Visual sensitivity is a stronger determinant of illusory processes than auditory cue parameters in the sound-induced flash illusion

Daniel P. Kumpik

Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, UK   ✉

Helen E. Roberts

Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, UK   ✉

Andrew J. King

Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, UK   ✉

Jennifer K. Bizley

Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, UK
UCL Ear Institute, London, UK   ✉

The sound-induced flash illusion (SIFI) is a multisensory perceptual phenomenon in which the number of brief visual stimuli perceived by an observer is influenced by the number of concurrently presented sounds. While the strength of this illusion has been shown to be modulated by the temporal congruence of the stimuli from each modality, there is conflicting evidence regarding its dependence upon their spatial congruence. We addressed this question by examining SIFIs under conditions in which the spatial reliability of the visual stimuli was degraded and different sound localization cues were presented using either free-field or closed-field stimulation. The likelihood of reporting a SIFI varied with the spatial cue composition of the auditory stimulus and was highest when binaural cues were presented over headphones. SIFIs were more common for small flashes than for large flashes, and for small flashes at peripheral locations, subjects experienced a greater number of illusory fusion events than fission events. However, the SIFI was not dependent on the spatial proximity of the audiovisual stimuli, but was instead determined primarily by differences in subjects' underlying sensitivity across the visual field to the number of flashes presented. Our findings indicate that the influence of auditory stimulation on visual numerosity judgments can occur independently of the spatial relationship between the stimuli.

Introduction

Cross-modal illusions can provide valuable insights into the ways in which sensory inputs are integrated within the brain (Alais & Burr, 2004; Bertelson & Radeau, 1981; Frissen, Vroomen, de Gelder, & Bertelson, 2005; Odgaard, Arieh, & Marks, 2004; Shams, Kamitani, & Shimojo, 2002; Violentyev, Shimojo, & Shams, 2005). In the sound-induced flash illusion (SIFI), first reported by Shams, Kamitani, and Shimojo (2000), an observer's perception of a visual stimulus is influenced by concurrently presented sounds. Thus, when a single brief flash is accompanied by multiple, rapidly presented auditory stimuli within a temporal window of about 100 ms (Shams et al., 2002), the probability of multiple flashes being reported increases substantially (flash fission illusion). Conversely, under some conditions, when flashes outnumber sounds, an observer is more likely to report seeing a single flash (flash fusion; Andersen, Tiippana, & Sams, 2004).

Recent studies have determined that similar effects can be elicited within a modality (Apthorp, Alais, & Boenke, 2013; Bizley, Shinn-Cunningham, & Lee, 2012; Chatterjee, Wu, & Sheth, 2011), suggesting that the same general principles, though not necessarily the same physiological mechanisms, exist for processing brief stimuli both within and across different sensory modalities. Specifically, Bizley et al. (2012) drew parallels between the principles

that guide object formation and those that apparently govern these illusions; under such a scheme SIFIs and their unisensory equivalents are the brain's best guess at grouping rapidly presented sensory information. These authors demonstrated that obligatory, but imperfect, binding of such rapidly presented stimuli occurred when subjects were presented with competing streams of auditory and/or visual stimuli and were directed to try and attend to one stream only when reporting the number of flashes and/or beeps present on each trial. They found that, as with object formation, spatial proximity was an important cue in determining how subjects perceive conflicting stimuli, i.e., a sound on the attended side was more likely to elicit a SIFI than one presented on the unattended side (Bizley et al., 2012). This contrasts with previous work that failed to demonstrate any influence of spatial proximity on the SIFI (Innes-Brown & Crewther, 2009). There are, however, two potentially important differences between these studies. First, the more complex task used by Bizley et al. (2012), where stimuli were presented as competing streams, may have imposed cross-modal grouping rules more strongly. Second, Innes-Brown and Crewther (2009) used tones as their auditory stimuli, whereas Bizley et al. (2012) employed more complex stimuli that potentially elicited a stronger localization percept.

In this study we systematically explore the spatial factors that determine the likelihood of an auditory stimulus influencing visual perception in the SIFI. To this end we manipulated three parameters: (a) the spatial congruence of the auditory and visual stimuli, (b) the type of sound localization cues available by either presenting broadband sounds in the free field or by limiting the binaural cues available over headphones, and (c) the eccentricity and size of the visual stimulus, in order to explore the extent to which underlying visual sensitivity dictated the likelihood of perceiving an illusory flash percept.

Methods

Ethics statement

All subjects provided written informed consent in accordance with the declaration of Helsinki. Ethical approval was provided by the Central University Research Ethics Committee of the University of Oxford.

Participants

A total of 32 subjects (23 female) took part in the study ($M = 20.2$ years, $SD = 1.3$). These were formed into four groups, one for each auditory localization cue type. All subjects were psychophysically naive at the start of testing and reported normal or corrected-to-normal vision and normal hearing.

Stimuli

All stimuli and tasks were generated and run using Matlab (r.2008a, The Mathworks, Natick, MA), and testing was conducted in a darkened, sound-attenuated room with an average background sound level of 35 dB_A. Closed-field auditory stimuli were passed to a Tucker-Davis Technologies (TDT, Alachua, FL) System 3 RM1 mobile processor and presented to subjects over headphones (Sennheiser HD650, Hannover, Germany). Closed-field level calibrations were performed using a Brüel and Kjaer (B&K, Nærum, Denmark) 4134 closed-field condenser microphone and a B&K 4153 artificial ear connected to a B&K 3110-003 measuring amplifier. Free-field auditory stimuli were passed to a TDT System 3 RP2.1 processor and run through an Alesis RA150 amplifier (Alesis, Cumberland, RI) before being presented over one of a pair of Audax (now AAC Acoustic Applications Composites, La Chartre-sur-le-Loir, France) TW025M0 loudspeakers. This setup was also used for free-field loudspeaker calibrations, which were conducted by presenting Golay codes (Zhou, Green, & Middlebrooks, 1992) and recording the loudspeaker output using the B&K 3110-003 measuring amplifier. In this manner, we ensured that the loudspeakers' frequency response characteristics were matched across the two spatial locations. A-weighted level calibrations were performed by positioning a B&K 4191 free-field condenser microphone at the point where the head would be during presentation of a stimulus (see below) and adjusting the presentation level so that it was equivalent for each loudspeaker. All auditory stimuli were presented to the signal processing hardware for digital-to-analogue conversion at a sample rate of 25 kHz. Visual stimuli were generated using the Matlab Psychophysics Toolbox (http://psychtoolbox.org; Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997), and presented using a 24-in. flat-screen LCD monitor at a refresh rate of 60 Hz. For the four test groups, the stimulus configurations were identical, with the exception of the auditory spatial information available.

Visual stimuli were 100%-contrast Gaussian blobs presented against a black background, with 2 SDs of the Gaussian contrast profile subtending either 6° or 13° of visual field at a viewing distance of 70 cm. The blobs were centered at one of four eccentricities (5° or 15° to the left or the right of the midline) at 15° below a fixation cross (see Figure 1A). Flash duration was 16.7 ms (i.e., one frame) and when two consecutive flashes were presented, their onsets were separated by 67 ms.

A

B

Auditory   8   59   8

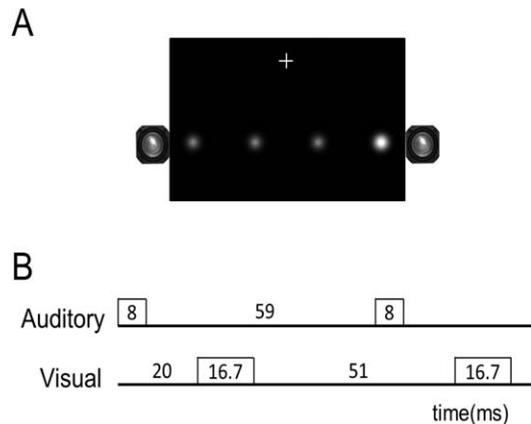Visual   20 16.7   51   16.7

time(ms)

Figure 1. Schematic showing the spatial (A) and temporal (B) structure of the stimuli used in this experiment. (A) Spatial layout of visual stimuli and of loudspeakers used to present free-field stimuli (not to scale). Gaussian blobs with small eccentricities (6°) are shown representing the presentation coordinates for the visual stimuli; grayed-out blobs show the potential positions a flash could take on any given trial, and the lighter blob on the right exemplifies the actual position on the current trial. (B) Temporal structure (in ms) of a trial where two flashes and two auditory stimuli were presented; with the exception of 1V2Ac catch trials (see text), this structure was valid for all trial types but with the second auditory or visual (or both) signal(s) absent when only one stimulus was presented in either modality. Numbers are times in milliseconds.

Subjects' head position was maintained using a forehead and chin rest so that their eyes were level with the fixation cross. On any given trial subjects were always presented with either one or two consecutive blobs in the same visual position.

Closed-field auditory stimuli were 8-ms Gaussian noise pulses with 3-ms cosine rise/fall ramps, presented at an average binaural level of 70 dB$_A$. When two noise pulses were presented, their onsets were separated by 67 ms and both were always presented from the same side. The stimulus-onset asynchrony (SOA) between the first flash and the first noise pulse in any stimulus sequence was 20 ms (controlled using an ActiveX software trigger to the TDT hardware), with the auditory stimulus preceding the flash to match previous studies (Bizley et al., 2012; Innes-Brown & Crewther, 2009; see Figure 1B). For the diotic group, auditory stimuli always had a 0 μs interaural time difference (ITD) and a 0 dB interaural level difference (ILD); for the ITD group, the auditory stimulus on any given trial was lateralized by 153 μs to either the left or the right. The ILD stimuli were high-pass filtered at 1.9 kHz using a finite impulse response filter before imposing an ILD of ±7.5 dB. These binaural cue values were chosen to roughly approximate a stimulus presented at 15° off the midline, so that they coincided with the position of our most lateralized visual stimuli.

Free-field auditory stimulus configurations were identical to those for the unfiltered closed-field stimuli, with the obvious difference being that they contained a rich array of matching ITD, ILD, and spectral localization cues. Loudspeakers were positioned either side of the computer monitor and as close as possible to the center of the most peripheral visual stimuli, the center of the loudspeaker was separated from the center of each visual stimulus by 13 cm (11°), 26 cm (20°), 39 cm (29°), and 52 cm (37°), respectively. Sound levels were calibrated (with the microphone at a location equivalent to the center of the subject's head) to 70 dB$_A$, as in the closed-field groups.

Procedure

Initial training phase

All subjects undertook three tasks designed to familiarize them with the stimuli and procedures and to obtain baseline data prior to entering the experimental phase. They first performed a flash-counting assessment task. This used identical visual stimuli to those used during experimental testing, but in the absence of auditory stimulation, and subjects were simply required to report whether they saw one or two flashes on any given trial. The fixation cross was presented for 2000 ms, and was then followed by the first flash of the sequence. Presentation of the flash(es) was followed by a 2000-ms response window, in which subjects indicated, using the numerical keys on the computer keyboard, whether they had perceived one or two flashes. Subjects attained a mean of 89% ± 0.07 correct.

A second training task was included for the closed-field groups in which subjects completed at least 260 trials in a single interval forced-choice left/right discrimination task, using a constant-stimulus paradigm with the binaural cue that was relevant to the experiment in which they took part. ITD values ranged from ±150 μs in 25 μs steps, and ILD values ranged from ±7.5 dB in 1.25 dB steps. Since both the diotic and ITD stimuli were broadband noise, subjects in the diotic group trained with ITDs to control for any procedural and perceptual learning that may have occurred in the ILD and ITD groups. Given the high localizability (determined from pilot data) and greater ecological validity of the free-field stimuli, simulating a constant-stimulus paradigm for this task was deemed unnecessary, and so on each of the 260 trials the auditory stimulus was simply presented from either one loudspeaker or the other. On each trial, subjects were presented with noise stimuli equivalent to those to be used in audiovisual testing, and were required to indicate, using a computer keyboard, whether they perceived the stimulus to have originated from the left or the right of the midline. Subjects received performance feedback on each trial (for the small number of

closed-field stimuli where the binaural cue values were 0 dB ILD or 0 μs ITD, feedback was pseudorandomized between the two alternatives). The ILD group achieved an overall score of 83.3 ± 6.6% correct ($M \pm SD$) in their final run, with the ITD group scoring 89.4 ± 4.2% correct. The diotic group scored 91.5 ± 3.8% and the free-field group 99.7 ± 0.006%.

The final training task was intended to assist subjects in spatially integrating the visual and (in some cases rather artificial) auditory stimuli that they would encounter in the later experiments. On any given trial, a single flash was presented to one of the two most peripheral positions in the left or right visual field. In the closed-field conditions, a sound was also presented containing either a 153 μs ITD or a 7.5 dB ILD corresponding to the same side as, or the opposite side to, the flash. For the free-field group, a concurrent noise burst was presented from one of the two loudspeakers. Subjects were simply asked to respond with a key-press if they perceived the visual and auditory stimuli to have originated from the same side. This session showed that subjects could discriminate well between congruent and noncongruent audiovisual stimuli, with scores in the final two training sessions of 89.5 ± 5.2% correct (ILD group), 88.1 ± 4.6% correct (ITD group), 84.2 ± 7.5% (diotic group), and 96 ± 8.5% correct (free-field group). In each of these training tasks, subjects were required to attain a score of ≥80% correct before being included in the next phase of the experiment.

Experimental phase

Aside from the nature of the auditory stimulus presentation, the protocol used for all our groups was identical. On any given trial, subjects were presented with a fixation cross for 2000 ms, followed by the trial sequence, which was itself followed by a response period of 2000 ms. Flashes were presented at one of four locations, in conjunction with auditory stimuli that were presented from either the same or the opposite side of the midline to the flash(es). Subjects were required to use the computer keyboard to indicate whether they saw one or two flashes. There were five trial types: one visual, one auditory (1V1A); one visual, two auditory (1V2A; flash fission trials); two visual, one auditory (2V1A; flash fusion trials); two visual, two auditory (2V2A); and catch trials (1V2Ac). This last trial class was identical to the 1V2A trials but with a larger visual-auditory SOA of 200 ms, to check whether subjects were merely biasing their judgments towards the number of auditory stimuli presented. All trial types were counterbalanced across all combinations of both visual stimulus size and presentation position and auditory stimulus position. Thirty repetitions of each of the first four trial types were conducted; for catch trials

this value was 15. Data were collected across 15 13-min blocks with no feedback.

Statistical analysis

We calculated signal detection statistics using a Gaussian model with equal variances for the signal and noise distributions, and with one of the two response alternatives arbitrarily assigned as the signal and the other as the noise. Perceptual sensitivity (d') was defined as $d' = z(\text{hit rate}) - z(\text{false alarm rate})$, where z is the inverse cumulative normal (Wickens, 2001). The position of the observer's decision criterion (response bias) was expressed relative to a point halfway between the signal and the noise distributions; thus, a negative criterion indicated a response bias towards judging that one flash had been presented, and a positive criterion indicated a bias towards judging that two flashes had been presented.

We conducted all parametric testing using R (R Foundation for Statistical Computing, r. 2.15.1) simplifying our statistical models in some cases by collapsing over factor levels that did not contribute significantly to the model deviance by refitting the model and comparing the Akaike Information Criterion (AIC) of each. AIC is a measure of the fit of a model; it penalizes superfluous parameters in the model by adding $2p$ to the deviance of the log-likelihood (where p is the number of model parameters). In the case of two models not differing in deviance, the model with the lowest AIC (i.e., that was the least penalized due to the number of parameters) indicates the most parsimonious model for explaining the variability present in the data (Sakamoto et al., 1986). In doing so, we used maximum-likelihood estimation fitting algorithms to allow for the comparison of models that had different fixed-effects structures before refitting our final model using restricted maximum-likelihood estimation (Crawley, 2002). We followed significant three-way interactions with tests of the simple interactions (unless each factor had two levels), and significant two-way interactions were followed by tests of simple main effects. Posthoc pairwise comparisons among means were made using Tukey's honestly significant difference (HSD) test.

To investigate the interplay between auditory spatial cue values and the reported number of visual illusions, trials were classified according to the degree of congruence (DoC) between the locations of the auditory and visual stimuli. DoC 1 indicates that the auditory and visual stimuli were presented on the same side with the flash at its most peripheral location (i.e., both stimuli from roughly the same location), while DoC 4 indicates that the stimuli for each modality were presented on opposite sides of space, again with the flash at its most peripheral location. DoC 2 and DoC 3
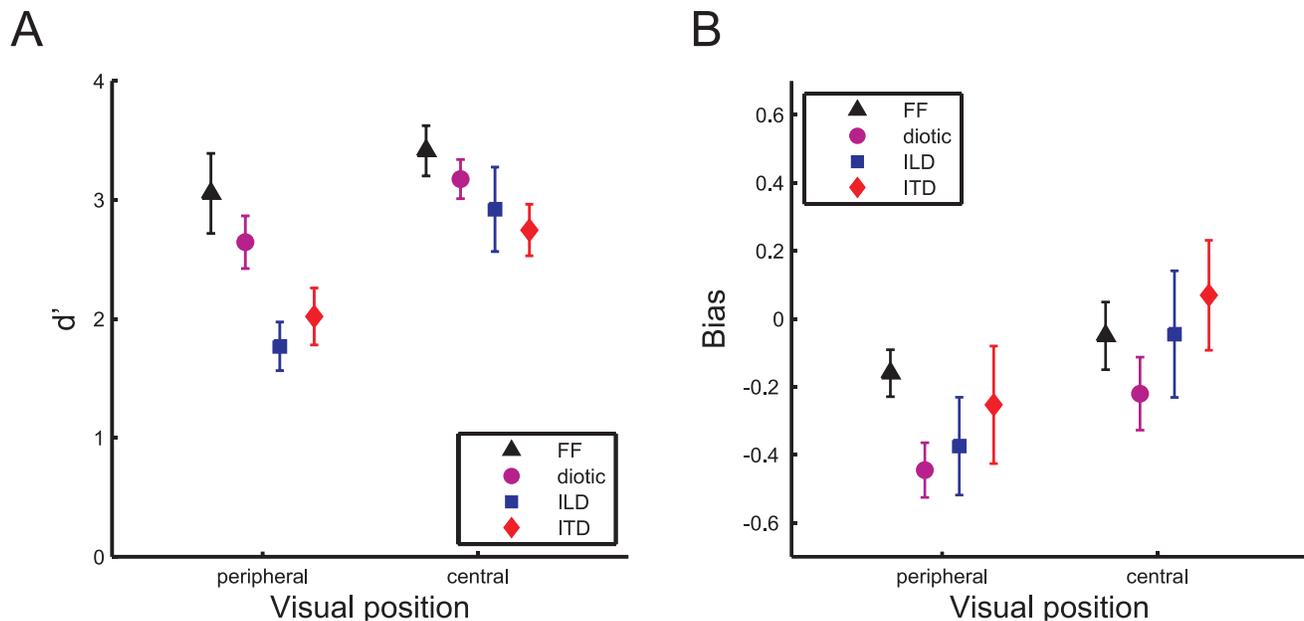
Figure 2. Mean unisensory flash discrimination d' (A) and decision criterion (B) values for peripheral and central visual stimulus locations, for the four auditory localization cue types used in this experiment. d' indicates strength of discrimination performance for one versus two flashes, with $d' = 0$ indicating that subjects could not discriminate the stimuli. Negative decision criterion values indicate that subjects were biased towards reporting one flash, and positive values indicate a bias towards reporting two flashes. FF = free field; ILD = interaural level difference; ITD = interaural time difference.

were the two intermediate conditions, where the flashes were presented at the two central locations: For DoC 2 the flash was presented on the same side of the midline as the auditory stimulus, and for DoC 3 it was presented on the opposite side of the midline. The DoC range used here thus represented a continuum of spatial congruence, but because of within-subject differences in visual sensitivity across the visual field, we treated DoC as a categorical variable in all parametric analyses. We therefore assumed no particular shape to the relationship between DoC and d' or response bias within our data. For the diotic group, where both ITD and ILD were 0, corresponding to sounds being at the center of the subject's head, the four DoC levels did not represent differences in degree of congruence, but instead indicated whether the flash was presented in a peripheral (DoC 1 and 4, less congruent) or a central (DoC 2 and 3, more congruent) position. For the sake of comparison with the groups that were exposed to lateralized auditory stimuli, however, we nevertheless retained this labeling convention for the diotic group.

Results

Baseline flash-only performance

In order to measure subjects' sensitivity to flash number throughout space we performed an initial flash-only assessment. The mean number of flashes reported was 1.08 ± 0.14 ($M \pm SD$) when one flash was presented and 1.86 ± 0.19 when two flashes were presented. We then used these data to analyze the extent to which sensitivity to flash number was determined by visual stimulus position and size. We used signal detection analysis to calculate discriminability (d') and bias indices (towards reporting one or two flashes) for our four experimental groups after collapsing the data across the two central and two peripheral positions and over flash size in order to have sufficient trials to perform this analysis for each subject. Sensitivity to flash number was significantly higher, $F(1, 56) = 15.1$, $p < 0.001$, and bias values significantly weaker (less negative), $F(1, 56) = 6.75$, $p < 0.05$, at the central than at the peripheral positions (Figure 2). Individual variations were observed, however, which led to differences in the mean performance of the experimental groups into which subjects were later placed, $F(3, 56) = 5.83$, $p < 0.01$. The sensitivity values to flash-only stimuli were significantly higher in the group in which free-field auditory stimuli were subsequently presented than in either the ILD ($p < 0.05$) or the ITD ($p < 0.05$) groups, although no other significant differences were found. In order to ensure that these pretest differences in underlying visual sensitivity did not affect our interpretation of differences in the frequency of the SIFI both across space and between groups, we used these training-phase data to control for this in later analysis.
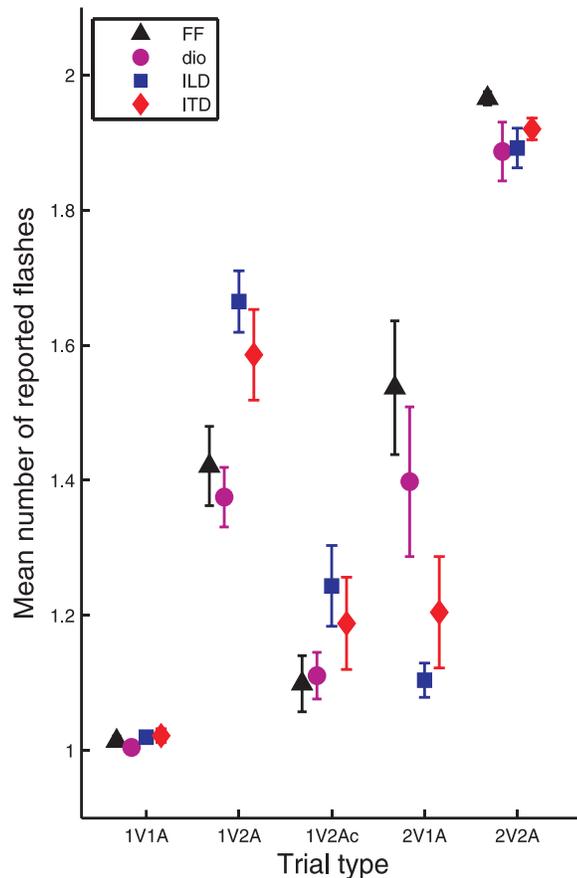
Figure 3. Mean number of flashes reported in each (visual and auditory) trial class for each audiovisual group (FF; dio = diotic presentation; ILD; ITD). Error bars are 1 *SE*. Data have been pooled over flash size and degree of spatial congruence after centering subjects' scores around their own mean to display relative differences between the different auditory localization cue types.

Sound-induced flash illusions

To assess the prevalence of the SIFI we initially calculated the mean number of reported flashes from all trials (i.e., all visual and auditory spatial positions) for each stimulus combination (Figure 3). For both 1V and 2V trials, an incongruent number of auditory stimuli presented in close temporal alignment with the visual stimuli shifted the number of flashes reported in the direction of the number of auditory stimuli presented. On catch trials, the effect of the sound on the number of reported flashes was significantly reduced compared to the 1V2A trials ($p < 0.001$ for each of the four auditory stimulus groups), indicating that illusory flashes were unlikely to have been caused simply by subjects biasing their responses towards the number of auditory stimuli presented.

Effect of the number of auditory stimuli

For this analysis we asked how the presence and nature of an auditory stimulus influenced the likelihood of experiencing the SIFI. To investigate the impact of one or two sounds on discrimination, we collapsed the audiovisual data over visual stimulus size, and over the two peripheral positions and the two central positions. We then computed signal detection statistics and averaged the resulting d' and response bias values over the peripheral and central eccentricities to simplify and focus the analysis.

For all groups, Figure 4A clearly shows that a concurrent auditory stimulus influences visual perception. When innate differences in visual sensitivity are taken into account, similar trends are seen irrespective of the type of auditory spatial information presented (free-field, ITD, ILD, or diotic). However, examination of Figure 4A suggests that the ILD and ITD stimuli might have had a greater effect on visual sensitivity than free-field or diotic sound presentation, since a larger change in d' values was seen when these stimuli were used. Therefore, in order to assess the impact of the presence and nature of the auditory stimuli, we controlled for individual differences by employing a mixed-effects ANOVA with subject as the random effect and with auditory stimulus type (4) and the number of auditory stimuli (3) as the fixed effects.

The number of sounds presented had a significant effect, $F(2, 56) = 37.96$, $p < 0.0001$, with posthoc tests revealing higher d' values when no sound was present than in both the one sound (1A, $p < 0.001$) and two sound (2A, $p < 0.001$) conditions. These analyses thus indicate that the presence of an auditory stimulus reduced the ability of subjects to discriminate one flash from two flashes. The type of auditory presentation used also had a significant effect on the d' values, $F(3, 28) = 6.54$, $p < 0.01$, with lower d' values in the ILD ($p < 0.0001$) and ITD ($p < 0.05$) groups than in the free-field group, although only the ILD group's sensitivity scores were lower than those of the diotic group (ILD: $p < 0.05$; ITD: $p = 0.17$). However, the relative effect of the number of auditory stimuli was the same across the four subject groups, $F(6, 56) = 1.02$, $p = 0.42$, regardless of individual differences in overall sensitivity between the groups.

Subjects' response biases were also affected by the number of sounds, $F(2, 56) = 216$, $p < 0.0001$, with the presence of one sound significantly biasing subjects' responses towards reporting that they had seen one flash and the presence of two sounds resulting in a bias towards reporting that they had seen two flashes (Figure 4B). No difference in response bias was seen among the four groups, $F(3, 28) = 2.47$, $p = 0.08$, and the effect of number of auditory stimuli on the bias was not affected by the type of auditory stimulus presentation, $F(6, 56) = 1.71$, $p = 0.13$.
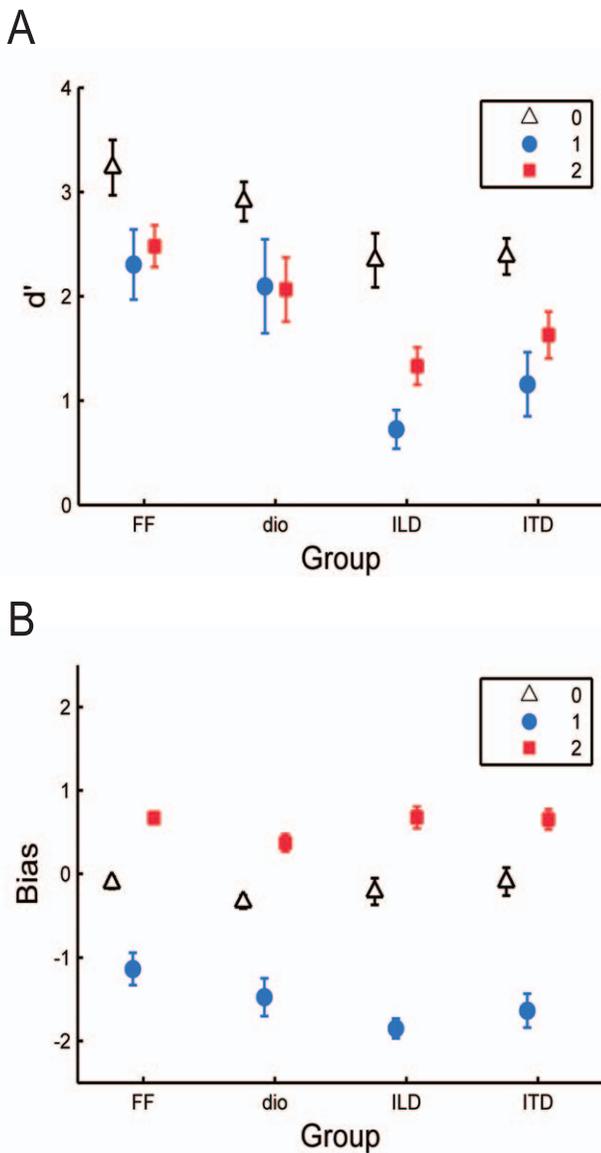
A

B

Figure 4. Mean flash discrimination d' (A) and decision criterion (B) values for each audiovisual group (FF; dio; ILD; ITD), plotted as a function of the number (0, 1, or 2) of concurrent auditory stimuli. d' and decision criterion values are defined as for Figure 2. Error bars are 1 *SE*.

Overall, this analysis indicates that (a) the presence of auditory stimuli both reduced visual perceptual sensitivity and biased subjects' responses towards the number presented, and (b) that both of these effects were consistent regardless of the type of sound localization or lateralization cue that was presented.

Effect of audiovisual spatial congruence

We next explored the extent to which spatial congruence between the visual and auditory stimuli influenced the SIFI. Because signal detection analysis

of our unisensory visual data demonstrated that subjects were generally more sensitive to, and had different relative biases for, central visual positions compared to those at the periphery, we used each subjects' flash-only visual sensitivity score as a random predictor within a mixed-effects ANOVA where the other factors—sound localization cue type (4) × DoC (4) × flash size (2) × number of auditory stimuli (2)— were fixed predictors. Prior to this we confirmed that splitting the data by left-sided versus right-sided flashes and including this as a parameter in the model did not improve the model fit, indicating that the data were symmetrical across the midline. Because each DoC was associated with two different visual positions that were always either peripheral or central, for each subject and DoC we used the flash-only d' or response bias values obtained—after collapsing across the midline for the peripheral positions and also for the central positions— as our random predictors. This allowed us to control the audiovisual data for differences in visual sensitivity and bias that were both position-specific (between central and peripheral positions) within an individual, and that were attributable to differences between individuals.

Fitting a conventional fixed-effects model to these data with no attempt to control for subjects' differences in visual sensitivity indicated a significant main effect of DoC, $F(3, 960) = 13.95$, $p < 0.0001$. However, it can clearly be seen from Figure 5A that spatial congruence does not influence sensitivity to flash number (which would most likely present as a difference in d' between DoCs 1 and 4). Instead, we found that sensitivity was highest for central compared to peripheral positions, as was also the case in the unisensory visual data. This occurred in all conditions apart from those associated with fission (2A) trials at the smaller flash size. By comparison with the ANOVA incorporating fixed effects only, when visual sensitivity was included as a random effect in the mixed-effects model we saw no effect of DoC on d', $F(3, 920) = 0.17$, $p = 0.91$.

The interaction between the effects of flash size and the number of auditory stimuli was not the same at each DoC, $F(3, 920) = 2.7$, $p < 0.05$. This can be seen in Figure 5A, with visual sensitivity remaining constant across DoC for each group in the fission condition (2A) at the smaller flash size (bottom left panel), whereas in the other three condition combinations, visual sensitivity was lower at peripheral visual positions than at central positions, giving an inverted "U" shape to the data. Prior to investigating this interaction further, we noted that the d' scores in Figure 5A are distributed almost symmetrically across the midline DoC values. We therefore simplified the statistical model by collapsing the levels of DoC into peripheral and central positions. We compared the model deviance values and confirmed that this simplification was justified, before
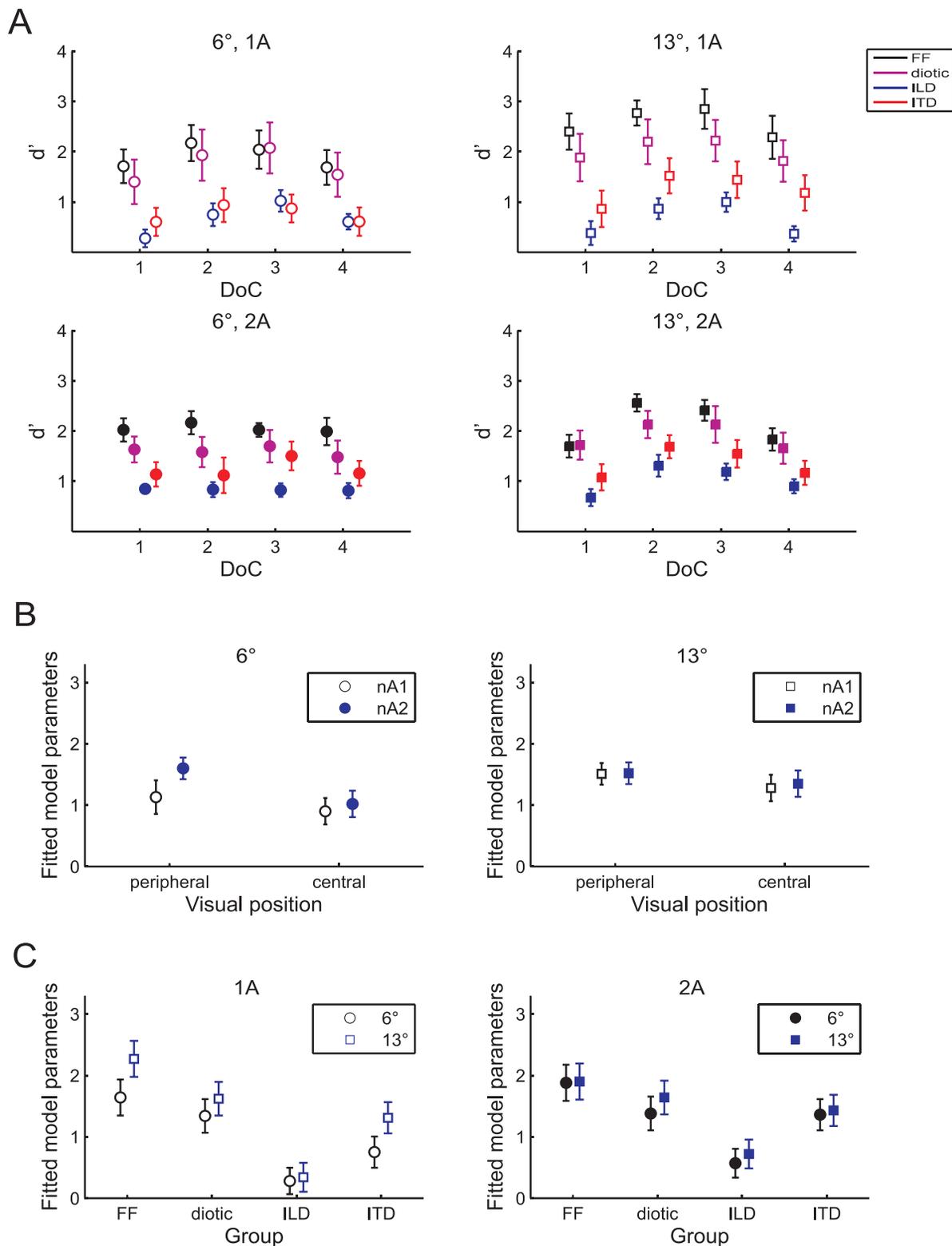
Figure 5. Visual sensitivity (d') scores for each audiovisual group (FF; dio; ILD; ITD). Circles denote small flashes and squares denote large flashes, while open symbols denote 1A stimuli and closed symbols denote 2A stimuli. (A) Mean flash discrimination d' values, plotted as a function of degree of spatial congruence (DoC), flash size, and the number of concurrent auditory pulses (1A or 2A). d' values are defined as for Figure 2. Error bars are 1 *SE*. (B) Fitted mixed-effects model parameters and associated 95% confidence intervals for the three-way interaction between visual position, flash size, and the number of concurrent auditory pulses. (C) Fitted mixed-effects model parameters and associated 95% confidence intervals for the three-way interaction between sound localization cue type, flash size, and the number of concurrent auditory pulses.

proceeding with pairwise comparisons conducted on the new model.

Figure 5B shows the model parameters (i.e., d' values) for the two visual flash sizes at central and peripheral visual stimulus locations after taking innate differences in visual sensitivity (Figure 2A) into account. Whereas the raw data in Figure 5A suggest that sensitivity in the presence of auditory stimuli is higher at central than at peripheral positions, this is not born out once this adjustment has been made. Thus, d' values did not differ between central and peripheral positions (small flash, 1A: $p = 0.42$; larger flashes: 1A, $p = 0.39$; 2A, $p = 0.79$), with the exception of 2A fission trials at the smaller flash size ($p < 0.0001$). Thus, differences in the likelihood of perceiving an illusion at different positions in the visual field can largely be accounted for by subjects' unisensory visual sensitivity.

As suggested by Figure 5B, there were no differences in sensitivity with the number of auditory stimuli at either peripheral ($p = 1$) or central ($p = 0.99$) locations for the larger flashes, whereas for smaller flashes, the difference between 1A and 2A trials was significant at peripheral locations ($p < 0.0001$) but not at central locations ($p = 0.88$). Flash size also affected d' values, $F(1, 920) = 38.3$, $p < 0.0001$, with higher sensitivity for larger visual stimuli at both central ($p < 0.001$) and peripheral ($p < 0.001$) locations on trials where one sound was presented. This was also the case for central ($p < 0.01$) but not peripheral ($p = 0.99$) trials accompanied by 2A stimuli. In summary, across all acoustic localization cues, visual sensitivity was a major contributing factor to the tendency for subjects to perceive sound-induced illusions. Additionally, the likelihood of perceiving a sound-induced illusion was greater for small flashes than for large flashes, and for small flashes at peripheral locations subjects experienced a greater number of illusory fusion events than fission events.

Sensitivity measures for fusion (1A) versus fission (2A) stimuli also differed with flash size across the four sound presentation groups, $F(3, 952) = 4.44$, $p < 0.01$ (Figure 5C). In particular, the ILD group had significantly lower sensitivity than the other groups for 1A and 2A stimuli when small (all ps < 0.05) and large flashes (all ps < 0.01) were presented. Moreover, differences in sensitivity between large and small flash sizes were observed for fusion stimuli ($p < 0.0001$) in two of the groups, with both the free-field ($p < 0.01$) and ITD ($p < 0.01$) groups displaying significantly better visual sensitivity with the larger flash size.

We found that subjects' response bias towards reporting one or two flashes in each condition was clearly affected by whether one or two auditory stimuli were present (see Figure 6A). As with the d' values, a significant main effect of DoC was observed, $F(3, 960) = 7.98$, $p < 0.0001$, whereas this was no longer the case

if each subject's unisensory bias values for peripheral or central locations were included as a random effect in the mixed-effects ANOVA, $F(3, 917) = 0.27$, $p = 0.84$. Therefore, the degree of spatial congruence between the auditory and visual stimuli did not significantly influence either the sensitivity or bias observed in our subjects. Because the bias appeared to be distributed symmetrically as a function of DoC (Figure 6A), we again simplified the model by collapsing DoC across the midline into peripheral and central visual presentation positions.

We observed a significant interaction between group and the number of auditory stimuli presented, $F(1, 949) = 68.96$, $p < 0.0001$ (Figure 6B). With one sound, the diotic and ITD groups produced a similar bias towards saying "one flash" ($p = 0.94$). Both of these groups exhibited less bias than the ILD group ($p < 0.05$ and $p < 0.01$, respectively), but more bias than the free-field group (both $p < 0.0001$). When two sounds were presented, only the diotic group differed significantly from any of the others (all $p < 0.001$).

A significant interaction between group and flash size was found, $F(3, 949) = 14.24$, $p > 0.0001$ (Figure 6C), although pairwise comparisons indicated that the bias profile across groups was similar for both small (all $p > 0.27$) and large (all $p > 0.2$) flashes. Only the ITD group displayed a difference in response bias with flash size ($p < 0.05$).

Discussion

Motivated by apparently conflicting observations in the literature (Bizley et al., 2012; Innes-Brown & Crewther, 2009), we aimed to determine whether spatial congruence between auditory and visual stimuli influences the likelihood of perceiving a sound-induced flash illusion. Our central finding is that an observer's underlying visual sensitivity to the presence of one or two light flashes is the best predictor of the likelihood of perceiving a sound-induced visual illusion. We found that diotic, dichotic, and free-field auditory stimuli all elicited SIFIs, indicating that the spatial cue content of the auditory stimulus is not a critical factor in affecting the auditory capture of visual stimuli in time, although the illusory percept was most likely to occur when binaural cues, particularly ILDs, were presented over headphones. We also found that the degree of spatial congruence between the auditory and visual stimuli does not influence the likelihood of a SIFI being experienced.

Human observers often act as Ideal Observers in the Bayesian sense (Alais & Burr, 2004; Deneve & Pouget, 2004). The reliability with which a stimulus can be detected or discriminated is known to influence the
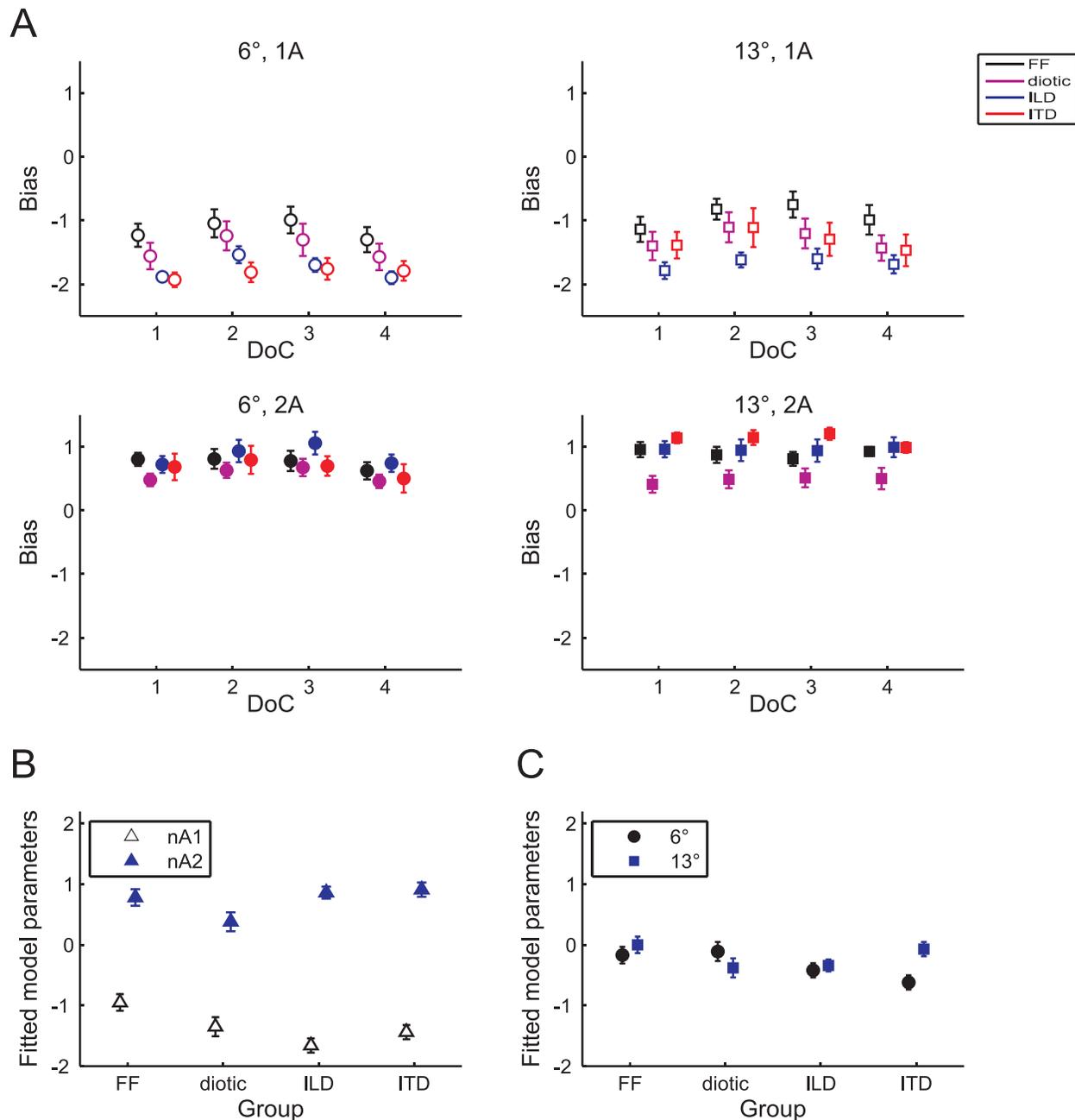
Figure 6. Response bias scores for each audiovisual group (FF; dio; ILD; ITD). Circles denote small flashes and squares denote large flashes, while open symbols denote 1A stimuli, and closed symbols denote 2A stimuli. (A) Mean flash discrimination decision criterion values, plotted as a function of DoC, flash size, and the number of concurrent auditory pulses. Decision criterion values are defined as for Figure 2. Error bars are 1 *SE*. (B) Fitted mixed-effects model parameters and associated 95% confidence intervals for the two-way interaction between sound localization cue type and the number of concurrent auditory stimuli. (C) Fitted mixed-effects model parameters and associated 95% confidence intervals for the two-way interaction between sound localization, cue type, and flash size.

weighting that stimulus is given in a multisensory decision-making process (Burr & Alais, 2006). Using a visual-only paradigm, we were able to estimate each subject's innate visual sensitivity for discriminating between one and two flashes at each of the flash locations tested. In keeping with other studies showing that visual sensitivity declines with eccentricity (Aubert

& Foerster, 1857; Daniel & Whitteridge, 1961), we showed using signal detection theory that subjects were both less able to identify the number of flashes presented and more likely to report having seen only one flash at peripheral locations. The presence of concurrent auditory stimuli affects both sensitivity and bias measures, with a conflicting number of auditory

and visual stimuli resulting in a decrease in visual sensitivity, single acoustic events biasing subjects towards reporting a single flash, and double acoustic events biasing subjects towards reporting two flashes. While we did not monitor subjects' eye position, had the subjects looked at the flashes, rather than the fixation point, their visual sensitivity would have been increased, reducing the likelihood of a SIFI.

These findings support a Bayesian framework for multisensory integration; the more unreliable the visual stimulus becomes, the greater the influence of the auditory stimulus on visual perception (Figure 5). Unless we took differences in visual sensitivity into account, systematic differences in the probability of a SIFI were observed with the degree of audiovisual spatial congruence. However, when we factored out the innate ability of each subject to discriminate flash number at each location, we were no longer able to observe any influence of either the location of the sound source, or the spatial congruence between the sound source and the visual flash, on discrimination ability. Although our auditory stimuli were not precisely matched to our flash stimuli in physical location, the failure to demonstrate an effect of spatial congruence on the SIFI is strikingly consistent across conditions—including the free-field condition, where two of the four possible visual positions were more closely matched to the physical locations of the two loudspeakers. Our pretesting demonstrated that subjects could accurately identify as congruent stimuli from the same side in all testing conditions. Therefore, we do not think that close matching the physical location of the auditory-visual stimulus pairs would have had led to a stronger effect of spatial proximity on the SIFI.

The findings reported here are in keeping with Innes-Brown and Crewther's (2009) report that spatial location did not influence the likelihood of perceiving a sound-induced illusion. However, in contrast to their study and the present results, when two competing, spatially separated stimulus streams are present, spatial location *does* influence the likelihood of perceiving a SIFI (Bizley et al., 2012). It seems likely that the key difference between these studies is the presence or absence of the competing stimulus stream, which imposes stronger spatial demands on what is otherwise a task that requires judging one from two rapidly presented events and therefore utilizes predominantly temporal information. In support of this argument, studies of temporal recalibration have demonstrated that while the effects on perceived simultaneity are robust to changes in the location of the adapting and test stimuli (Keetels & Vroomen, 2007; Roseboom & Arnold, 2011), spatially specific temporal recalibration can occur when conflicting asynchronies are presented at two locations simultaneously (Heron, Roach,

Hanson, McGraw, & Whitaker, 2012). Another further possibility is raised by the fact that audiovisual binding can be context-dependent (Nahorna, Berthommier, & Schwartz, 2012).

Spence (2013) recently argued that for spatial congruence to influence task performance, that task must usually be explicitly spatial in nature (e.g., spatial discrimination, overt orienting), or at least involve covert spatial cueing of attention. Because subjects in our task could not predict which of the four possible locations the flashes would occur at, they were required to deploy attention across the frontal portion of space bounded by the monitor and loudspeakers. This differs from the study of Bizley et al. (2012), where subjects were required to engage endogenous spatial attention in a location-specific manner and to perform an explicitly spatial task. Under such conditions the acoustic stimulus on the attended side of space had a more powerful effect on visual perception than the acoustic stimulus on the to-be-ignored side. However, in keeping with the current study, for experiments using unisensory stimuli, spatial locations away from the attended location are able to elicit illusory percepts (Apthorp et al., 2013; Bizley et al., 2012). For example, when subjects are presented with only visual or only acoustic stimuli and cued to report the number of stimuli on one side, events on the to-be-ignored side are able to elicit illusory flashes/sounds at the attended location (Bizley et al., 2012). The lack of an effect of spatial congruence in the absence of a second competing stimulus stream suggests that, at least for concurrent audiovisual stimuli with a highly salient temporal component (i.e., short stimulus duration and interstimulus interval), the ability of a highly localizable auditory stimulus to influence visual numerosity judgments does not depend on whether those stimuli have a common spatial origin or are displaced in space to a large degree.

Despite the large differences in protocols and stimulus design between the two studies, our data agree well with the findings of Innes-Brown and Crewther (2009). Although the minimum spatial separation between the auditory and visual stimuli employed in this experiment is plausibly within the range over which the ventriloquism illusion operates (Bertelson & Radeau, 1981), unpublished observations from our pilot studies suggest that these results hold over a far greater range of sound locations (i.e., throughout the whole of auditory space, including the posterior hemifield). This suggests that, for rapidly presented cues, object formation is mainly determined by temporal features—principally temporally correlated onsets—and that once stimuli are automatically bound, conflicting cue values are resolved by more heavily weighting the more reliable auditory estimate. The weighting given to the auditory stimulus is dependent on the relative reliability

of the visual stimulus; as visual sensitivity declines in the periphery the auditory cues are given increasing weight (Figures 2 and 5). In support of a role for object formation or perceptual grouping in guiding SIFIs, a recent study has demonstrated that when the two illusion-inducing sounds have different features, they no longer elicit an illusory flash (Roseboom, Kawabe, & Nishida, 2013). Taken together with evidence for the importance of stimulus context and the absence of a requirement for spatial co-localization, these observations support the idea that multisensory integration of such stimuli is a mid-level process (Fujisaki, Koene, Arnold, Johnston, & Nishida, 2006; Roseboom et al., 2013; Roseboom, Nishida, Fujisaki, & Arnold, 2011), rather than a low-level, automatic or high-level, cognitive process.

We found that subjects experienced at least as many fusion illusions as fission illusions. Previous studies vary in the proportion of fusion illusions reported. In some studies, these have been observed infrequently or not at all (Innes-Brown & Crewther, 2009; Shams et al., 2000; Watkins, Shams, Tanaka, Haynes, & Rees, 2006), whereas, in others, they were equally common as or even more prevalent than fission illusions (Andersen et al., 2004; Apthorp et al., 2013; Bizley et al., 2012). Apthorp et al. (2013) suggested that sound-induced fusion illusions might be more likely if the auditory stimuli are more salient than the visual stimuli, which could be the case when Gaussian blob stimuli are used. Like Apthorp et al. (2013), we also used Gaussian blobs, but this in itself is unlikely to be the underlying reason since fusion illusions were also reported by Andersen et al. (2004) and Bizley et al. (2012), who used hard-edged disks that are likely to be more salient visual stimuli.

While spatial congruence did not influence the likelihood of perceiving a SIFI, the type of lateralization cue did appear to influence the proportion of trials on which an illusion was reported. ILD stimuli were more likely to induce SIFIs than diotic or free-field stimuli, with ITD stimuli eliciting an intermediate number of SIFIs. Apthorp et al. (2013) proposed that the relative salience of the inducer and target stimuli determine the strength of the SIFI. While we have no direct measure of saliency, our closed-field ILD stimuli will have provided the loudest monaural stimulus (77.5 dB$_A$, in the louder channel, 63.5 dB$_A$ in the quieter channel). This is because we calibrated the sound levels in the free-field setup to match the average binaural level for closed field stimuli (70 dB$_A$). Because previous work has demonstrated that the level of the acoustic cue can have a strong influence on the likelihood of a SIFI, with very quiet sounds generating a reverse flash-beep illusion (Andersen et al., 2004), we suggest that the increased tendency for ILD stimuli to elicit SIFIs relates to the more intense monaural sound

levels in this condition, which provide potentially more salient stimuli. Consistent with saliency playing a key role in determining multisensory integration, the importance of saliency in auditory-visual binding has previously been demonstrated for the detection of auditory-visual synchrony (Fujisaki et al., 2006).

The mechanisms underlying illusory flash percepts remain uncertain; while the illusory flash has been shown to have a measurable contrast (McCormick & Mamassian, 2008), experienced subjects are able to determine differences between real and illusory flashes when given a third ("not one, not two") response option (van Erp, Philippi, & Werkhoven, 2013). Neural imaging experiments demonstrate that illusory flashes are accompanied by activity in visual cortex, but that this activity differs from that elicited by real flashes (Mishra, Martinez, & Hillyard, 2008; Mishra, Martinez, Sejnowski, & Hillyard, 2007; Shams, Iwaki, Chawla, & Bhattacharya, 2005; Shams, Kamitani, Thompson, & Shimojo, 2001). Similar illusions can be elicited in both visual-only (Apthorp et al., 2013; Chatterjee et al., 2011; Bizley et al., 2012) and auditory-only (Bizley et al., 2012) conditions. Our data therefore support the findings of previous studies (Apthorp et al., 2013; Bizley et al., 2012), which have argued that SIFIs are perceptual confusions that result from uncertainty about rapidly presented stimuli.

Conclusions

In summary, our data demonstrate that the SIFI can be elicited by auditory stimuli presented either in the free field or over headphones with more limited spatial cues. By varying stimulus position and size on flash-only trials, we were able to show that the ability of our subjects to discriminate between one and two flashes declined at peripheral locations, and that this variation in visual sensitivity across space was the primary factor determining whether they perceived the SIFI. After taking this into account, we found no effect of stimulus location or binaural cue value on the likelihood of experiencing the SIFI, suggesting that, for rapidly presented stimuli, spatial congruence per se is a poor determinant of audiovisual binding. Although the SIFI is unlikely to reflect a generalized sound-induced enhancement of attention (Shams et al., 2002), there is evidence that attentional mechanisms can modulate the likelihood of the illusion occurring (Kamke, Vieth, Cottrell, & Mattingley, 2012; Mishra, Martinez, & Hillyard, 2010). Combining our findings with other recent evidence (e.g., Bizley et al., 2012) suggests that the presence of a competing stimulus stream and/or spatial attention may be key requirements for observing an influence of spatial congruence on the SIFI.

Acknowledgments

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*(3), 257–262.

Andersen, T. S., Tiippana, K., & Sams, M. (2004). Factors influencing audiovisual fission and fusion illusions. *Brain Research: Cognitive Brain Research*, *21*(3), 301–308.

Apthorp, D., Alais, D., & Boenke, L. T. (2013). Flash illusions induced by visual, auditory, and audiovisual stimuli. *Journal of Vision*, *13*(5):3, 1–15, http://www.ncbi.nlm.nih.gov/pubmed/23547105, doi:10.1167/13.5.3. [PubMed] [Article]

Aubert, H., & Foerster, R. (1857). Beitraege zur Kenntnis des indirekten Sehens. I. Ueber den Raumsinn der Retina [Trans. Contributions to the knowledge of indirect sight. I. About the retina's spatial sense]. *Albert von Graefes Archliv fuer Ophthalmologie*, *3*, 1–37.

Bertelson, P., & Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. *Perception & Psychophysics*, *29*(6), 578–584.

Bizley, J. K., Shinn-Cunningham, B. G., & Lee, A. K. (2012). Nothing is irrelevant in a noisy world: Sensory illusions reveal obligatory within- and across-modality integration. *Journal of Neuroscience*, *32*(39), 13402–13410.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436.

Burr, D., & Alais, D. (2006). Combining visual and auditory information. *Progress in Brain Research*, *155*, 243–258.

Chatterjee, G., Wu, D. A., & Sheth, B. R. (2011). Phantom flashes caused by interactions across visual space. *Journal of Vision*, *11*(2):14, 1–17, http://www.journalofvision.org/content/11/2/14, doi:10.1167/11.2.14. [PubMed] [Article]

Crawley, M. J. (2002). *Statistical computing: An introduction to data analysis using S-Plus.* New York: John Wiley & Sons.

Daniel, P. M., & Whitteridge, D. (1961). The representation of the visual field on the cerebral cortex in monkeys. *Journal of Physiology*, *159*, 203–221.

Deneve, S., & Pouget, A. (2004). Bayesian multisensory integration and cross-modal spatial links. *Journal of Physiology, Paris*, *98*(1–3), 249–258.

Frissen, I., Vroomen, J., de Gelder, B., & Bertelson, P. (2005). The aftereffects of ventriloquism: Generalization across sound-frequencies. *Acta Psychologica (Amsterdam)*, *118*(1–2), 93–100.

Fujisaki, W., Koene, A., Arnold, D., Johnston, A., & Nishida, S. (2006). Visual search for a target changing in synchrony with an auditory signal. *Proceedings of the Royal Society B: Biological Sciences*, *273*(1588), 865–874.

Heron, J., Roach, N. W., Hanson, J. V., McGraw, P. V., & Whitaker, D. (2012). Audiovisual time perception is spatially specific. *Experimental Brain Research*, *218*(3), 477–485.

Innes-Brown, H., & Crewther, D. (2009). The impact of spatial incongruence on an auditory-visual illusion. *PLoS One*, *4*(7), e6450.

Kamke, M. R., Vieth, H. E., Cottrell, D., & Mattingley, J. B. (2012). Parietal disruption alters audiovisual binding in the sound-induced flash illusion. *NeuroImage*, *62*(3), 1334–1341.

Keetels, M., & Vroomen, J. (2007). No effect of auditory-visual spatial disparity on temporal recalibration. *Experimental Brain Research*, *182*(4), 559–565.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3?. *Perception 36 ECVP Abstract Supplement.*

McCormick, D., & Mamassian, P. (2008). What does the illusory-flash look like? *Vision Research*, *48*(1), 63–69.

Mishra, J., Martinez, A., & Hillyard, S. A. (2008). Cortical processes underlying sound-induced flash fusion. *Brain Research*, *1242*, 102–115.

Mishra, J., Martinez, A., & Hillyard, S. A. (2010). Effect of attention on early cortical processes associated with the sound-induced extra flash illusion. *Journal of Cognitive Neuroscience*, *22*(8), 1714–1729.

Mishra, J., Martinez, A., Sejnowski, T. J., & Hillyard, S. A. (2007). Early cross-modal interactions in auditory and visual cortex underlie a sound-induced visual illusion. *Journal of Neuroscience*, *27*(15), 4120–4131.

Nahorna, O., Berthommier, F., & Schwartz, J. L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. *Journal of the Acoustical Society of America*, *132*(2), 1061–1077.

Odgaard, E. C., Arieh, Y., & Marks, L. E. (2004). Brighter noise: Sensory enhancement of perceived loudness by concurrent visual stimulation. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(2), 127–132.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.

Roseboom, W., & Arnold, D. H. (2011). Twice upon a time: Multiple concurrent temporal recalibrations of audiovisual speech. *Psychological Science*, *22*(7), 872–877.

Roseboom, W., Kawabe, T., & Nishida, S. (2013). The cross-modal double flash illusion depends on featural similarity between cross-modal inducers. *Science Reports*, *3*, 3437.

Roseboom, W., Nishida, S., Fujisaki, W., & Arnold, D. H. (2011). Audio-visual speech timing sensitivity is enhanced in cluttered conditions. *PloS One*, *6*(4), e18309.

Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike information criterion statistics*. Tokyo: KTK Scientific Publishers.

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.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions: What you see is what you hear. *Nature*, *408*(6814), 788.

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Brain Research: Cognitive Brain Research*, *14*(1), 147–152.

Shams, L., Kamitani, Y., Thompson, S., & Shimojo, S. (2001). Sound alters visual evoked potentials in humans. *Neuroreport*, *12*(17), 3849–3852.

Spence, C. (2013). Just how important is spatial coincidence to multisensory integration? Evaluating the spatial rule. *Annal of the New York Academy of Sciences*, *1296*, 31–49.

van Erp, J. B., Philippi, T. G., & Werkhoven, P. (2013). Observers can reliably identify illusory flashes in the illusory flash paradigm. *Experimental Brain Research*, *226*(1), 73–79.

Violentyev, A., Shimojo, S., & Shams, L. (2005). Touch-induced visual illusion. *Neuroreport*, *16*(10), 1107–1110.

Watkins, S., Shams, L., Tanaka, S., Haynes, J. D., & Rees, G. (2006). Sound alters activity in human V1 in association with illusory visual perception. *Neuroimage*, *31*(3), 1247–1256.

Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.

Zhou, B., Green, D. M., & Middelbrooks, J. C. (1992). Characterization of external ear impulse responses using Golay codes. *The Journal of the Acoustical Society of America*, *92*(2, Pt 1), 1169-1171.