On the Mechanisms of Audio-Visual Integration for Manual Interception

Philipp Kreyenmeier[1,2], Anna Schroeger[3,4], Rouwen Cañal-Bruland[4], Markus Raab[5,6], & Miriam Spering[2,7,8]

[1]Graduate Program in Neuroscience, University of British Columbia

[2]Department of Ophthalmology & Visual Sciences, University of British Columbia

[3]Deptartment of Psychology, Justus-Liebig University

[4]Deptartment for the Psychology of Human Movement and Sport, Friedrich-Schiller University

[5]Deptartment of Performance Psychology, German Sport University

[6]School of Applied Sciences, London South Bank University

[7]Djavad Mowafaghian Centre for Brain Health, University of British Columbia

[8]Institute for Computing, Information, and Cognitive Systems, University of British Columbia

Author Note

P.K. and A.S. share first authorship. The authors thank Anna Montagnini for helpful discussions. Correspondence should be addressed to Philipp Kreyenmeier, Koerner Pavilion, UBC Hospital, S257-211 Wesbrook Mall, Vancouver, BC, V6K 1E1, Canada.

The study reported in this article was not preregistered. Data and analyses will be made available on the Open Science Framework upon publication of the manuscript. The authors declare no competing financial interests.

Abstract

Natural interception actions such as catching a ball require accurate motion prediction and typically involve multiple senses. Laboratory interception studies focus on vision and commonly ignore sound. Here we ask how visual and auditory signals are integrated to guide interception. Observers tracked the brief launch of a simulated baseball randomly paired with batting sounds of varying intensities and made a quick pointing movement at the ball. Interception endpoints revealed systematic overestimation of target speed when ball launch was paired with a loud versus a quiet sound, even though sound was never informative. We applied a Bayesian observer model showing that audio-visual integration, strongest at the shortest presentation duration, depended on visual certainty. Continuous eye movements uncovered that sound influenced visual speed estimates within 125 ms of ball launch. Our findings suggest that auditory and visual signals are rapidly and optimally weighed to guide interception.

Keywords: multisensory integration, prediction, eye movements, interception, perception-action

On the Mechanisms of Audio-Visual Integration for Manual Interception

When intercepting a rapidly moving object—swapping a fly, catching a ball, or hunting prey—we rely heavily on vision. It is well known that humans and other animals direct their eyes at moving objects of interest to sample critical visual information and to increase performance accuracy (e.g., Borghuis & Leonardo, 2015; Diaz et al., 2013; Michaiel et al., 2020; Spering et al., 2011; for review, see Fooken et al., 2021). However, other sensory modalities also supply information that might be used to guide behaviour in interception tasks. Indeed, in goalball—an interceptive sport for visually-impaired athletes—players rely solely on auditory information to locate and intercept a ball (International Blind Sports Federation, 2022). Our study addresses the question whether and how vision-guided interceptive actions rely on sound information in normally-sighted observers.

In our natural environment, object motion is almost always accompanied by sound. Hitting a ball with a bat or racket creates an impact sound, and its volume provides information about hit intensity and launch speed. Accordingly, impact sounds can bias perceived ball-bounce locations and perceptual ball speed judgements, suggesting that observers use auditory information to predict ball trajectories (Cañal-Bruland et al., 2018; 2022). In general, foundational studies in human perception have shown that sounds can alter visual motion judgements. When two identical targets move horizontally toward each other and cross, observers might perceive either a bounce or a pass-through (motion-bounce illusion; Metzger, 1934). A sound coinciding with the crossing point commonly results in a perceived bounce, indicating that sound can critically alter the interpretation of a visual event and thus the perceived direction of

motion (Sekuler et al., 1997). Similar sound-based alterations of visual perception have been obtained for tasks involving visual target speed (Meyerhoff et al., 2022) and acceleration (Wessels et al., 2022). In summary, there is ample evidence supporting the idea that visual and auditory information are integrated to inform perceptual judgements of an object's motion (Chaplin et al., 2016; Soto-Faraco et al., 2003).

Arguably, the effect of auditory information on perception could be tightly linked to visual uncertainty. Auditory signals bias perception the most in tasks in which visual information is ambiguous (such as the motion-bounce illusion), or when visual information is sparse. Generally, cues from different sensory modalities are combined by weighting them according to their internal noise, i.e., their uncertainty (Alais & Burr, 2004; Ernst & Banks, 2002; Körding et al., 2007). If viewing conditions are poor, and a visual signal is thus uncertain, observers may rely more on auditory information than they would if visual certainty was high.

Our study probes this interaction between visual certainty and auditory cues in a real-world inspired, fast-paced movement interception task. It also tests directly whether principles of audio-visual integration, previously demonstrated for pure perceptual tasks, apply to motor actions. We manipulated visual presentation time of a simulated baseball flight curve and sound intensity associated with its simulated launch. In this challenging task, we measured observers' eye and hand movements towards the ball as indicators of their ability to estimate its speed, and to predict its trajectory. We hypothesized, first, that auditory cues would systematically bias ball speed estimation, implying that observers would overestimate speed when ball launch was accompanied by a loud as compared to a quiet batting sound. Second, we expected that the auditory cue's

influence would scale with visual certainty, implying that observers rely more on the auditory cue when visual presentation times are short. We applied a Bayesian Observer Model to our results to reveal the computational cognitive processes that might explain audio-visual integration. Finally, measuring eye movements during this track-intercept task allows us to investigate the temporal dynamics of audio-visual integration. Eye movements in response to the moving ball are continuous—tracking the ball and predicting its trajectory—and can therefore provide a readout of the extent to which observers rely on auditory information in this visual task at different points in time.

Methods

Participants

We show data from 16 healthy adults (age: *M* = 25.5, *SD* = 4.7; eleven females; two authors). This sample size was determined using an a priori power analysis in G*Power (Faul et al., 2009; *power* = 0.80; *alpha* = 0.05) with an effect size of $\eta_p^2 = 0.34$ (main effect of stimulus volume on estimated target speed) derived from pilot data. All observers had normal or corrected-to-normal visual acuity. The study protocols were approved by the UBC Behavioural Research Ethics Board and were in line with the Declaration of Helsinki. Observers gave written informed consent and were compensated at a rate of $10 CAD/hour.

Apparatus

The experimental setup combined a high-resolution stimulus display with eye and hand tracking. Display and data collection were controlled by a PC (NVIDIA GeForce

GTX 1060 graphics card) using Matlab and the Psychophysics and Eyelink toolboxes (Cornelissen et al., 2002; Kleiner et al., 2007). Stimuli were back-projected onto a 41.8 × 33.4 cm translucent screen with a PROPixx video projector at a resolution of 1,280 × 1,024 pixels (120 Hz; VPixx Technologies, Saint-Bruno, QC, Canada). Two speakers (S-0264A; Logitech, Newark, CA, USA), located 40 cm to the left and right of the screen centre displayed the sound. Observers viewed stimuli binocularly at a distance of 44 cm while their right eye was recorded with an Eyelink 1000 Tower Mount eyetracker (1 kHz; SR Research, Kanata, ON, Canada). The 3D position of each observer's right index finger was recorded with a 3D Guidance trakSTAR (120 Hz; Ascension Technology, Shelburne, VT).

Stimuli and Procedure

In each trial, we displayed a small black disc (0.35 degrees of visual angle [°] in diameter) that moved along the trajectory of a simulated batted baseball (Fooken et al., 2016). The ball was launched at one of five speeds, resulting in five unique trajectories (**Fig. 1A**), from a start position to the left of the screen centre so that it would always move into the observer's ipsilateral reach space. The screen was separated into two zones by varying background luminance; the darker right side served as the "hit zone", in which observers were asked to intercept the ball (**Fig. 1B**). The sound of a baseball hitting a wooden bat was retrieved from a free online sound library (SocializedArtist45, 2015; 44,100 Hz). The sound was played at one of three different volumes (75, 78.5, 82 dB) at the time of ball launch and lasted approx. 50 milliseconds (ms).

Each trial began with a random-duration fixation period (**Fig. 1B**), during which observers had to fixate their eyes on a line segment that marked the ball-launch position (**Fig. 1B**) and place their hand at a start position on the table. The ball then appeared and moved for either 100 or 300 ms before disappearing from view. Observers were instructed to manually intercept the ball anywhere along its occluded trajectory within the hit zone. Upon interception, we showed a red dot, indicating the finger's interception location, and a black dot, showing actual ball position at interception (**Fig. 1B**), as visual feedback for the observer.

Before the start of the experiment, observers performed nine practice trials to familiarize themselves with setup and task. For the experiment, we used a 3 (sound volume) x 2 (presentation time) x 5 (target speed) within-subject experimental design with batting sound, presentation times, and target speed randomly selected for each trial. The experiment consisted of 420 trials total, divided into seven experimental blocks of 60 trials each (14 trials per condition) and took approximately 90 minutes to complete.

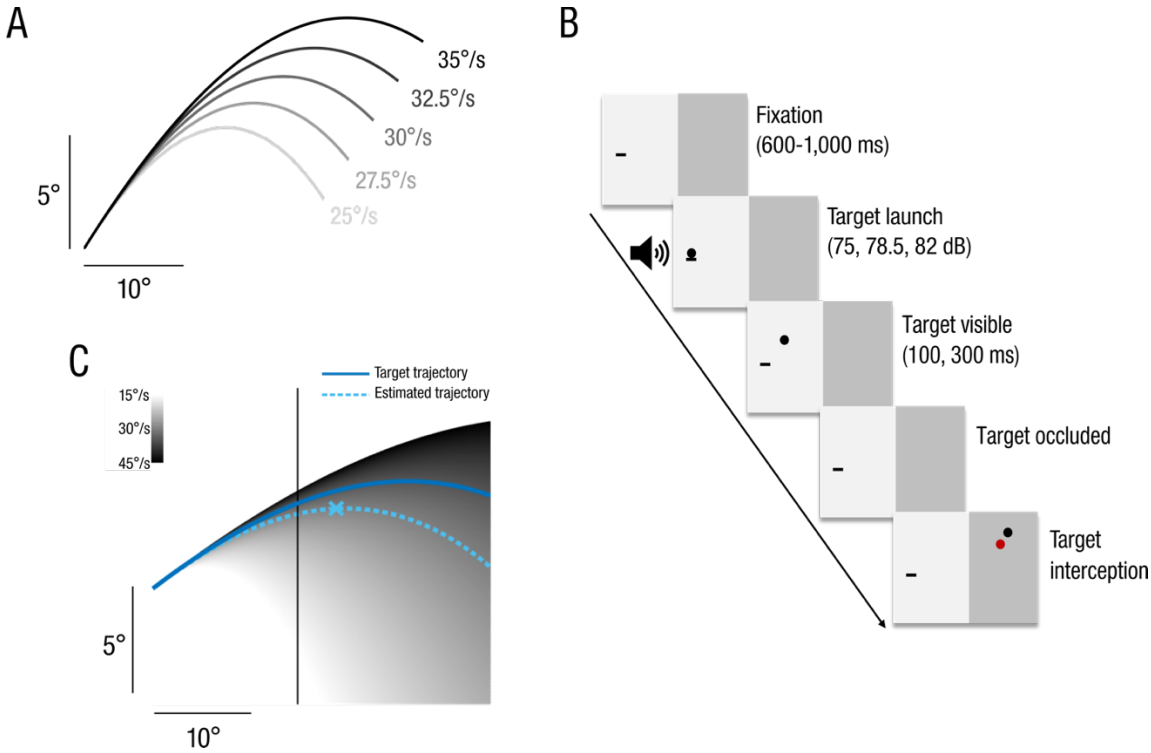

Figure 1. (A) Simulated target trajectories. **(B)** Timeline of a single trial. **(C)** Analysis of estimated target speed. The solid blue line shows the actual target trajectory in a given trial, the bright blue cross shows interception position in the same hypothetical trial. The shaded area illustrates the space comprised by simulating target speeds. The target speed that produced the smallest Euclidian distance to the interception position is labelled estimated target speed for this trial (dashed, blue line).

Eye and Hand Movement Recordings and Analyses

Eye and hand movement data were pre-processed offline using custom-made routines in MATLAB. Filtered eye movement traces (second-order Butterworth filters with 15 Hz [position] and 30 Hz [velocity] cut-off frequencies) were aligned offline to the target start position. Saccades were detected when five consecutive frames exceeded a fixed velocity criterion of 30°/s; saccade on- and offsets were then determined as the nearest reversal in the sign of acceleration.

Hand position data were up-sampled to 1 kHz by linear interpolation and filtered using a second-order Butterworth filter (15 Hz cut-off). Hand latency was computed as

the first sample exceeding 5% of the peak hand velocity in that trial. Hand movement offset was detected when the finger landed within ± 0.80 mm of the screen. If no interception was detected online, we determined the interception time and position offline as the maximum hand position in the z-dimension. We inspected all trials manually and excluded trials in which observers blinked or when the eye tracker lost the signal (3.2% of trials across participants).

Bayesian Observer Model

To reveal the interaction between visual certainty (presentation times) and the use of auditory signals (sound volume), we applied a Bayesian Observer Model (**Supplementary Materials**) integrating an observer's estimate (likelihood) of visual and auditory signals with a speed prior (Chang & Jazayeri, 2018; Jazayeri & Shadlen, 2010). First, we modelled the prior, P(S), as a continuous uniform distribution with equal probability from the lowest to the highest target speed. Second, we modelled the visual and auditory likelihood distributions, $P(S_v|S)$ and $P(S_a|S)$, around the tested speed with scalar Gaussian noise σ. We hypothesized that short stimulus presentation increases uncertainty. Thus, we expected wider likelihood distributions (larger σ) at the short presentation duration (100 ms condition), compared to the 300-ms condition, and therefore implemented two visual variances, σ_{v100} and σ_{v300}. We assume that observers will naturally associate the quiet, intermediate, and loud sounds with a slow, medium, and fast target speed, and thus modelled the auditory likelihood around the lowest (25°/s), intermediate (30°/s), and highest target speed (35°/s) with variance σ_a.

Assuming conditional independence between the visual and auditory measurements, the posterior is given by:

$$P(S|S_v,S_a) \propto P(S)P(S_v|S)P(S_a|S). \quad (1)$$

Finally, we predicted observers' estimate of target speed based on the Bayes-least-squares (BLS) estimator (i.e., the mean of the posterior) and compared the model predictions to our empirical data. The model's three free parameters (σ_{v100}, σ_{v300}, σ_a) were estimated by fitting the model to our empirical data using a residual sum square cost function and the Matlab function fminsearchbnd. Comparing Bayesian model predictions with our empirical results allows us to determine the extent to which the use of auditory signals depends on visual uncertainty.

Statistical Data Analyses

Our results are based on the analyses of three dependent variables. (1) We calculated estimated target speed from the 2D hand interception position. For each individual trial, we determined which target trajectory best fitted the observed interception position, as follows. We simulated 600 target trajectories with launch speeds ranging from 0.1°/s to 60°/s in 0.1°/s steps (**Fig. 1C**). We then determined the trajectory that produced the smallest Euclidian distance to the interception position. The corresponding target speed that best fit to the observed interception position was labelled the estimated target speed for that trial. (2) To compare effects of sound volume and presentation time between eye and hand, we repeated the same analysis for the eye position at the time of interception. (3) We analysed the amplitudes of the detected saccades during each trial to obtain a readout of audio-visual signal integration

at different time points. Because trajectories showed their strongest differences in their vertical component, we restricted this analysis on the vertical saccade amplitudes. For each dependent variable we applied a within-subject z-score outlier detection (data points were excluded if data points were >3 SD from an observer's mean).

To assess effects of stimulus volume and presentation time on our dependent variables, we calculated observers' means per condition and fed the data into repeated measures (rm) ANOVA in R (R Core Team, 2022) with an alpha level of .05. To correct for multiple comparisons within multiway ANOVA, we applied a sequential Bonferroni-correction (Cramer et al., 2016). Bonferroni-correction is applied to all post-hoc pairwise comparisons.

Results

Observers intercepted the ball with a quick pointing movement within the hit zone (**Fig. 2A**). Mean 2D interception positions strongly regress toward the intermediate trajectory in the 100-ms condition **(Fig. 2B)**, indicating poor ability to discriminate between the different target trajectories at the shortest presentation duration. In contrast, in the 300-ms condition, observers intercepted the balls more accurately along their trajectories (**Fig. 2C**).

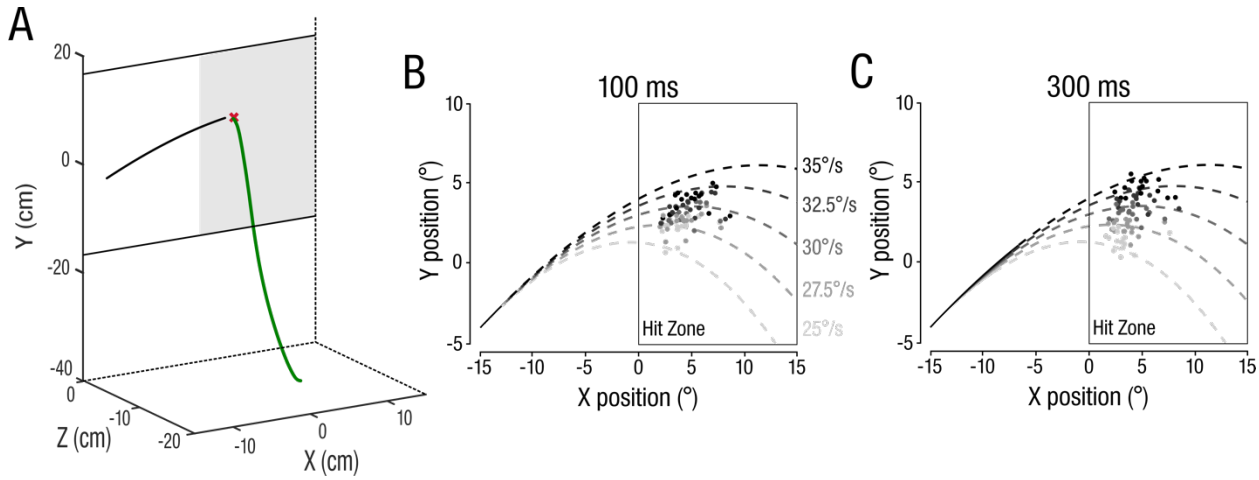

Figure 2. (A) Example of a hand (green) position trace. Black line represents the 2D target position and the grey zone represents the hit zone. Red cross indicates the interception position. **(B-C)** Mean individual observer 2D interception positions for the **(B)** 100 ms and **(C)** 300 ms presentation durations. Each data point indicates one observer's mean interception position per each of the five target speeds for *n* = 16.

Auditory Cues Affect Visual Estimates of Target Speed

In our task, observers had to rely on the visual information of the ball's speed during its brief presentation to intercept the ball along its predicted trajectory. **Figure 3** shows observers' estimated target speed as a function of physical target speed, separately for each sound volume. If speed estimates were accurate, they would fall along the diagonal (dashed line). Because sound volume was randomly assigned to target speeds and thus not informative, an ideal observer would rely solely on the visual information and ignore the auditory cue. In contrast, we found that observers systematically underestimated target speed when ball launch was paired with a quiet batting sound and overestimated target speed when the ball was paired with a loud batting sound. This effect was consistent across all target speeds at short presentation duration (**Fig. 3A**). Conversely, at long presentation duration, sound volume did not systematically affect estimated target speed (**Fig. 3B**). To assess the differential effects

of sound volume depending on presentation duration, we calculated each observer's bias in speed estimation across target speeds (mean difference between estimated and physical target speed; **Fig. 3C**). A 2 x 3 rmANOVA with factors presentation time and sound volume revealed a significant main effect of sound volume ($F(2, 30) = 4.91$, $p = .029$, $\eta_p^2 = 0.25$) and no main effect of presentation time ($F(1, 15) = 0.60$, $p = .45$, $\eta_p^2 = 0.04$). A significant interaction between presentation time and sound volume ($F(1.43, 21.47) = 20.28$, $p < .001$, $\eta_p^2 = 0.57$) confirms the profound effect of auditory cues on manual interception when visual information is sparse, but not when the target is presented sufficiently long to base speed estimation for interception on visual information alone.

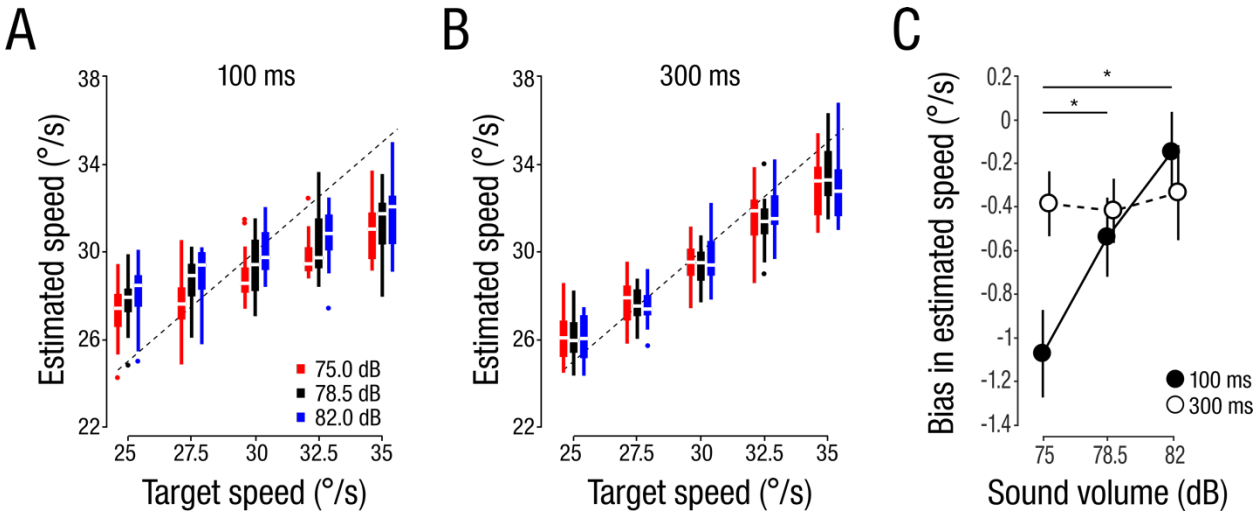

Figure 3. (**A-B**) Boxplots of estimated target speed (n = 16) as a function of physical target speed. Colours denote sound volume conditions and dashed lines indicates veridical estimates. (**B**) 100 ms condition. (**C**) 300 ms condition. (**C**) Effect of stimulus volume on the bias in the estimated target speed averaged across target speeds, separately for the 100 ms (filled circles) and 300 ms (open circles) condition. Circles and error bars denote the mean ± 1 standard error of the mean (SEM). Asterisks denote significant post-hoc comparisons ($p < .05$).

Influence of Auditory Cues Depends on Uncertainty of Visual Speed Signals

The use of auditory information in the short presentation condition is associated with strong regression toward the mean. Generally, a strong bias toward the mean indicates that an observer might have relied on an estimation strategy that follows Bayesian integration (Petzschner et al., 2015). According to this strategy, an observer in our task might have reduced uncertainty in their visual speed judgement by relying on (a) prior knowledge about the observed target speed range, and (b) the auditory cue. To further test the idea that the use of the auditory signal depended on the degree of visual uncertainty, we fitted a Bayesian Observer Model (**Fig. 4A**) to our empirical data.

The model reproduces the main finding of our data: the systematic influence of the auditory cue, and the strong regression toward the mean in the short presentation condition (**Fig. 4B,** left panel) and a weaker influence of the auditory cue and less regression toward the mean in the long presentation condition (**Fig. 4B,** right panel).

Using this model, we are now able to answer the question to what extent visual uncertainty mediates the influence of the auditory cue. Such an effect of visual uncertainty would be reflected as differences in the weight of the visual signal between the presentation conditions (i.e., different standard deviations of the visual likelihood distributions; σ_{v100} and σ_{v300}). In line with this assumption, the model produced a best estimate for σ_{v100} of 4.09, which was more than twice as large as the estimate for σ_{v300} of 1.89. Given that the variance of the auditory likelihood (σ_a = 8.26) was held constant between the presentation conditions, this finding implies that only the difference in the weight of the visual signal can explain the reliance on prior knowledge (regression toward the mean) and the influence of the auditory cue.

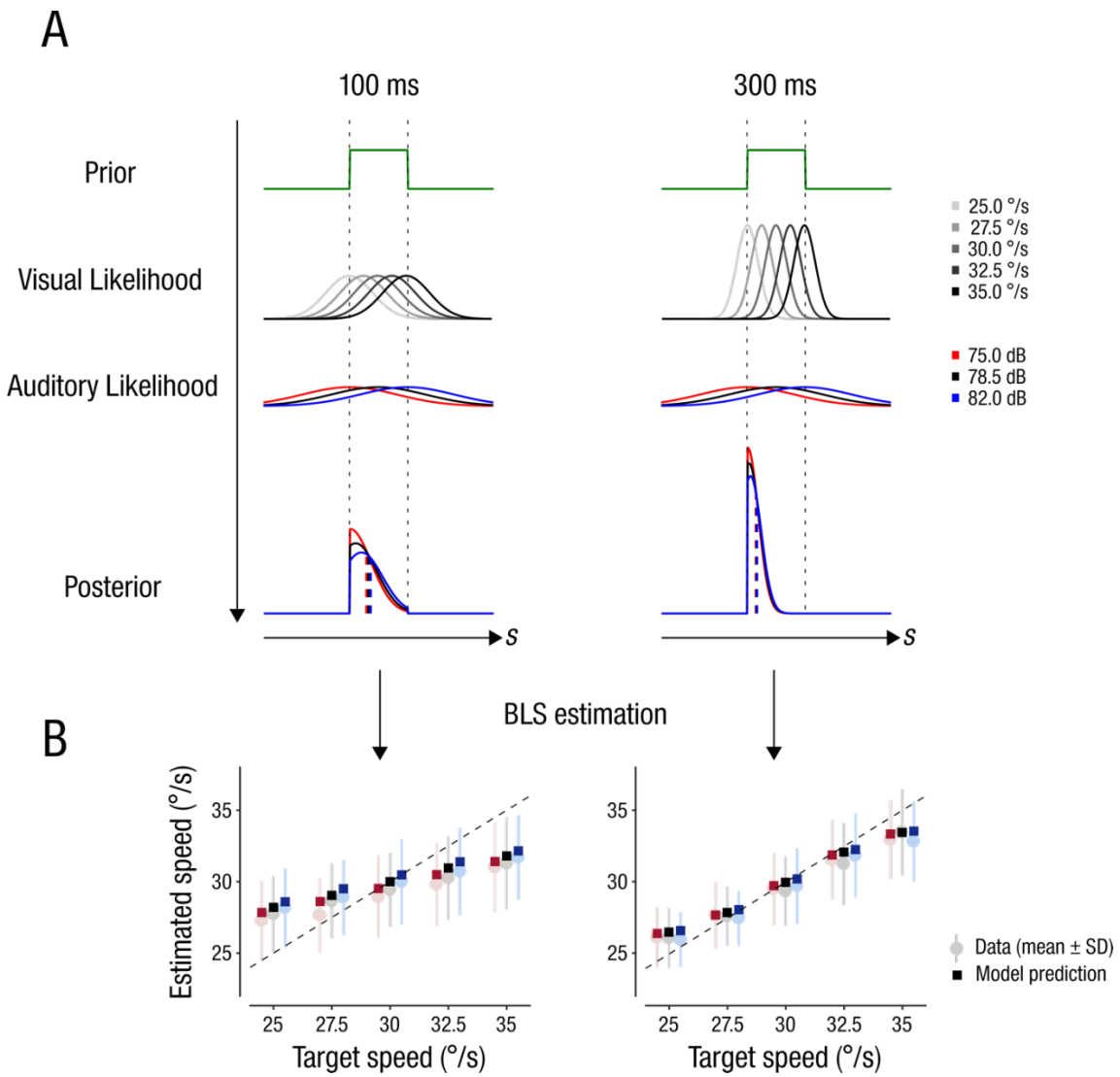

Figure 4. **(A)** Bayesian Observer Model. For better visibility, posterior distributions (fourth row) are only shown for the lowest target speed condition (25°/s). A BLS estimator (coloured dashed lines) is used to produce estimations of target speed. **(B)** Comparison between model predictions (dark squares) and empirical data (bright disks). Data points denote pooled means across observers and trials; error bars are ±1 SD.

Eye Movements Reveal Temporal Dynamics of Audio-Visual Integration

The extent to which observers relied on the auditory signal depended on the certainty of the visual speed signal. In our experiment, the degree of visual certainty was determined by the target's presentation duration (low certainty for short, high certainty for longer presentations); impact of the auditory signal decreased with

increasing duration. In this section, we analysed observers' continuous eye movements as an indicator of audio-visual integration over time.

Observers tracked the simulated baseball with their eyes using a combination of smooth pursuit and saccadic eye movements (**Fig. 5A-B**). They typically made an early catch-up saccade shortly after target onset (M = 124.19 ms, SD = 23.73 ms). Subsequent saccades were made after target disappearance to the predicted interception location.

Eye movement endpoints, based on the 2D eye position at the time of interception, reflect observers' ability to estimate target speed. **Figure 5c** indicates that observers underestimate speed in the presence of a quiet sound and overestimate it when it is paired with a louder sound, akin to what we observed for manual interception responses (**Fig. 3C**). Accordingly, a presentation time x stimulus volume rmANOVA revealed a significant main effect of stimulus volume ($F(2, 30) = 10.37$, $p = .001$, $\eta_p^2 = 0.41$), a significant stimulus volume x presentation time interaction effect ($F(2, 30) = 6.04$, $p = .012$, $\eta_p^2 = 0.29$), and no main effect of presentation time ($F(1, 15) = 1.45$, $p = .247$, $\eta_p^2 = 0.09$). Moreover, target speed estimates based on eye and hand movement endpoints were strongly correlated on a trial-by-trial basis with a median correlation of $r = 0.70$ (**Fig. 5D**). This finding indicates that the known tight link between eye and hand movements extends to multisensory stimulus environments.

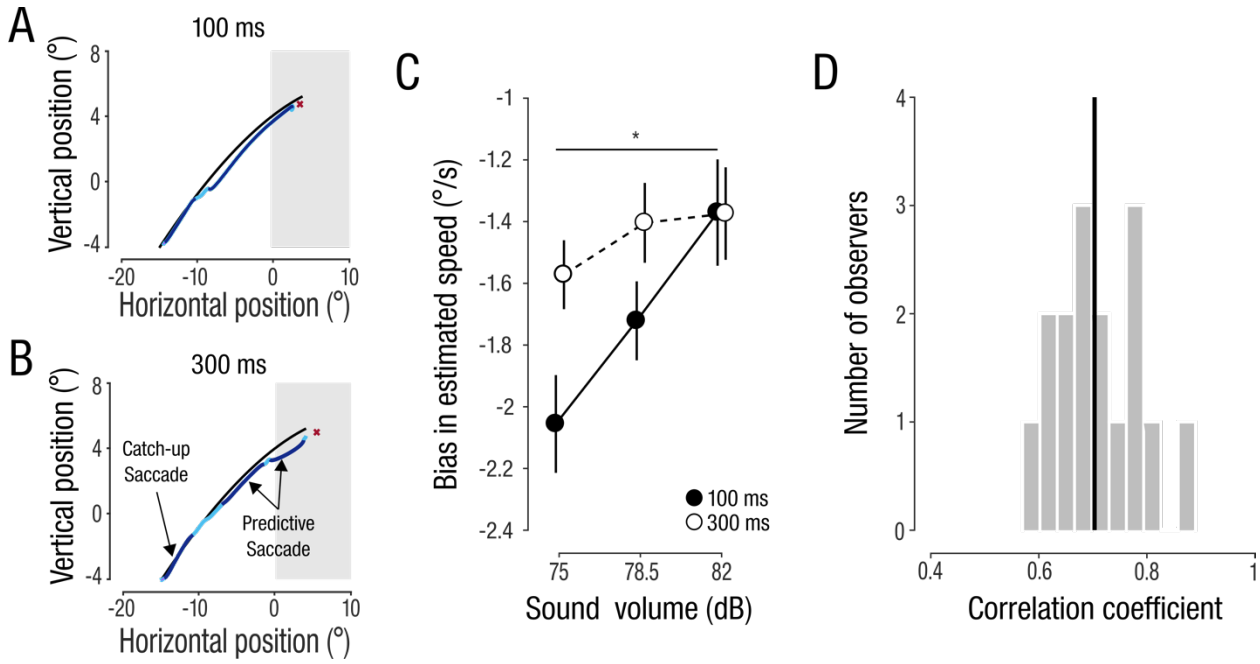

Figure 5. (**A-B**) 2D eye position traces of two individual trials for the 100ms **(A)** and 300 ms **(B)** presentation times. Bright blue segments are smooth pursuit, eyes' continuous tracking of moving targets, dark blue segments are saccades. (**C**) Effect of sound volume on estimated target speed based on final eye position. Circles indicate mean across observers; error bars are ± 1 SEM. Asterisk denotes significant post-hoc comparison ($p < .05$). (**D**) Histogram of the trial-by-trial correlation coefficients from all observers. Black line indicates the median across observers.

We next assessed whether eye movements reveal the time course of audio-visual integration by analysing the vertical amplitude of the first catch-up saccade (early integration) and the combined amplitude of subsequent, predictive saccades (later integration). For this analysis, we excluded trials where the first catch-up saccade was made in anticipation of target onset (<50 ms latency; 3.92% of trials). We found an influence of sound volume at an early timepoint (main effect of sound volume in a sound volume x presentation time rmANOVA: $F(2, 30) = 29.04$; $p = <.001$; $\eta_p^2 = 0.66$; **Fig. 6A**), consistently observed across and finely tuned to physical target speeds (**Fig. 6B**). Neither the main effect of presentation time nor the interaction term was significant (all *p*

> .487), indicating that the influence of the auditory cue occurred before any differences in presentation duration could impact visual speed judgments.

By contrast, saccades that occurred later in the trial had increased amplitudes with increasing sound volume in the 100 ms condition but had decreased amplitudes for the longer presentation duration (**Fig. 6C**). This finding was consistent across target speeds (**Fig. 6D**) and confirmed by a significant sound volume x presentation time interaction effect ($F(2, 30) = 20.85$; $p < .001$; $\eta_p^2 = 0.58$). A significant main effect of presentation time ($F(1, 15) = 9.78$; $p = .014$; $\eta_p^2 = 0.39$) is likely due to smaller saccade sizes in the 300-ms condition, which generally elicits stronger pursuit; the main effect of sound volume was not significant ($F(1.38, 20.75) = 0.60$; $p = .501$; $\eta_p^2 = 0.04$). These opposing effects in the 100- and 300-ms condition indicate a correction of the early auditory influence with the availability of additional visual information.

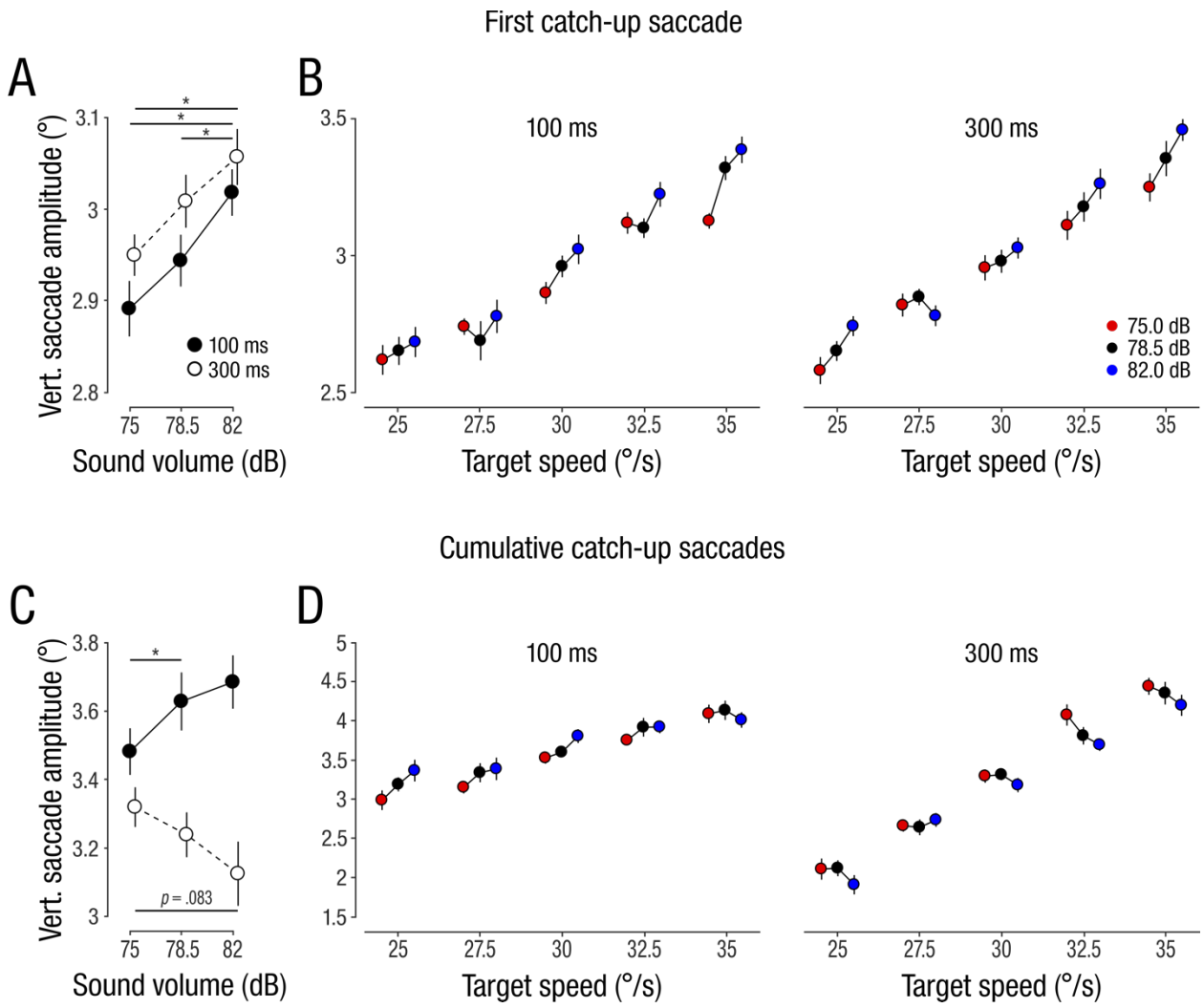

Figure 6. Saccade analyses. (**A-B**) Effect of sound volume on vertical saccade amplitudes for the first catch-up saccades after target onset, averaged across target speeds (**A**) and separately for each target speed (**B**). (**C-D**) Cumulative saccade amplitudes of all subsequent saccades. Circles and error bars show means ± 1 SEM. Asterisks denote significant post-hoc comparisons ($p < .05$).

Discussion

Predicting the trajectory of a moving object is a fundamental ability that allows us to accurately hit, catch, or otherwise intercept targets (Fiehler et al., 2019). Most research on interception has focused solely on the critical role of vision to form trajectory predictions and guide interceptive hand movements (for reviews, see Brenner &

Smeets, 2018; Fooken et al., 2021). Yet, in our natural environment, object motion is typically accompanied by sounds that can provide additional information about an object's motion. Here we show that auditory signals are used in combination with visual motion information to estimate target speed for interceptive actions. Using a rapid track-intercept task in which a visual trajectory was paired with batting sounds of varying intensities, and applying a Bayesian observer model, we present three key findings: (1) Sound intensity of bat-ball contact systematically influences interception responses. (2) The integration of auditory cues and visual information depends on the certainty of the visual signal: Auditory cues influence target speed estimates only when visual information is sparse. (3) Audio-visual integration occurred within the first 125 ms of stimulus presentation and is corrected later with the availability of additional visual information. The temporal dynamics of audio-visual integration was revealed by analyzing continuous eye movements during this task. In our task, sound intensity was never informative of physical target speed, precluding the possibility that our results were solely caused by learning to associate certain sound volumes to certain target trajectories. Together, these findings highlight the important contribution of auditory cues for vision-guided actions, particularly in situations where visual information is sparse or highly uncertain (e.g., in poor viewing conditions). These results build on a long line of literature on audio-visual signal integration for perceptual tasks (Ernst & Bülthoff, 2004). The novelty of our findings lies in discovering how auditory information contributes to vision-guided interception, a fundamental ability for everyday interactions.

Signal Uncertainty Modulates Audio-Visual Integration

By manipulating the visual presentation duration of the target, we revealed that the use of auditory cues critically depends on the uncertainty of the visual motion signal. A Bayesian Observer Model fitted to our empirical data showed that under conditions of high visual uncertainty (short presentation duration) observers relied on the auditory cue to estimate target speed. In contrast, when visual information was sufficient to accurately judge ball trajectories (long presentation), observers' interceptive hand movements were not affected by the auditory cue. Bayesian Observer Models have been applied to a variety of psychophysical data to reveal strategies of perceptual estimates under uncertainty (e.g., Körding & Wolpert, 2004) and to understand how cues from different sensory modalities are combined (Alais & Burr, 2004; Ernst & Banks, 2002; Körding et al., 2007). Under uncertainty, perceptual estimates generally indicate a strong influence of prior knowledge, which can result in a regression toward the mean (Petzschner et al., 2015). The nature of the prior depends on the sensory signal and task. When observers judge visual speed at very low contrast, they appear to rely on a "slow prior", given the statistics of the natural environment in which most objects are stationary or move slowly (Stocker & Simoncelli, 2006). Here we showed a strong regression toward the mean, indicating that observers primarily used the internal statistics of the experiment as their prior. Thus, instead of implementing a slow bias in our model, we formulized a prior based on the uniform distribution of the tested speed range (for similar approaches, see Chang & Jazayeri, 2018; Jazayeri & Shadlen, 2010). A strength of this approach is that it allowed us to reproduce our main findings without additional free parameters and assumptions about observers' speed priors.

Outside of the laboratory, Bayesian estimation models have recently been applied to real-world Major League Baseball data to elucidate the visual cues baseball batters rely on—e.g., a pitcher's posture and hand position—when estimating where to swing (Brantley & Körding, 2022). Simple cues and heuristics are critical in baseball, where hitters only have a few hundred milliseconds to decide whether and where to swing. In this or similar rapid-decision making contexts, auditory cues may provide a critical advantage, because combining them with visual cues can reduce uncertainty (Alais & Burr, 2004). Although our interception task was inspired by baseball, future studies are needed to test whether our findings and model predictions translate to how athletes integrate sounds of bat-ball contact and visual ball motion during real-world interceptive sports.

Eye Movements Elucidate Dynamics of Audio-Visual Integration

Eye movements are a natural, instinctive behavior in tasks that require fine-motor interactions with a visual object. When manually intercepting, hitting, or catching an object, observers track its trajectory until the point of interception (Fooken et al., 2021; Mrotek & Soechting, 2007). The continuous nature of these movements provides an opportunity to relate their kinematics to ongoing cognitive task processes, such as decision-making (Spering, 2022). Here we used the continuous eye movements observers made during the interception task to probe the temporal dynamics of audio-visual integration. We observed a systematic influence of the auditory cue on the first catch-up saccade, made within 125 ms of target onset. At this early time point, louder sound volumes evoked larger saccade amplitudes. However, when additional visual

information was available (long presentation duration), subsequent saccades corrected for this early auditory effect. This finding suggests that the integration of auditory and visual signals can occur at a very short timescale, in line with findings showing early effects of audio-visual cues on pupil dilation and saccade behavior in a simple saccadic decision task (Wang et al., 2017). Previous studies have identified the superior colliculus—a midbrain structure that is also involved in the control of eye movements (Sparks, 1999)—as a key hub of audio-visual integration (Stein & Stanford, 2008). Visual and auditory signals reach this brain structure within 80 ms (Ito et al., 2021), making this area an excellent candidate for short-latency audio-visual integration. Alternatively, or in parallel, visual and auditory signals could also be integrated in cortical sensory areas such as the middle temporal cortex, an area traditionally dedicated to early visual motion processing (Rezk et al., 2020).

We conclude that using non-invasive, time-sensitive eye movement measurements can provide new behavioural evidence for early integration of auditory and visual signals. This integration must occur in brain areas that receive these inputs in less than 125 ms. Auditory signals significantly and systematically impact interceptive actions in an optimal way, weighing each sensory input according to its informational value.

References

Alais D. & Burr D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14(3), P257-P262. https://doi.org/10.1016/j.cub.2004.01.029

Borghuis B.G. & Leonardo A. (2015). The Role of Motion Extrapolation in Amphibian Prey Capture. *Journal of Neuroscience*, 35(46), 15430-15441. https://doi.org/10.1523/JNEUROSCI.3189-15.2015

Brantley J.A. & Körding K.P. (2022). *Bayesball: Bayesian integration in professional baseball batters*. bioRxiv. https://doi.org/10.1101/2022.10.12.511934

Brenner E. & Smeets J.B.J. (2018). Continuously updating one's predictions underlies successful interception. *Journal of Neurophysiology*, 120(6), 3257-3274. https://doi.org/10.1152/jn.00517.2018

Cañal-Bruland R., Meyerhoff H.S., & Müller F. (2022). Context modulates the impact of auditory information on visual anticipation. *Cognitive Research: Principles & Implications*, 7, 76. https://doi.org/10.1186/s41235-022-00425-2

Cañal-Bruland R., Müller F., Lach B., & Spence C. (2018). Auditory contributions to visual anticipation in tennis. *Psychology of Sport and Exercise,* 36, 100–103. https://doi.org/10.1016/j.psychsport.2018.02.001

Chang C.J. & Jazayeri M. (2018). Integration of speed and time for estimating time to contact. *Proceedings of the National Academy of Sciences USA*, 115(12), E2879-E2887. https://doi.org/10.1073/pnas.1713316115

Chaplin T.A., Rosa M.G.P., & Lui L.L. (2016). Auditory and visual motion processing and integration in the primate cerebral cortex. *Frontiers in Neural Circuits*, 12, 23. https://doi.org/10.3389/fncir.2018.00093

Cornelissen F.W., Peters E.M., & Palmer, J. (2002). The Eyelink Toolbox: eye tracking with MATLAB and the Psychophysics Toolbox. *Behavior Research Methods, Instruments, & Computers,* 34, 613–617. https://doi.org/10.3758/BF03195489

Cramer A.O.J., van Ravenzwaaij D., Matzke D., Steingroever J., Wetzels R., Grasman R.P.P.P., Waldorp L.J., Wagenmakers J.E. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, 23(2), 640-647. https://doi.org/10.3758/s13423-015-0913-5

Diaz G.J., Cooper J., Rothkopf C., & Hayhoe M. (2013). Saccades to future ball location reveal memory-based prediction in a virtual-reality interception task. *Journal of Vision*, 13(1), 1-14. https://doi.org/10.1167/13.1.20

Ernst M. & Banks M. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature,* 415(6870), 429–433. https://doi.org/10.1038/415429a

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

Faul F., Erdfelder E., Lang A.G., & Buchner A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavioral Research Methods*, 39(2), 175-191. https://doi.org/10.3758/bf03193146

Fiehler K., Brenner E., & Spering M. (2019). Prediction in goal-directed action. *Journal of Vision*, 19(9), 1-21. https://doi.org/10.1167/19.9.10

Fooken J., Kreyenmeier P., & Spering M. (2021). The role of eye movements in manual interception: a mini-review. *Vision Research*, 183, 81–90. https://doi.org/10.1016/j.visres.2021.02.007

Fooken J., Yeo S.H., Pai D.K., & Spering M. (2016). Eye movement accuracy determines natural interception strategies. *Journal of Vision*, 16(14), 1–15. https://doi.org/10.1167/16.14.1

International Blind Sports Federation (2022, February 4). *Goalball Rules and Regulations 2022-2024 (v1.1)*. https://goalball.sport/wp-content/uploads/2022/04/IBSA-Goalball-Rules-and-Regulations-2022-2024-v1.1-4-Feb-22.docx-Summary-Change-document.pdf

Ito S., Si Y., Litke A.M. & Feldheim D.A. (2021). Nonlinear visuoauditory integration in the mouse superior colliculus. *PLoS Computational Biology*, 17(11), e1009181. https://.doi.org/10.1371/journal.pcbi.1009181

Jazayeri M. & Shadlen M.N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*,13(8), 1020-1026. https://doi.org/10.1038/nn.2590

Kleiner M., Brainard D., Pelli D., Ingling A., Murray R., & Broussard C. (2007). What's new in psychtoolbox-3. *Perception*, 36(14), 1-16.

Körding K.P., Beierholm U., Ma W.J., Quartz S., Tenenbaum J.R., & Shams L. (2007). Causal inference in multisensory perception. *PLoS ONE*, 2(9), e943. https://doi.org/10.1371/journal.pone.0000943

Körding K.P. & Wolpert D.M. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427(6971), 244-247. https://doi.org/10.1038/nature02169

Metzger W. (1934). Beobachtungen über phänomenale Identität. *Psychologische Forschung,* 19*,* 1-60. https://doi.org/10.1007/BF02409733

Meyerhoff H.S., Gehrer N.A., Merz S., & Frings C. (2022). The beep-speed illusion: Non-spatial tones increase perceived speed of visual objects in a forced-choice paradigm. *Cognition*, 219, 104978. https://doi.org/10.1016/j.cognition.2021.104978

Michaiel A.M., Abe E.T.T., & Niell C.M. (2020). Dynamics of gaze control during prey capture in freely moving mice. *eLife,* 9, e57458. https://doi.org/10.7554/eLife.57458

Mrotek L.A. & Soechting J.F. (2007). Target interception: hand-eye coordination and strategies. *Journal of Neuroscience,* 27(27), 7297-309. https://doi.org/10.1523/JNEUROSCI.2046-07.2007

Petzschner F.H., Glasauer S., Stephan K.E. (2015). A Bayesian perspective on magnitude estimation. *Trends in Cognitive Science*,19(5), 285-293. https://doi.org/10.1016/j.tics.2015.03.002

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org

Rezk M., Cattoir S., Battal C., Occelli V., Mattioni S. & Collignon O. (2020). Shared Representation of Visual and Auditory Motion Directions in the Human Middle-Temporal Cortex. *Current Biology*, 30(12), 2289-2299. https://doi.org/10.1016/j.cub.2020.04.039

Sekuler R., Sekuler A., & Lau, R. (1997). Sound alters visual motion perception. *Nature,* 385(6614), 308. https://doi.org/10.1038/385308a0

SocializedArtist45 (2015). *Baseball Hitting Bat*. Freesound.org. https://freesound.org/people/SocializedArtist45/sounds/266595/

Soto-Faraco S., Kingstone A., & Spence C. (2003). Multisensory contributions to the perception of motion. *Neuropsychologia*, 41(13), 1847-1862. https://doi.org/10.1016/s0028-3932(03)00185-4

Sparks D.L. (1999). Conceptual issues related to the role of the superior colliculus in the control of gaze. *Current Opinion in Neurobiology*, 9(6), 698-707. https://doi.org/10.1016/s0959-4388(99)00039-2

Spering M. (2022). Eye Movements as a window into decision-making. *Annual Review of Vision Science*, 8, 427-448. https://doi.org/10.1146/annurev-vision-100720-125029

Spering M., Schütz A.C., Braun D.I., & Gegenfurtner K.R. (2011). Keep your eyes on the ball: smooth pursuit eye movements enhance prediction of visual motion. *Journal of Neurophysiology*, 105(4), 1756–1767. https://doi.org/10.1152/jn.00344.2010

Stein B.E. & Stanford T.R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, 9(4), 255-266. https://doi.org/10.1038/nrn2331

Stocker A.A. & Simoncelli E.P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9(4), 578-585. https://doi.org/10.1038/nn1669

Wang C.A., Blohm G., Huang J., Boehnke S.E., & Munoz D.P. (2017). Multisensory integration in orienting behavior: Pupil size, microsaccades, and saccades. *Biological Psychology*, 129, 36-44. https://doi.org/10.1016/j.biopsycho.2017.07.024

Wessels M., Zähme C., & Oberfeld D. (2022). Auditory information improves time-to-collision estimation for accelerating vehicles. *Current Psychology*. https://doi.org/10.1007/s12144-022-03375-6

On the mechanisms of audio-visual integration for manual interception

Supplementary Materials

Philipp Kreyenmeier[1,2*†], Anna Schroeger[3,4*], Rouwen Cañal-Bruland[4], Markus Raab[5,6], & Miriam Spering[2,7,8]

Bayesian Observer Model

We built a Bayesian Observer Model that integrates the measurements of the visual and auditory signals with a prior based on the statistics used in the experiment. For simplicity, we modelled the prior as a continuous uniform distribution from the lowest (s^{min}) to the highest (s^{max}) tested target speed (s):

$$(1) \quad p(s) = \begin{cases} \frac{1}{(s^{max} - s^{min})} & \text{for } s^{min} \geq s \leq s^{max} \\ 0 & \text{otherwise} \end{cases}.$$

Visual and auditory likelihood distributions were modelled around the tested stimulus *s* and with scalar Gaussian noise σ. We expected the likelihood distribution to be more variable in the 100 ms conditions, compared to the 300 ms condition and implemented two separate variances, σ_{v100} and σ_{v300}, resulting in two separate visual likelihood distributions:

$$(2) \quad p_{v100}(s_v|s) = \frac{1}{\sqrt{2\pi \cdot \sigma_{v100}{}^2}} \times e^{\frac{-(s_v - s)^2}{2 \cdot \sigma_{v100}{}^2}},$$

$$(3) \quad p_{v300}(s_v|s) = \frac{1}{\sqrt{2\pi \cdot \sigma_{v300}{}^2}} \times e^{\frac{-(s_v - s)^2}{2 \cdot \sigma_{v300}{}^2}}.$$

The auditory likelihood distribution was modelled with variance σ_a:

$$(4) \quad p_a(s_a|s) = \frac{1}{\sqrt{2\pi \cdot \sigma_a{}^2}} \times e^{\frac{-(s_a - s)^2}{2 \cdot \sigma_a{}^2}}.$$

Because sound volume did not directly relate to a physical target speed, we assumed that the quiet, intermediate, and loud sounds corresponded to the lowest (25°/s), intermediate (30°/s), and highest target speed (35°/s). Assuming conditional independence between the visual and auditory measurements, the posterior is given by:

$$(5) \quad p(s|s_v, s_a) = \frac{p(s)p(s_v|s)p(s_a|s)}{\int p(s)p(s_v|s)p(s_a|s)ds}.$$

Finally, we predicted observers' estimate of target speed based on the Bayes-least-squares estimator (BLS; i.e., the mean of the posterior) and compared the model predictions to our empirical data. Our model contained three free parameters, the variances of the likelihood distributions (σ_{v100}, σ_{v300}, and σ_a), that were estimated by fitting the model to the observed means per condition using a residual sum square cost function and the Matlab function fminsearchbnd. Because performance in fast-paced interception tasks with disappearing targets is typically largely variable across observers, we only used the means calculated across all observers and trials per condition to fit our model (**Fig. 4B**).