsets. Similar to experiment 2 the visual stimuli could either precede the sound, be synchronous with it or be delayed, and the sounds were presented from the same five locations. However, in the conditions AVsVd, AVsVa, and AVaVd the two visual stimuli could have spatial offsets in opposing directions and of same magnitude similar to experiments 1 and 2 ([+9,-9],[-9,+9], [+3,-3],[-3,+3]), or have opposing directions and different magnitudes ([+9,-3],[-3,+9],[-9,+3],[3,-9]), or could point in the same directions but have different magnitudes (([+9,+3,[+3,+9],[-9,-3],[-3,-9]). As a result, the spatial discrepancies of the two visual stimuli in a given condition were independent of each other, unlike in experiments 1 and 2. As in experiment 2, the vast majority of AV trials required participants to judge the sound. Participants performed 900 trials each, split across 5 blocks. These included 360 AV trials with an auditory judgement, 360 A trials, and 180 AV trials requiring participants to judge one of the visual stimuli. We collected data from 21 participants.

Data analysis

We investigated the ventriloquism bias obtained from the responses in AV trials and the aftereffect obtained from the responses in A trials. We quantified each bias based on the deviation between the participant's trial-wise response and the expected response given the task-relevant stimulus and its location (Wozny and Shams, 2011a, b; Park and Kayser, 2019). This expected response was defined by taking potential response tendencies into account, such as a to judge stimuli closer to the midline of the screen (Kording et al., 2007; Wozny and Shams, 2011a). Practically, the expected response was computed, per participant, stimulus location and judgement modality, as the average response across all relevant trials with stimuli at this location: for auditory judgements we computed the average response across all A trials for a given sound location, and used this expected response both in AV and A trials. For visual judgements in AV trials of experiment 1 we computed the mean response for a given location of a visual stimulus across all AV trials with the respective judgement. While the expected response captures general response tendencies (e.g. towards the midline) it itself is not shaped by the ventriloquism bias given the overall balancing of stimulus positions and audio-visual discrepancies around each sound location. The trial-wise judgement error was then defined as the difference of the actual response minus this expected response.

We removed few individual trials with judgement errors larger than 150% of the largest spatial discrepancy included in each experiment. These trials can reflect lapses of attention or premature clicks when submitting the response. We visually inspected the resulting biases for all participants. This revealed very large biases for 4 datasets, which can for example arise when participants do not follow task instructions (e.g. localizing the visual rather than the auditory stimulus). These datasets were excluded from analysis and the reported results are from n=22, 20, 24 and 21 datasets respectively for the pilot and experiments 1 to 3. For those datasets the removed trials constituted 0.22±0.1% of all trials for experiment 1 (mean ± s.e.m. across participants), 0.25±0.1 % for experiment 2 and 0.5±0.1 in experiment 3.

The judgement errors were then analyzed as a function of the spatial discrepancy defined as the location of the visual stimulus minus that of the sound in AV trials. For AV trials featuring two visual stimuli, the condition indicated first in the label was used as reference (e.g. judgment errors in AVsVd trials were analyzed as functions of Vs discrepancy). For a group-level visualization we show judgement errors as a function of spatial discrepancy and condition (Fig. 2, 3 left and middle panels). For subsequent analysis we reduced the participant-wise data to a single measure of bias per condition: this was defined as the linear slope of the

trial-wise judgement error against the spatial discrepancy across trials for each condition - referred to as bias slope in the following (Debats et al., 2021; Park and Kayser, 2022). It is important to note that for most conditions in experiments 1 and 2 the two visual stimuli had opposing discrepancies of same magnitude. In these conditions, the slope of the judgement error as a function of the discrepancy of the one visual stimulus has the same magnitude but opposite sign as that relative to the discrepancy of the other visual stimulus. For example, the bias slope in condition AVsVd is expressed relative to the discrepancy of Vs as reference, and the bias slope relative to the discrepancy of Vd would be the negative of this. Note that in experiment 3 this constraint was removed and the two visual stimuli featured independent spatial discrepancies. This makes it difficult to compare the bias slopes for conditions with two visual stimuli between the experiments 1&2 and experiment 3.

The ventriloquism bias can in principle taper off at large discrepancies, reflecting the reduced binding of spatially very discrepant signals as predicted by models of Bayesian multisensory causal inference (Kording et al., 2007; Rohe and Noppeney, 2015; Shams and Beierholm, 2022). While this is seen in the paradigm used here when larger spatial discrepancies are included (Park and Kayser, 2020; Park et al., 2021; Park and Kayser, 2022), the present experiments were restricted to a smaller range of discrepancies and the resulting biases were largely linear. We verified this by comparing the quality of fit of a linear slope for the bias with a model featuring a linear plus a quadratic term, similar as we have used in previous studies (Cao et al., 2019; Park et al., 2021; Park and Kayser, 2022). This quadratic term was proportional to the square-root of the discrepancy and can describe the tapering off at very large spatial discrepancies. For AV trials with one visual stimulus in experiments 1 and 2 the average r-square fit improved only marginally from 0.130±0.020 to 0.138±0.019 (mean±s.e.m. of the participant wise qualities of fit averaged across AV conditions with a single visual stimulus).

Modelling

We implemented modeling analyses to determine how the two visual stimuli collectively shape the ventriloquism bias. To this end we predicted the judgement errors observed in the combined condition (e.g. in AVsVd) based on the biases observed when only one visual stimulus was present (i.e. AVs and AVd). More specifically, we used the bias slopes obtained in conditions with only one visual stimulus as measures of the influence of a single visual stimulus at a given temporal offset. We then estimated how these two individual influences need to be combined in order to approximate the bias observed when both visual stimuli are present in the same trial. This allowed us to estimate the relative weight of the bias induced by each visual stimulus to the combined effect and to arbitrate between different candidate models that may explain the collective bias: e.g. a winner-takes-all model in which only one visual stimulus contributes, a simple linear summation of both biases or some form of weighted averaging.

We modelled the trial-wise judgement errors in condition AVsVd (as example) as follows:

$$\text{error}(t) \sim w_{vs} * b_{AVs} * \Delta_{Vs}(t) + w_{vd} * b_{AVd} * \Delta_{Vd}(t) \qquad \text{(eq. 1)}$$

where t denotes the individual trial, b_{AVs} is the slope obtained in condition AVs, $\Delta_{Vs}(t)$ is the trial-wise spatial discrepancy for stimulus Vs, and w_{vs} and w_{vd} denote the weights by which the bias induced by each visual stimulus contributes to the combined effect. Note that in experiments 1 and 2 the two visual stimuli featured opposing discrepancies, hence $\Delta_{Vs}=-\Delta_{Vd}$, while for experiment 3 the two discrepancies were independent. The weights were fit for each participant and condition by maximizing the log-likelihood of this model using multidimensional constrained nonlinear minimization in Matlab (fminsearchbnd, the weights were assumed to be bound in a range of -0.4 to +1.4). We did not constrain the weights to sum to one to also allow a model that simply reflects the direct additive influence of each visual stimulus. To display the biases predicted by these models we derived the slopes for the predicted data in the same manner as obtained for the actual

data (Fig. 3A). We also investigated the distribution of the respective weights (Fig. 3B). We tested whether the weights differed from zero and one and whether they differed from each other.

We also confirmed that the model featuring participant-specific weights predicts the observed data better than a set of priori specified models: a winner-takes-all model in which the visual stimulus with the individually stronger bias serves as only predictor (in the example this would be w_{vs}=1 and w_{vd}=0), a strict summation (w_{vs}=1 and w_{vd}=1) and a strict averaging (w_{vs}=0.5 and w_{vd}=0.5). For these models we also obtained their log likelihoods of explaining the data. We then compared all four models based on the Akaike information criterion (AIC) as measure of model fit and the Akaike weights of relative model probabilities given the experimentally observed data (Burnham and Anderson, 2004; Gelman et al., 2014). For model comparison we summed the AICs across participants and report the AIC difference between the winning and the second-best model.

Statistical analysis

The bias slopes were subjected to comparisons of interest for each individual experiment. These were implemented using linear models or paired t-tests, for which we report F or t-values, p-values and Bayes factors (BF) for or against the null hypothesis (evidence in favor of the null is coded negatively). Bayes factors for t-tests were obtained using the BayesFactor toolbox for Matlab (DOI:10.5281/zenodo.7006300), for linear models these were obtained by comparing the BIC between full and reduced models omitting the factor of interest. For some individual comparisons one could in principle have combined the data from all three experiments into a single linear model. However, we refrained from doing so for two reasons: first, the context of individual experiments can potentially induce experiment-specific factors that may be lost in such an analysis (Park and Kayser, 2022); and because the data were collected anonymously some participants may have participated in more than one of these experiments, which makes the handling of this factor in a combined analysis difficult. Model weights were compared using paired two-sided t-tests. When interpreting BF's we refer to the nomenclature of Raftery (Raftery, 1995): we interpreted BF between 1 and 3 as 'weak', between 3 and 20 as 'positive', between 20 and 150 as 'strong', and BF > 150 as 'very strong' evidence.

Results

We conducted three experiments to probe how two visual and one acoustic stimulus presented in the same (AV) trial can shape the spatial ventriloquism bias and the aftereffect. To this end we presented a sound in the AV trial together with either one or two visual stimuli that featured spatial and temporal offsets relative to the sound. To ensure that the two visual stimuli can in principle induce biases of different strength we presented them at different temporal offsets to the sound, the magnitude of which was based on a pilot study. Besides addressing the main question about multisensory interactions for two vs. one visual stimulus, each experiment manipulated one additional dimension of interest. In experiment 1 we tested the effect of judgment modality in AV trials. For this we asked participants to either judge the position of the sound or of one of the visual stimuli. While in experiment 1 the visual stimuli were either synchronous with the sound or delayed, experiment 2 included visual stimuli that preceded the sound, were synchronous, or delayed. Hence both the number of visual stimuli and their timing in AV trials was unpredictable. While in experiments 1 and 2 the two visual stimuli featured spatial offsets of the same magnitude but different directions relative to the sound, similar to previous studies (Bertelson et al., 2000; Debats et al., 2021), experiment 3 rendered the spatial discrepancies of the two visual stimuli independent by placing them at different discrepancies and either both on the same or different sides of the sound.

Pilot study

The pilot was implemented to determine suitable stimulus onset asynchronies (SOA) between the visual and auditory stimuli in the audio-visual (AV) trials. To this end we tested five SOAs: ±600ms, ±300ms, and 0ms. We expected the ventriloquism effect to follow the general principle of a multisensory temporal binding window and hence to be strongest for synchronous AV stimuli (van Eijk et al., 2008; Diederich and Colonius, 2009; Stevenson and Wallace, 2013; Wallace and Stevenson, 2014). Indeed, a linear model provided very strong evidence in favor of an effect of SOA magnitude on bias slope ($F(2,107)=38.2$, $p<10^{-5}$, $BF\sim10^{8}$, $n=22$; the individual group-level slopes were as follows: 0.16±0.03, 0.22±0.04, 0.32±0.04, 0.18±0.03, 0.16±0.025; mean±s.e.m ordered from -600 to +600ms). Subsequent comparisons showed that the bias was comparable between visual-leading and auditory-leading trials (±300ms: $t=1.37$, $p=0.19$, $BF=-2.0$; ±600ms $t=0.42$, $p=0.68$, $BF=-4.1$; $n=22$), was smaller for ±300ms compared to synchronous stimuli ($t=5.55$, $p<10^{-5}$, $BF=1378$) and was smaller for ±600ms compared to ±300ms ($t=5.03$, $p=0.0001$, $BF=465$). Based on these results we selected an SOA of ±300ms for the subsequent experiments, as it induces sufficient multisensory binding of visual and auditory stimuli. In addition, the bias observed for synchronous visual stimuli is stronger than for asynchronous ones, which would allow to clearly see the dominance of a synchronous stimulus in the combined condition, if such a dominance exists.

Ventriloquism bias for auditory judgements

We here focus on AV trials with auditory judgements, which were common to all experiments. In Figure 2 we show the group-level judgement errors as a function of spatial discrepancy between the auditory and visual stimuli (left and middle column) and their respective slopes against this discrepancy (right).

In trials with one visual stimulus we observed the expected effects of temporal delay (Fig. 2 left panels). A comparison of bias slopes returned very strong evidence for a stronger bias for synchronous AV pairs (Exp1 AVs vs. AVd: $t=5.70$, $p<10\text{-}5$, $BF=1531$; Exp2 AVs vs. mean of AVd and AVa: $t=5.68$, $p<10^{-5}$, $BF=2469$; Exp3 AVs vs. mean of AVd and AVa: $t=3.49$, $p=0.0026$, $BF=15$). Similar as for the pilot study we observed no clear effect of temporal order for the asynchronous AV pairs (Exp2 AVa vs. AVd: $t=-2.22$, $p=0.04$, $BF=1.54$; Exp3 AVd vs. AVa: $t=0.97$, $p=0.44$, $BF=-3.3$).

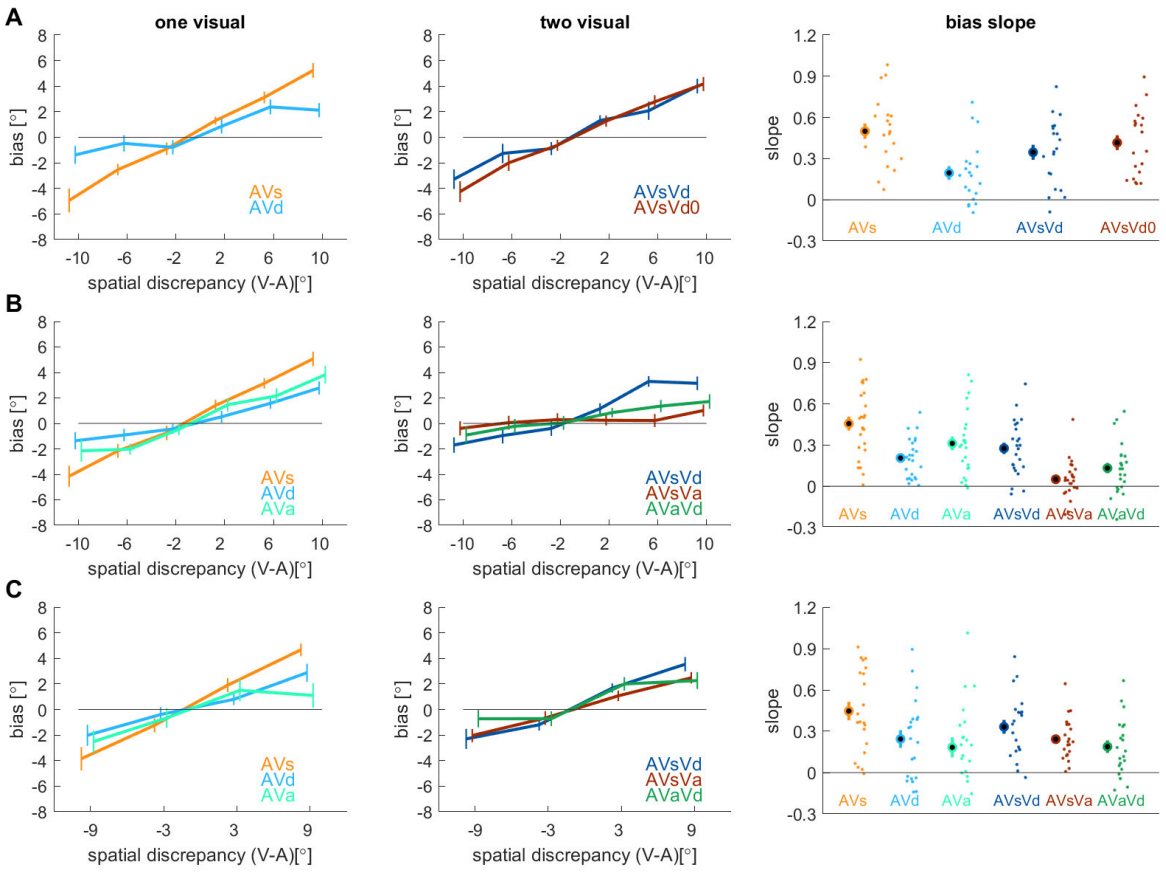

Figure 2. Ventriloquism bias in AV trials with auditory judgements. Left: Group-level judgement errors as a function of spatial discrepancy for AV trials with one visual stimulus. Middle: same for AV trials with two visual stimuli. Right: Bias slope for each condition. **A**: Data from experiment 1 (n=20). **B**: Data from experiment 2 (n=24). **C**: Data from experiment 3 (n=21). Spatial discrepancy is defined as the position of the visual stimulus (when one was presented, or the one indicated first in the condition label) minus the position of the sound. Note that in experiments 1 and 2 the two visual stimuli presented in the same trial had opposing spatial discrepancies (hence the bias in condition AVsVd towards Vd has the same magnitude but opposing sign as towards Vd). In experiment 3 the two visual stimuli had independent discrepancies, which makes it difficult to directly compare the precise slopes in conditions with two visual stimuli between experiments. For example, the vanishing bias for AVsVa in experiment 2 indicates a comparable influence of both visual stimuli, but the same effect would not lead to a vanishing bias in experiment 3. Lines and circles indicate the group mean, error-bars the s.e.m. across participants, dots the individual participants.

Focusing on our main question, we found that the presence of two visual stimuli in an AV trial reduced the bias compared to the presence of only a single visual stimulus, but did so differently depending on their timing to the sound. In experiments 1 and 2 both visual stimuli had opposing discrepancies to the sound. Hence, an equal influence of both visual stimuli on auditory judgements should lead to a direct cancellation of their individual influences and a vanishing slope in the combined condition. In contrast, a differential influence of each visual stimulus should not lead to a vanishing bias. For trials involving one visual stimulus synchronous with the sound (AVsVd and AVsVa) the bias pointed towards Vs and we obtained very strong evidence for a reduced bias in the combined condition (Table 1). The collective bias differed from zero for AVsVd while the data provide no clear evidence for such a difference from zero for AVsVa. When both visual stimuli were temporally offset (AVaVd) the bias pointed towards Va (hence away from Vd). The slope was again reduced compared to the presence of only one visual stimulus but the data provide positive evidence for a non-vanishing combined effect (Table 1). These results suggest that the ventriloquism bias in the presence of two visual stimuli is shaped not only by one of these (e.g. the synchronous one) but by the collective visual evidence.

To further characterize the interaction of two visual stimuli, experiment 1 featured a condition in which Vd was presented at the same location as the sound (hence only Vs was spatially offset; condition AVsVd0). If Vd affects the bias directly in proportion to its spatial discrepancy, one would expect the bias to differ between the control condition (no spatial offset of Vd) and the condition in which both Vs and Vd are spatially offset: the data indeed offer positive evidence for this hypothesis (AVsVd vs. AVsVd0: t=2.55, p=0.019, BF=2.9). Yet, the data also provide weak evidence for the bias in this control condition to differ from that obtained with only Vs present (AVs vs. AVsVd0: t=2.45, p=0.024, BF=2.5), suggesting that the second visual stimulus shapes the bias not only in direct proportion to its spatial offset.

For experiment 3 the two visual stimuli featured independent spatial offsets that could vary in magnitude and direction. Hence, directly opposing influences of two visual stimuli do not necessarily lead to a vanishing bias slope in the combined condition. We show the respective biases in Figure 3 for the sake of completeness and point to the modelling section for the analysis of this dataset.

	AVsVd vs. AVs			AVsVd vs. 0			AVsVa vs. AVs			AVsVa vs. 0			AVaVd vs. AVa			AVaVd vs. 0		
	t	p	BF	t	p	BF	t	p	BF	t	p	BF	t	p	BF	t	p	BF
Exp1 n=20	-4.03	0.0007	49	6.30	$<10^{-5}$	4332												
Exp2 n=24	-4.65	0.0001	216	6.72	$<10^{-5}$	23337	-6.90	$<10^{-5}$	28080	1.80	0.086	-1	-5.08	$<10^{-5}$	570	3.47	0.002	19

Table 1. Statistical analysis of ventriloquism slopes. Results are obtained from two-sided paired t-tests within experiments that either contrasted conditions or individual conditions versus the null hypothesis of a zero bias. Empty cells indicate conditions not present in the respective experiment. Sample sizes are indicated. BF=Bayes factor.

Model based analysis

Conceptually the ventriloquism bias could behave according to different models. For example, one of the visual stimuli could dominate and shape the bias independently of the other visual signal (e.g. the synchronous one). Such a winner-takes-all model could be expected based on the reported 'capturing' of multisensory perception by one modality (Colavita, 1974; Spence, 2009; Chen and Spence, 2017). Alternatively, the influences of two visual stimuli could superimpose. A near linear summation could be expected based on the additive effects of multisensory stimuli on arousal (Van der Stoep et al., 2021), response properties of neurons in the superior colliculus (Alvarado et al., 2007; Stanford and Stein, 2007; Stein and Stanford, 2008; Stein et al., 2009) or signatures of within modality integration in multisensory circuits (Ohshiro et al., 2011). However, a superposition could also be skewed towards one stimulus, as seen in multisensory paradigms that involve one stimulus per modality and consider each stimulus equally task relevant (Ernst and Bulthoff, 2004; Kording et al., 2007; Angelaki et al., 2009). We hence compared the experimentally observed data to several model-based predictions: a winner-takes-all model, a direct sum, averaging and a model in which we optimized the weights attributed to each visual stimulus to best reflect the participant-specific bias pattern. To this end we modelled the participant-wise judgement errors in conditions with two visual stimuli as a linear weighted superposition of the two influences of the visual stimuli obtained when they were presented individually with the sound.

For all three experiments and all conditions the data were best explained by the participant-wise optimized model, which provides clear evidence against a winner-takes all model or a superposition with fixed weights (the smallest delta AIC between the optimized model and any other model was 47.2, the respective Akaike weights for the optimized model were all ~ 1). The model with the participant-derived weights explained the observed bias slopes well (Figure 3A) and the r-square between the observed and predicted slopes computed across participants and all eights conditions was 0.98 (total n=112).

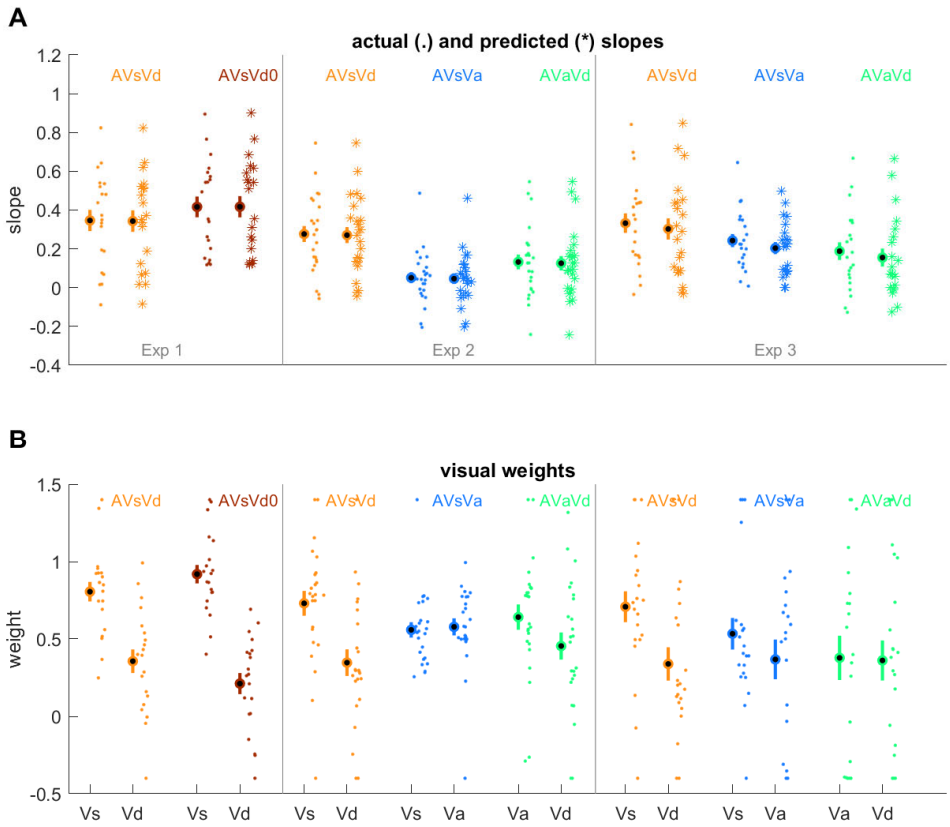

Figure 3: Modelling results for the ventriloquism bias. The judgement error reflecting the ventriloquism bias in AV trials featuring two visual stimuli was predicted based on the influence of the respective visual stimuli when paired individually with the sound. **A:** Slopes of the ventriloquism bias observed in experiments 1 to 3 and those predicted by the model relying on participant-wise derived weights. The conditions are ordered from left to right for experiments 1 to 3.**B:** The weights assigned to the bias induced by each visual stimulus individually in predicting the combined bias. Error bars indicate the group mean and s.e.m. across participants, dots and stars individual participants.

	AVsVd			AVsVa			AVaVd			AVsVd0		
	t	p	BF	t	p	BF	t	p	BF	t	p	BF
Exp1 n=20	3.56	0.002	19							6.12	$<10^{-5}$	3057
Exp2 n=24	2.31	0.03	2.0	-0.19	0.84	-4.6	1.12	0.27	-2.7			
Exp3 n=21	2.74	0.012	4.2	0.96	0.34	-2.9	0.08	0.93	-4.3			

Table 2. Statistical comparisons of model weights for the predicted ventriloquism slopes. Results are obtained from two-sided paired t-tests within experiments. Empty cells indicate conditions not present in the respective experiment. Sample sizes are indicated. BF=Bayes factor.

The weights for each pair of visual signals are shown in Figure 3B. We obtained positive to strong evidence for these to differ from zero (min BF=3.4), and all except for one condition all differed from one (min BF= 5.7); only for condition AVsVd0 did we obtain weak evidence for the weight of AVs to be comparable to one (t=1.35, p=0.19, BF=-1.93). Importantly, we then contrasted the weights between the two visual stimuli to determine the relative influence of each to the combined effect (Table 2). This revealed that the relative influence of each visual signal differed with the overall context defined by the relative timing of the visual stimuli presented in that condition. For the condition with a synchronous and a delayed stimulus combined

(condition AVsVd) we obtained weak to positive evidence for the weight of the synchronous visual stimulus to be stronger compared to the delayed stimulus. In contrast, for the condition combining a synchronous stimulus with one preceding the sound we obtained weak or positive evidence in favor of comparable weights (condition AVsVa). Similarly, in the condition with one visual stimulus preceding and one following the sound we obtained weak or positive evidence in favor of comparable weights (condition AVaVd).

To further characterize this context (i.e. condition) dependence we compared the weights of each visual stimulus between conditions. For example, we asked whether the contribution of Vs is comparable between conditions AVsVd and AVsVa. Note that this is sensible as the weights were not normalized to unity and hence capture the absolute (rather than relative) influence of each visual stimulus. For Vd this provided weak to positive evidence for comparable weights between contexts in both experiments (weight of Vd in AVsVd vs. AVaVd, exp 2: t=-1.37, p = 0.18, BF=-2.0 ; exp 3: t=-0.13, p=0.90, BF=-4.4) and the same was true for Va (weight of Va in AVsVa vs. AVaVd exp 2: t=-0.57, p=0.57, BF=-4.0, exp 3: t=-0.09, p=0.92, BF=-4.4). In contrast, for Vs this provided weak evidence in favor of the influence of Vs to be stronger when combined with a delayed rather than a preceding visual stimulus (weight of Vs in AVsVd vs AVsVa, exp 2: t=1.95, p=0.06, BF=1.1, exp3: t=2.2, p=0.04, BF=1.7). All in all, these findings show in particular that the influences of a synchronous and a delayed visual signal differ, but also that the overall influence of a synchronous visual stimulus varies with context.

Ventriloquism bias for visual judgements

In experiment 1 we also probed the impact of judgement modality in the AV trials and asked participants to either localize the sound or one of the visual stimuli. For AV trials with visual judgements we observed a robust ventriloquism bias that pointed towards the sound (condition AVs with visual judgements 0.09±0.01 mean±s.e.m.; AVd with visual judgements 0.12±0.02; note that here we use a positive sign to indicate a bias towards the sound). The biases for visual judgements were smaller compared to auditory judgements (AVs t=7.16, $p<10^{-5}$, $BF=2*10^{5}$; AVd t=2.64, p=0.016, BF=3.5), which could be expected given the higher reliability of the visual stimuli in these experiments. For conditions with two visual stimuli the slopes were enhanced in the presence of a second visual stimulus (AVsVd vs. AVs both judging Vs: t=2.09, p=0.049, BF=1.4; AVsVd vs. AVd both judging Vd: t=2.84, p=0.01, BF=4.95). Given that Vs and Vd had opposing offsets to the sound, a bias from Vs towards the sound points also towards Vd. Hence, the stronger bias in the combined condition may either reflect the combined attraction of the sound and the second visual stimulus or a change in the relative weight of the visual stimulus in the presence of a visual distractor. The present data cannot disambiguate these effects.

Aftereffect following AV trials

The trial-wise ventriloquism aftereffect is a weak but robust bias visible in auditory judgements following a single AV trial. The present experiments allowed us to probe i) how this aftereffect is shaped by the relative timing of the visual and auditory stimuli in the AV trials, ii) whether it emerges following AV trials featuring two visual stimuli and iii) whether it emerges following AV trials featuring a visual judgement.

To answer the first question, we focus on A trials following AV trials featuring a single visual stimulus (Figure 4). This revealed positive evidence against an effect of relative timing (Exp 1 AVs vs AVd: t=0.37, p=0.71, BF=-4.0; Exp 2 AVs vs mean of AVd and AVa: t=1.28, p=0.21, BF=-2.2; Exp 3 AVs vs. mean of AVd and AVa: t=1.02, p=0.31, BF=-2.76) and positive evidence against an effect of temporal order (Exp 2 AVd vs. AVa: t=-0.48, p=0.64, BF=-4.2; Exp 3: t=0.23, p=0.81, BF=-4.29).

To answer the second question, we focus on the aftereffect following AV trials with two visual stimuli. For these conditions the aftereffects differed only marginally from zero, in particular for condition AVsVa (Table 3). In this case we also observed strong evidence for a clear reduction in the aftereffect compared to the presence of Vs only. Hence, the overall picture emerging is one similar as for the ventriloquism effect, with

visual signals prior to and following a sound to have a differential impact. However, given the overall weak aftereffect slopes this result is less robust than for the ventriloquism bias.

	AVsVd vs. AVs			AVsVd vs. 0			AVsVa vs. AVs			AVsVa vs. 0			AVaVd vs. AVa			AVaVd vs. 0		
	t	p	BF	t	p	BF	t	p	BF	t	p	BF	t	p	BF	t	p	BF
Exp1	-1.8	0.07	1.1	2.81	0.011	5												
Exp2	-1.79	0.085	-1.2	2.03	0.054	1	-3.34	0.003	14.2	1.06	0.299	-3	-1.46	0.16	-1.81	2.06	0.051	1

Table 3. Statistical analysis of aftereffect slopes. Results are obtained from two-sided paired t-tests within experiments that either contrasted conditions or individual conditions versus the null hypothesis of a zero bias. Empty cells indicate conditions not present in the respective experiment. Sample sizes are indicated. BF=Bayes factor.

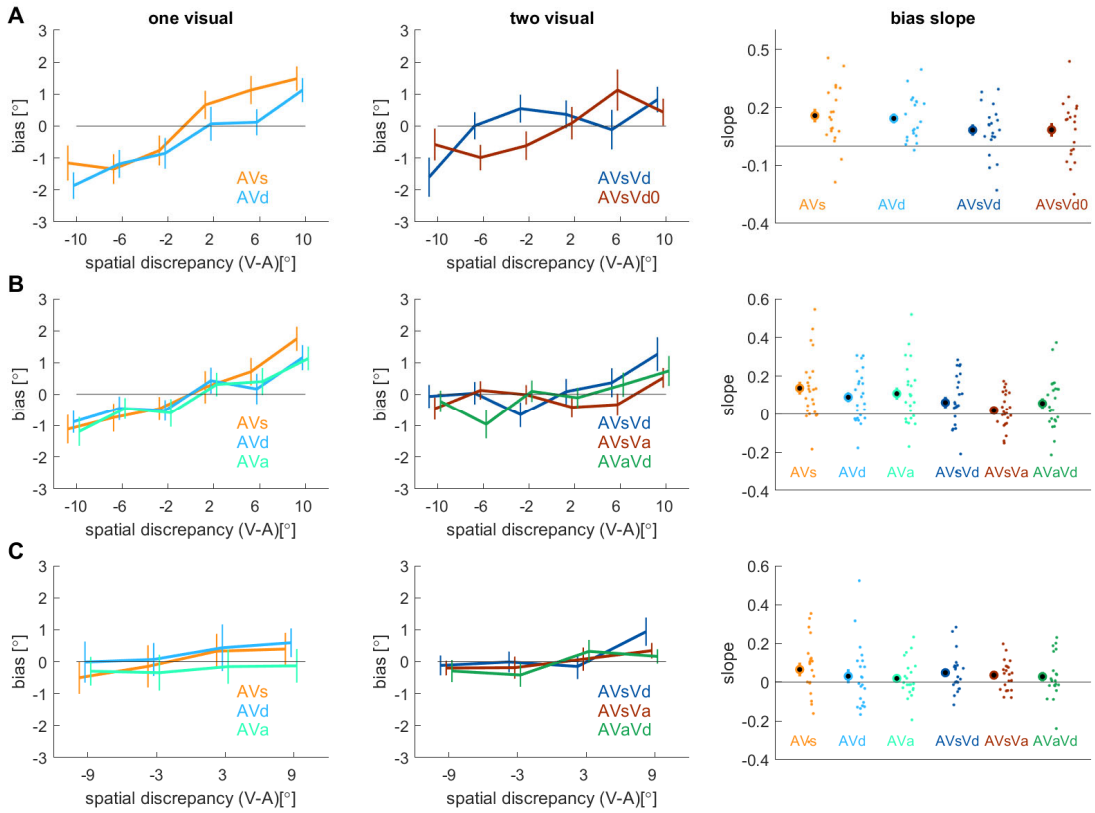

Figure 4. Aftereffect following AV trials with auditory judgements. Left: Group-level bias as a function of spatial discrepancy for AV trials with one visual stimulus. Middle: Group-level bias for AV trials with two visual stimuli. Right: Bias strength for each condition. **A**: Data from experiment 1 (n=20). **B**: Data from experiment 2 (n=24). **C**: Data from experiment 3 (n=21). Lines and circles indicate the group mean, error-bars the s.e.m. across participants, dots individual participants.

To answer the third question, we focus on AV trials with visual judgements (Figure 5). Following trials with one visual stimulus we also observed a reliable aftereffect (Exp1 AVs: 0.12±0.40, AVd: 0.12±0.02), which pointed towards the visual stimulus and hence in the same direction as following AV trials with auditory judgements. Comparing the aftereffect following AV trials with auditory and visual judgements revealed positive evidence for a comparable slope (Exp 1 Vs: t=0.73, p=0.47, BF=-3.4; Vd: t=0.73, p=0.48, BF=-3.4), hence showing that the aftereffect emerges independently of the judgments made in preceding AV trials. Following AV with two visual stimuli and a visual judgement, however, the aftereffect vanished (Exp 1 AVsVd judging Vs: mean±s.e.m. 0.04±0.03; t=1.21, p=0.23, BF=-2; AVsVd judging Vd: 0.02±0.04; t=0.78, p=0.44, BF=-

3) and we found positive evidence for the slope to be smaller compared to conditions featuring only a single visual stimulus (AVsVd vs. AVs judging Vs: t=-3.09, p=0.006, BF=7.8; AVsVd vs. AVd judging Vd: t=-2.68, p=0.015, BF=3.7). This suggests that the aftereffect emerges irrespectively of the task relevant modality in the AV trial as long as only one visual signal is present, but competing visual stimuli are detrimental to this aftereffect.

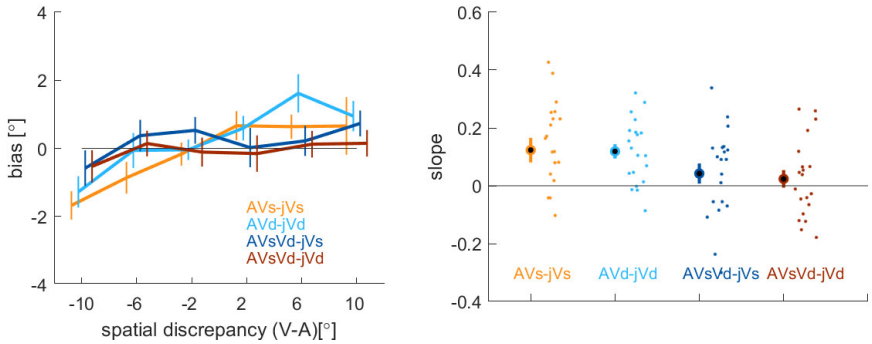

Figure 5. Aftereffect following AV trials with visual judgements. Left: Group-level bias as a function of spatial discrepancy for AV trials with one or two visual stimuli. Right: Bias strength for each condition. Spatial discrepancy is defined as the sound position minus the position of either the single visual stimulus, or Vs. The conditions labels include the respective judgment in the preceding AV trial (jVs or jVd). Lines and circles indicate the group mean, error-bars the s.e.m. across participants, dots individual participants (data from experiment 1).

Discussion

We studied how the presence of two visual stimuli shapes multisensory perception in the audio-visual spatial ventriloquism paradigm. Our results provide a number of insights about the ventriloquism effect and the trial-wise aftereffect. Foremost, they show how two visual stimuli collectively shape these two biases. For AV trials with an auditory judgement our data suggest that the influence of an individual visual stimulus is shaped by its own relative timing to the sound, but importantly, also by the temporal context provided by another visual signal. Depending on this overall context, two visual signals can influence the collective bias with comparable or with asymmetric weights. Concerning the aftereffect our results show that this emerges following both auditory and visual judgements in the preceding AV trials and also following asynchronous AV stimulus pairs, confirming the idea that this aftereffect is a rather automatic adjustment to spatial discrepancies.

Ventriloquism effect for multiple visual stimuli

Our results show that when two visual stimuli are combined with a sound, both visual signals exert an influence on the ventriloquism bias. This was most pronounced when the sound was judged in the AV trials, likely because the acoustic stimuli carried less reliable spatial information in the present experiments and multisensory biases tend to scale with the relative reliability of each sensory modality (Ernst and Bulthoff, 2004; Rohe and Noppeney, 2015). Importantly, our data show that the influence of a visual stimulus on the ventriloquism bias scales with its spatial distance to the sound, with its specific relative timing to the sound, and importantly, depends also on the overall temporal context implemented by the other visual signals. We interpret this context-sensitivity of multisensory integration as a consequence of sensory causal inference shaping the binding process based on the collective evidence.

Multisensory causal inference guides perception by fostering the binding of those sensory cues that are deemed to be causally related and to originate from a common sensory object (Kording et al., 2007; Odegaard and Shams, 2016; Odegaard et al., 2017; Noppeney, 2021; Shams and Beierholm, 2022). It is known that causal inference is context dependent (Odegaard and Shams, 2016; Odegaard et al., 2017; Park and Kayser, 2022), reflecting for example the variable likelihood of a face and a voice belonging to the same person depending on the total number of people in the room. The condition dependence of the ventriloquism bias for two visual stimuli in our view directly reflects the outcome of such a context sensitive causal inference process. During the course of an individual trial the time of occurrence of one or more visual signals was unpredictable. As a result, perceptual processes may integrate the available visual evidence with a constant weight up to the timepoint when strong evidence for a direct causal relation becomes evident. Here this would be the occurrence of a visual stimulus synchronous with the sound (i.e. Vs). Once evidence for a specific causal relation is available subsequent sensory signals receive a reduced weight, given that subsequent signals are delayed and hence less likely causally linked to the sound. This temporally asymmetric emergence of evidence for putative causal relations directly results in an asymmetric contribution of visual signals to the collective bias. However, the evidence for a causal relation between audio-visual depends also on the absolute audio-visual delay, as evident in the AV conditions featuring a single visual stimulus. As a result, the overall influence of each visual stimulus depends both on its own specific features and the overall context provided by all visual signals. This model directly predicts a comparable influence of Va and Vs when only these are present, a reduced influence of Vd when combined with Vs, and in absolute terms a stronger influence of Vs when being presented as first visual stimulus compared to when being presented second.

Our data directly support this model. In condition AVsVd the influence of Vd was clearly reduced compared to that of Vs. This is evident by their differential weights for the collective bias and the finding that the combined influence did not cancel when both stimuli featured directly opposing spatial discrepancies (in experiment 2). In contrast, when combining Va and Vs with a sound, their weights were comparable and their opposing biases directly cancelled (experiment 2). Furthermore, the absolute influence of Vs on the

combined bias differed between conditions, with the evidence pointing to a stronger influence of Vs when accompanied by a visual stimulus following the sound. And finally, when two temporally offset visual stimuli were present, their weights were largely comparable though still the collective bias did not cancel, speaking in favor of a slightly stronger influence of the preceding visual signal.

Such an asymmetry of a multisensory temporal binding window fits with studies investigating temporal simultaneity detection (SJT) or temporal order judgements (TOJ), which often report asymmetric effects (Powers et al., 2009; Stevenson et al., 2012; Stevenson and Wallace, 2013; Bedard and Barnett-Cowan, 2016). In these tasks visual and auditory stimuli are systematically temporally offset and the observed response patterns (e.g. fraction of synchrony judgements) differs between temporal orders at fixed SOA magnitudes. Typically, the binding window is longer when visual stimuli lead over the sound. This asymmetry of the multisensory binding window in audio-visual temporal tasks fits well with the stronger influence of preceding over delayed visual stimuli in AV trials with two visual stimuli observed here.

The differential influence of visual signals prior to and subsequent to a sound that we observed in conditions featuring two visual stimuli was not visible in trials featuring only one visual stimulus. Consistently across all three experiments we found the ventriloquism bias for individual visual stimuli to be symmetric with regards to the sign of the SOA. This in our view further substantiates the context sensitive influence of visual signals at different SOAs, likely shaped by sensory causal inference. In the present experiments both the timing and the number of visual stimuli in each trial were unpredictable. The time lag between sound and a visual signal may be more relevant for determining putative causal relations when multiple rather than one candidate signals are present (Van der Burg et al., 2008b; Talsma et al., 2010). In particular, in the presence of multiple visual signals the overall evidence for causal relations may depend not only on the absolute SOA of each visual signal but may engage additional selection processes that determine which of two visual signals better relates to the sound (Kording et al., 2007; Noppeney, 2021; Shams and Beierholm, 2022). Hence the emergence of temporal asymmetries in conditions featuring two visual signals may not be a qualitative difference to the more symmetric temporal pattern seen for individual visual stimuli, but simply the consequence of the elaborate computations shaping sensory causal inference in more complex multisensory environments.

The present analysis was constrained to biases that were linearly proportional to the degree of spatial discrepancy and which combined the influence of temporally dispersed visual stimuli in a linear manner. It is well known that the ventriloquism bias (and also the aftereffect so some degree) can scale in a nonlinear manner with the spatial discrepancy (Kording et al., 2007; Rohe and Noppeney, 2015; Shams and Beierholm, 2022). This nonlinear dependency is predicted by models of Bayesian causal inference and becomes prominent at large spatial discrepancies, which we avoided in the present study. Similarly, the temporal binding window when considered over longer SOAs than those included here may reflect a nonlinear interaction of multiple modalities (Stevenson and Wallace, 2013). This poses limitations over the spatial range and time lags over which the present results may generalize, and it is possible that spatial and temporal features interact nonlinearly over a larger range. Mapping the combined spatio-temporal binding window for the ventriloquism bias over a larger range of spatial and temporal scales remains to be implemented in the future.

Multisensory binding follows a weighted sensory combination rather than sensory capture

Our results speak against the idea of multisensory capture shaping perceptual judgements in the present paradigm. Multisensory capture refers to the phenomenon that one stimulus, or more precisely here one pair of crossmodal stimuli, will capture attention and capitalize the available processing resources in favor of the processing of this stimulus pair (Van der Burg et al., 2008a; Talsma et al., 2010). A prominent example is the pip and pop effect, in which the synchronization of one out of many visual signals to a sound fosters the perceptual pop-out of this one visual stimulus. In the ventriloquism paradigm, it was previously shown that

when two visual stimuli of same spatial offset but distinct saliency are presented with one sound, the more salient visual stimulus dominates the ventriloquism bias (Bertelson et al., 2000). Our data in contrast show than when multiple visual stimuli of comparable saliency are present, the specific synchronization of one of them to the sound does not lead to a similar capturing of the multisensory bias. Rather, the collective data is very consistent with a spatio-temporally weighted contribution of the individual signals.

To further understand the contribution of spatial and temporal dimensions, experiment 1 included a control condition in which one visual stimulus featured only a temporal offset but was located at the same position as the sound (i.e. had a spatial discrepancy of zero degrees). The bias in this condition (AVsVd0) was reduced compared to the condition where only Vs was present. In our view this demonstrates that neither the spatial nor the temporal dimension dominates over the other in shaping the respective weight of each visual stimulus. Theoretically, the lack of spatial offset of Vd0 could have captured the binding process, which would have led to a negligible bias given it's zero spatial offset. All in all, the present data speak against the notion that multisensory perception in the ventriloquism paradigm is dominated by a single signal and rather point to a systematic weighting of all potentially task-relevant audio-visual signals given the context of all sensory signals available.

Aftereffect emerges robustly across task relevance and temporal delays

Similar to previous studies we observed a robust trial-wise (also called immediate) ventriloquism aftereffect in unisensory auditory trials following the AV trials (Radeau and Bertelson, 1974; Recanzone, 1998; Bruns et al., 2011; Park and Kayser, 2019, 2021). For AV trials featuring one visual stimulus this aftereffect reflects a bias in the localization of a sound towards the direction of the spatial discrepancy experienced the AV trial (Wozny and Shams, 2011b; Park et al., 2016). We here corroborate the robustness of this aftereffect and extend previous work in a number of ways.

First, we show that the aftereffect emerges also when the visual and auditory stimuli are asynchronous, at least on the scale of 300ms used here. As for the ventriloquism bias we did not find a sign of clear temporal asymmetry for individual visual stimuli in the respective AV trials. Second, we show that this bias in sound localization also emerges when the participants localized a single visual stimulus in the preceding AV. Also following visual judgements did the aftereffect point towards the direction in which the visual stimulus was offset relative to the sound in the preceding AV trial. This corroborates that the aftereffect emerges automatically and independently of the precise task instructions, possibly as a result of the belief in a modality specific bias (Wozny and Shams, 2011a; Noppeney, 2021; Park and Kayser, 2021). Other studies have shown that the aftereffect emerges also automatically after a series of passive exposure trials in which often no judgement is required (Wozny and Shams, 2011b; Frissen et al., 2012; Bruns and Roder, 2015; Bosen et al., 2017), and the present data support the notion that the task-relevance of the audio-visual exposure trials inducing the aftereffect is not relevant.

However, our data also suggest that this robustness to the task-relevance in the previous AV trials persists only as long as the sensory-causal relations in this AV trial are unambiguous. In fact, following AV trials with two visual stimuli and a visual judgment the aftereffect vanished. In our view this result supports the notion that the aftereffect reflects an adjustment in response to a belief in a modality specific (i.e. auditory) bias. Potential evidence for such a bias emerges in the presence of individual audio-visual stimulus pairs, regardless of which stimulus was directly task relevant. However, when multiple visual signals are present it may be less clear whether or which one provides evidence about a modality specific bias and this is even reinforced further when the visual modality itself is task relevant. When performing a task based on an ambiguous or cluttered visual environment there may be no evidence for a modality specific bias in the auditory domain. This influence of task- or behavioral-relevance on the aftereffect is also supported by a study that rendered one of two visual stimuli behaviorally relevant over the course of an entire experiment and found an aftereffect bias specifically towards the relevant visual stimulus (Tong et al., 2020).

Conclusion

In every day life we are faced with multiple signals in each sensory modality and combining these into a coherent percept is a major feat achieved by the brain. While many laboratory studies employ only a single stimulus per modality, the present experiments pave the way to understand how multisensory perception behaves under more realistic conditions involving multiple stimuli in one modality. We suggest that the underlying sensory computations are guided by multisensory causal inference and the present data show how this results in context-sensitive multisensory biases in the spatial ventriloquism paradigm. Here, the influence of individual visual signals is shaped by their individual properties such as the spatial distance and temporal delay to a sound, but also by the temporal context provided by other visual signals. These results highlight the context-sensitive nature of multisensory perception, which so far has received comparatively little attention.

Acknowledgements

This study was funded by the Deutsche Forschungsgemeinschaft (DFG KA2661/2-1). The authors declare that they have no conflict of interests. We thank Jenny Jakisch and Lisa Stetza for their help during data collection.

References

Alvarado JC, Vaughan JW, Stanford TR, Stein BE (2007) Multisensory versus Unisensory Integration: Contrasting Modes in the Superior Colliculus. J Neurophysiol 97:3193-3205.

Angelaki DE, Gu Y, DeAngelis GC (2009) Multisensory integration: psychophysics, neurophysiology, and computation. Curr Opin Neurobiol 19:452-458.

Bedard G, Barnett-Cowan M (2016) Impaired timing of audiovisual events in the elderly. Exp Brain Res 234:331-340.

Bertelson P, Radeau M (1981) Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. Percept Psychophys 29:578-584.

Bertelson P, Vroomen J, de Gelder B, Driver J (2000) The ventriloquist effect does not depend on the direction of deliberate visual attention. Percept Psychophys 62:321-332.

Bosen AK, Fleming JT, Allen PD, O'Neill WE, Paige GD (2017) Accumulation and decay of visual capture and the ventriloquism aftereffect caused by brief audio-visual disparities. Exp Brain Res 235:585-595.

Brainard DH (1997) The Psychophysics Toolbox. Spat Vis 10:433-436.

Bruns P (2019) The Ventriloquist Illusion as a Tool to Study Multisensory Processing: An Update. Front Integr Neurosci 13:51.

Bruns P, Roder B (2010) Tactile capture of auditory localization is modulated by hand posture. Exp Psychol 57:267-274.

Bruns P, Roder B (2015) Sensory recalibration integrates information from the immediate and the cumulative past. Sci Rep 5:12739.

Bruns P, Liebnau R, Roder B (2011) Cross-modal training induces changes in spatial representations early in the auditory processing pathway. Psychol Sci 22:1120-1126.

Burnham KP, Anderson DR (2004) Multimodel Inference: Understanding AIC and BIC in Model Selection. Sociological Methods & Research 33:261-304.

Cao Y, Summerfield C, Park H, Giordano BL, Kayser C (2019) Causal Inference in the Multisensory Brain. Neuron 102:1076-1087.

Chen YC, Spence C (2017) Assessing the Role of the 'Unity Assumption' on Multisensory Integration: A Review. Front Psychol 8:445.

Colavita FB (1974) Human sensory dominance. . Perception & Psychophysics 16:409-412.

Debats NB, Heuer H, Kayser C (2021) Visuo-proprioceptive integration and recalibration with multiple visual stimuli. Sci Rep 11:21640.

Diederich A, Colonius H (2009) Crossmodal interaction in speeded responses: time window of integration model. Prog Brain Res 174:119-135.

Ernst MO, Bulthoff HH (2004) Merging the senses into a robust percept. Trends Cogn Sci 8:162-169.

Frissen I, Vroomen J, de Gelder B (2012) The aftereffects of ventriloquism: the time course of the visual recalibration of auditory localization. Seeing Perceiving 25:1-14.

Gelman A, Hwang J, Vehtari A (2014) Understanding predictive information criteria for Bayesian models. Statistics and Computing 24:997-1016.

Howard IP, Templeton WB (1966) Human spatial orientation. London: Wiley.

Kanaya S, Yokosawa K (2011) Perceptual congruency of audio-visual speech affects ventriloquism with bilateral visual stimuli. Psychon Bull Rev 18:123-128.

Kording KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L (2007) Causal inference in multisensory perception. PLoS ONE 2:e943.

Krause H, Schneider TR, Engel AK, Senkowski D (2012) Capture of visual attention interferes with multisensory speech processing. Front Integr Neurosci 6:67.

Lewkowicz DJ, Schmuckler M, Agrawal V (2021) The multisensory cocktail party problem in adults: Perceptual segregation of talking faces on the basis of audiovisual temporal synchrony. Cognition 214:104743.

Macaluso E (2010) Orienting of spatial attention and the interplay between the senses. Cortex 46:282-297.

Noppeney U (2021) Perceptual Inference, Learning, and Attention in a Multisensory World. Annu Rev Neurosci 44:449-473.

Odegaard B, Shams L (2016) The Brain's Tendency to Bind Audiovisual Signals Is Stable but Not General. Psychol Sci 27:583-591.
Odegaard B, Wozny DR, Shams L (2017) A simple and efficient method to enhance audiovisual binding tendencies. PeerJ 5:e3143.
Ohshiro T, Angelaki DE, Deangelis GC (2011) A normalization model of multisensory integration. Nat Neurosci 14:775-782.
Park H, Kayser C (2019) Shared neural underpinnings of multisensory integration and trial-by-trial perceptual recalibration in humans. eLife 8.
Park H, Kayser C (2020) Robust spatial ventriloquism effect and trial-by-trial aftereffect under memory interference. Sci Rep 10:20826.
Park H, Kayser C (2021) The Neurophysiological Basis of the Trial-Wise and Cumulative Ventriloquism Aftereffects. J Neurosci 41:1068-1079.
Park H, Kayser C (2022) The context of experienced sensory discrepancies shapes multisensory integration and recalibration differently. Cognition 225:105092.
Park H, Nannt J, Kayser C (2021) Sensory- and memory-related drivers for altered ventriloquism effects and aftereffects in older adults. Cortex 135:298-310.
Park H, Kayser C, Thut G, Gross J (2016) Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. eLife 5.
Powers AR, 3rd, Hillock AR, Wallace MT (2009) Perceptual training narrows the temporal window of multisensory binding. J Neurosci 29:12265-12274.
Radeau M, Bertelson P (1974) The after-effects of ventriloquism. The Quarterly journal of experimental psychology 26:63-71.
Raftery AE (1995) Bayesian Model Selection in Social Research. . Sociological Methodology 25:111-163.
Recanzone GH (1998) Rapidly induced auditory plasticity: the ventriloquism aftereffect. Proc Natl Acad Sci U S A 95:869-875.
Rohe T, Noppeney U (2015) Sensory reliability shapes perceptual inference via two mechanisms. J Vis 15:22.
Shams L, Beierholm U (2022) Bayesian causal inference: A unifying neuroscience theory. Neurosci Biobehav Rev 137:104619.
Shavit-Cohen K, Zion Golumbic E (2019) The Dynamics of Attention Shifts Among Concurrent Speech in a Naturalistic Multi-speaker Virtual Environment. Front Hum Neurosci 13:386.
Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 22:1359-1366.
Soto-Faraco S, Lyons J, Gazzaniga M, Spence C, Kingstone A (2002) The ventriloquist in motion: illusory capture of dynamic information across sensory modalities. Brain Res Cogn Brain Res 14:139-146.
Spence C (2009) Explaining the Colavita visual dominance effect. Prog Brain Res 176:245-258.
Spence C, Santangelo V (2009) Capturing spatial attention with multisensory cues: a review. Hear Res 258:134-142.
Stanford TR, Stein BE (2007) Superadditivity in multisensory integration: putting the computation in context. Neuroreport 18:787-792.
Stein BE, Stanford TR (2008) Multisensory integration: current issues from the perspective of the single neuron. Nat Rev Neurosci 9:255-266.
Stein BE, Stanford TR, Ramachandran R, Perrault TJ, Jr., Rowland BA (2009) Challenges in quantifying multisensory integration: alternative criteria, models, and inverse effectiveness. Exp Brain Res 198:113-126.
Stevenson RA, Wallace MT (2013) Multisensory temporal integration: task and stimulus dependencies. Exp Brain Res 227:249-261.
Stevenson RA, Zemtsov RK, Wallace MT (2012) Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. J Exp Psychol Hum Percept Perform 38:1517-1529.
Talsma D, Senkowski D, Soto-Faraco S, Woldorff MG (2010) The multifaceted interplay between attention and multisensory integration. Trends Cogn Sci 14:400-410.

Tong J, Li L, Bruns P, Röder B (2020) Crossmodal associations modulate multisensory spatial integration. Atten Percept Psychophys 82:3490-3506.
Van der Burg E, Olivers CN, Bronkhorst AW, Theeuwes J (2008a) Pip and pop: nonspatial auditory signals improve spatial visual search. J Exp Psychol Hum Percept Perform 34:1053-1065.
Van der Burg E, Olivers CN, Bronkhorst AW, Theeuwes J (2008b) Audiovisual events capture attention: evidence from temporal order judgments. J Vis 8:2.1-10.
Van der Stoep N, Van der Smagt MJ, Notaro C, Spock Z, Naber M (2021) The additive nature of the human multisensory evoked pupil response. Sci Rep 11:707.
van Eijk RL, Kohlrausch A, Juola JF, van de Par S (2008) Audiovisual synchrony and temporal order judgments: effects of experimental method and stimulus type. Percept Psychophys 70:955-968.
Wallace MT, Stevenson RA (2014) The construct of the multisensory temporal binding window and its dysregulation in developmental disabilities. Neuropsychologia 64:105-123.
Wozny DR, Shams L (2011a) Computational characterization of visually induced auditory spatial adaptation. Front Integr Neurosci 5:75.
Wozny DR, Shams L (2011b) Recalibration of auditory space following milliseconds of cross-modal discrepancy. J Neurosci 31:4607-4612.