A Model of Audio-Visual Motion Integration During Active Self-Movement

Maria Gallagher[1], Joshua D. Haynes[2,3], John F. Culling[2] & Tom C. A. Freeman[2]

[1]School of Psychology, University of Kent, Canterbury, UK

[2]School of Psychology, Cardiff University, Cardiff, UK

[3]School of Health Sciences, University of Manchester, Manchester, UK

Corresponding authors:
Maria Gallagher and Tom C. A. Freeman

Abstract

Despite good evidence for optimal audio-visual integration in stationary observers, few studies have considered the impact of self-movement on this process. When the head and/or eyes move, the integration of vision and hearing is complicated as the sensory measurements begin in different coordinate frames. To successfully integrate these signals, they must first be transformed into the same coordinate frame. We propose that audio and visual motion cues are separately transformed using self-movement signals, before being integrated as body-centred cues to audio-visual motion. We tested this hypothesis using a psychophysical audio-visual integration task in which participants made left/right judgements of audio, visual, or audio-visual targets during self-generated yaw head rotations. Estimates of precision and bias from the audio and visual conditions were used to predict performance in the audio-visual conditions. We found that audio-visual performances were predicted well by models which suggested the transformation of cues into common coordinates, but could not be explained by a model that did not rely on coordinate transformation before integration. We also found that precisions specifically were better predicted by a model that accounted for shared noise arising from signals encoding head movement. Taken together, our findings suggest that motion perception in active observers is based on the integration of partially-correlated body-centred signals.

Keywords: Multisensory Integration, Motion Perception, Self-Movement, Active Movement, Audio-Visual Integration

Despite good evidence for optimal audio-visual integration of spatial cues to direction and movement in stationary observers, few studies have considered the impact of self-movement on this process. Early auditory and visual signals are represented in different spatial coordinate frames, with auditory cues starting out in head-centred coordinates and visual signals in eye-centred coordinates. The coordinate frames align in observers who keep their eyes and head stationary, which makes the integration of audio-visual spatial cues relatively straightforward. Observers appear to use an optimal integration strategy, based on the maximum likelihood principles original developed to explain the integration of depth cues both within and across modalities (Ernst & Banks, 2002; Ernst & Bülthoff, 2004; Landy et al., 1995). Accordingly, more reliable estimates are given a higher weighting than less reliable ones, resulting in a combined percept that is more precise than either cue alone. The maximum likelihood principle explains why the spatial localisation of stationary audio-visual targets is dominated by the visual location when visual reliability is high, but gradually shifts towards the auditory location as visual reliability decreases (Alais & Burr, 2004b; Bolognini et al., 2007). It also explains why localisation is more precise for audio-visual targets presented in isolation compared to auditory or visual stimuli alone (Alais & Burr, 2004b; Hairston et al., 2003; Wuerger et al., 2010). The optimal integration exhibited by observers is not limited to stationary targets, with the perceived direction of moving auditory and visual targets shifting towards the direction of the more reliable sense (Meyer & Wuerger, 2001; Soto-Faraco et al., 2002, 2004).

The integration of audio-visual cues is made more complicated when the observer moves their head and/or eyes because self-movement impacts vision and hearing in different ways. For example, with the head stationary, moving the eyes across a stationary audio-visual object results in motion across the retina, but none across the ears. In this case, the coordinate frames in which visual and auditory information reside no longer coincide. As such, additional computational steps are required, firstly to ‘compensate’ for movement of the eyes, and secondly to transform both modalities into a common coordinate frame. Only then is

integration possible (Andersen et al., 1993; Burns & Blohm, 2010; Harris, 1994; Landy et al., 1995; Sober & Sabes, 2003).

Previous research suggests that both auditory and visual signals can be transformed into eye-, head-, and/or body-centred coordinate frames, depending on the task and stimuli used (Cohen & Andersen, 2002; Collins et al., 2010; Furman & Gur, 2012; Hairston et al., 2003; Kopinska & Harris, 2003; Wallach, 1987; Wertheim, 1994). During smooth pursuit eye-movements, information from eye-muscle proprioception and efference copy is integrated with retinal information to distinguish retinal motion resulting from self-movement and object movement (Freeman & Banks, 1998; Furman & Gur, 2012; Sperry, 1950; von Holst & Mittelstaedt, 1950). It is likely that the compensation for eye-movements occurs in the medial-superior temporal (MST) cortical region, which receives extra-retinal input from proprioception, the vestibular system, and efference copy, and represents visual signals in both eye- and head-centred reference frames (Furman & Gur, 2012; Gu et al., 2008; Newsome et al., 1988; Ono & Mustari, 2012). Less attention has been given to how the auditory system compensates for head movements, despite evidence that individuals are able to localise auditory stimuli during head and eye-movements (Genzel et al., 2016; Goossens & van Opstal, 1999). It is possible that head-movement compensation occurs via similar processes to eye-movement compensation, requiring the integration of auditory spatial information with “extra-cochlear” signals from proprioception, the vestibular system, and efference copy (Freeman et al., 2017). Previous research has demonstrated that such signals can influence the localisation of auditory cues (Lewald et al., 1999; Lewald & Karnath, 2000), with evidence suggesting that auditory signals are transformed from head- to body-centred coordinates (Genzel et al., 2016; Goossens & van Opstal, 1999; Vliegen et al., 2004).

Our hypothesis is that audio-visual motion integration during self-movement is based on the optimal combination of compensated auditory and visual cues, as shown in Figure 1. According to this hypothesis, auditory and visual cues are first transformed into a common coordinate frame, based on self-movement compensation mechanisms described above, and

then integrated into a coherent audio-visual percept using optimal integration of the compensated cues. We make the assumption that the cues are transformed into a body-centred reference frame partly because this reference frame remains stationary during head and eye movements in our experiments, and also previous literature shows that both auditory and visual cues are transformed into this frame (Furman & Gur, 2012; Goossens & van Opstal, 1999; Kopinska & Harris, 2003; Lewald et al., 1999). For example, Kopinska and Harris (2003) demonstrated that both auditory and visual localisation were impacted by head-on-body position, but were not affected by eye-in-head or body-in-space positions, implying that such localisation judgements are based on body-centred coordinates. This is echoed in gaze-orienting behaviour towards sequences of auditory and visual targets (Goossens & van Opstal, 1999). Crucially, as the figure shows, the transformed cues are partially correlated because they share a source of noise that depends on the precision of the signals encoding head movement (the body-centred cues may be also biased, a point taken up below). To account for this shared noise, we used a modified version of the standard optimal cue combination model proposed by Oruç et al. (2003). Instead of including the shared noise as a free parameter, however, we used a recently-developed technique for measuring the precision of this self-movement signal, one specifically designed to measure signal noise when self-movement is self-controlled as in the current paper (Haynes et al., 2024).

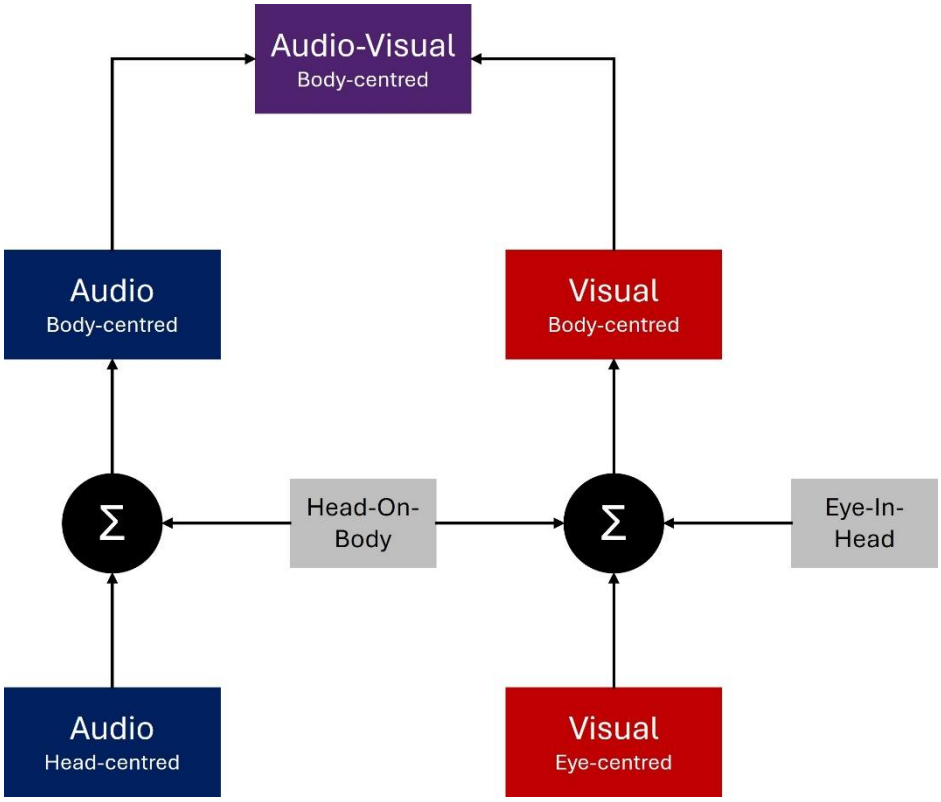

Figure 1. *Outline of the proposed framework for integrating of audio-visual cues during self-movement. The framework is based on the integration of compensated cues. For hearing, compensation is carried out using head-on-body signals to transform audio-motion signals from head-centred coordinates to body-centred coordinates. For vision, both head-on-body and additional eye-in-head signals are used to transform visual motion signals from eye-centred coordinates to body-centred coordinates. These 'compensated' audio and visual signals are then integrated, resulting in a body-centred audio-visual estimate of motion.*

We tested the framework sketched in Figure 1 using a psychophysical audio-visual integration task in which participants freely rotated their heads around a vertical axis while maintaining their gaze on a head-stationary fixation point. Although an unusual fixation strategy, this type of gaze behaviour guarantees that vision must transform retinal image motion into body-centred coordinates in order to make the correct spatial judgment, in the same way that hearing must transform auditory images. Participants were asked to judge the direction of motion of audio, visual, and audio-visual targets in interleaved conditions. In some conditions, we also added external noise to the stimuli to manipulate the reliability (i.e. precision) of the cues directly (Bentvelzen et al., 2009; Ernst & Banks, 2002; Girshick et al., 2011). We fit the data with three models: a standard cue integration model (Ernst & Banks,

2002; Ernst & Bülthoff, 2004) based on transformed auditory and visual signals, but which does not account for shared noise between the cues (Body-Centred Integration); a modified optimal integration model (Oruç et al., 2003), based on transformed auditory and visual signals, which also accounts for shared noise (Body-Centred Integration, adjusted for shared noise); and finally an optimal integration model based on uncompensated cues i.e. auditory and visual image motion alone (Image-Centred Integration). In anticipation of the results, we found that the data was accounted for well by models based on transformed auditory and visual signals, but could not be explained by a model based on uncompensated cues i.e., auditory and visual image motion alone.

Methods

Participants

Six participants (three female, mean age = 36.83, SD = 14.74) completed the experiment. Two participants were naïve to the purposes of the study, four participants were the study authors. No participants reported neurological or psychiatric conditions. All participants were right-handed.

Ethics

The study procedure was approved by the Cardiff University ethics committee (EC.12.04.03.3123GRA2), and was conducted in accordance with the Declaration of Helsinki. All participants provided informed consent prior to commencing the study.

Stimuli and Equipment

A photograph of the experimental setup can be seen Haynes et al. (2024). The experiment took place in a carpeted room, sound treated with wall and ceiling tiles with absorption coefficients of 0.9, resulting in a reverberation time of ~60 ms. Auditory stimuli were presented via a 2.4-m diameter ring of 48 speakers (Cambridge Audio Minx), controlled by two MoTU 24-channel sound cards, each linked to four six-channel Auna amplifiers. Sound intensity was normalised across all speakers with a RadioShack 33-2055 digital meter on the dB A setting using 0.2-20 kHz noise. Auditory stimuli were spatially updated at 240 Hz. Visual stimuli were presented via an Adafruit Neopixel strip of 342 RGB LEDs subtending 256°, controlled by an Arduino Uno microcontroller. The LED strip was covered with a 1.2f neutral density filter and three layers of diffuser gel at a distance of 35 mm. Visual stimuli were updated at 40 Hz. The distance from the participant's head to the LED/Speaker array was 1.2 meters. Head movements were recorded via a Liberty Polhemus tracker, sampled at 240 Hz. Changes in head movement direction were detected using a smoothed derivative of head position, which was achieved by convolving tracker samples with a finite difference filter (13 samples long). This meant that head turns were detected 7 frames (~30 ms) after the turn was made (Figure 2). Eye movements were recorded via a PupilLabs Pupil Core eye tracker sampled at

30 Hz (Participants 1, 2, and first four sessions of Participant 4) or 120 Hz (Participants 5 and 6, remainder of Participant 4). The front-facing camera was used to calibrate eye position using a 3x2 array of calibration points, allowing conversion of normalised units to degrees. Stimulus presentation and response collection was conducted via custom MATLAB r2018b scripts.

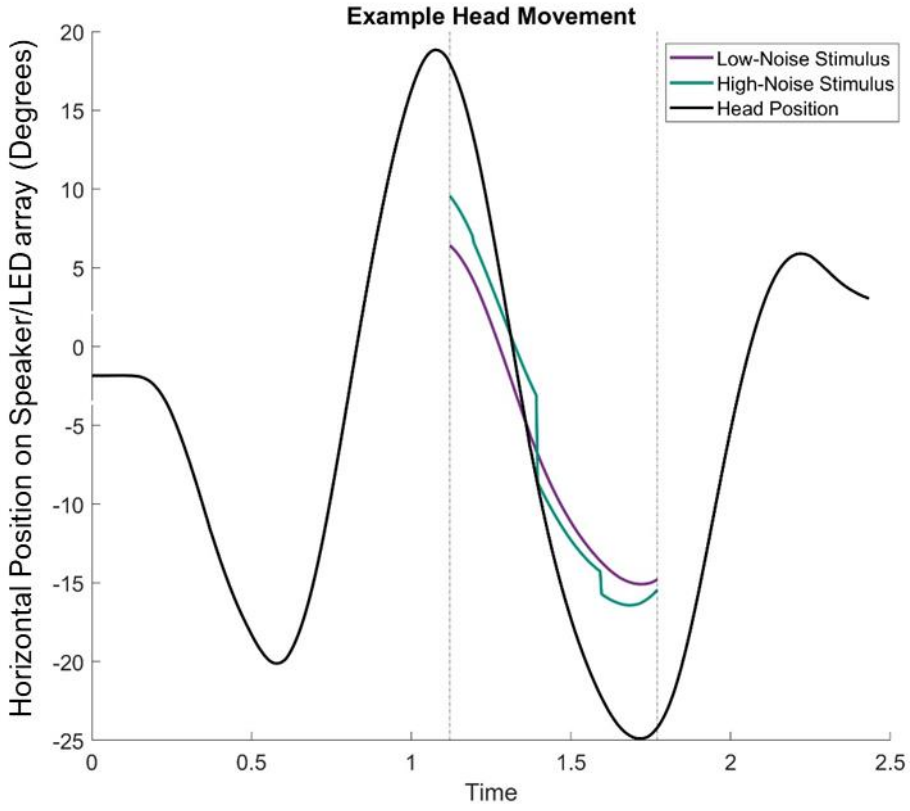

Figure 2. *Example head movement and stimulus position over time. Vertical lines indicate the start and end of the third head sweep, during which time the stimulus to be judged was visible. The purple line indicates a low-noise stimulus, while the green line indicates a high-noise (i.e., jittered) stimulus.*

A visual cue was used to indicate the beginning of each trial. The cue was a diffuse blue LED blob, spatially windowed by a Gaussian distribution with σ = 1.05°. This light lasted for 500 ms, or until the participant moved the head (whichever was soonest).

The fixation point was a diffuse green LED blob windowed by a Gaussian with σ = 1.05°, increased to σ = 1.07° by the diffuser, with a peak luminance of ~0.042 cd/m^2. The

fixation point was yoked to the participants' head movement, such that the eyes were always straight forward in the head (i.e., the eyes moved with respect to the speaker/LED ring).

The visual stimulus to be judged was a diffuse red LED blob based windowed by a Gaussian with σ = 2.25° The auditory stimulus was a white noise burst spatially windowed by a gaussian function with an SD of 5.25° in power (σ = 7.5° amplitude). The noise was sampled at 48 kHz, with a peak of 70 dB. In Audio-Visual conditions, auditory and visual stimuli were presented at the same location with the same speed.

The Method of Constant Stimuli was used to measure precision and accuracy. Audio and visual stimuli moved at a proportion of the head speed (movement gain), in the same or opposite direction as the head. Movement gains ranged from 0 to ± 0.5 of the proportion of head speed in 7 steps (13 speeds total). Each speed was presented 30 times. Head movement speeds were entirely paced by the participants. However, no fixation or stimuli would appear if the participants' head speed fell below a threshold (15°/s), as determined by a leaky integrator. This integrator made the amplitude/luminance of the stimulus decrease by 50% for each frame spent below the threshold speed.

The precision of the stimuli could be modified by adding positional jitter (Bentvelzen et al., 2009) randomly drawn from a rectangular distribution that was ±7.5° wide. The jitter was updated at 5 Hz, and was added to each modality to create 'low-noise' and 'high-noise' conditions (Figure 2, Table 1).

Table 1.
Conditions used in the experiment.

Modality	Positional Jitter
Visual	None Visual
Audio	None Audio
Audio-Visual	None Visual Audio Audio & Visual

Procedure

Each trial began with a cue light to indicate that the participants should position their head looking straight ahead and aligned with the body. Once the participant was in the correct position, the cue light disappeared, and participants were instructed to begin moving the head. Participants made self-paced back-and-forth (yaw) head movements while maintaining fixation on the green fixation light. On the third head 'sweep', a stimulus to be judged was presented, which moved in the same or opposite direction as the head, at a speed defined by the movement gain selected for that trial. The stimulus could be audio, visual, or audio-visual, and could be presented at high or low reliabilities (with low reliability based on the addition of positional jitter). Thus all the conditions defined by Table 1 were interleaved in any one session. The participant judged whether the stimulus moved to the left or right using a key press. Once the press was recorded, the next trial began once the participant's head was approximately centred (within ±7.5° of straight ahead).

Eight conditions were tested (Table 1), with all trial types interleaved. In total, the experiment consisted of 3,120 trials per participant (8 conditions × 13 target speeds × 30 target repetitions), split into approximately 24 10-minute blocks over 6-8 sessions on separate days.

Psychophysical Analysis

Analyses were carried out using MATLAB 2022a (MATLAB Version 9.12.0.1884302 (R2022a), 2022) and the Palamedes Toolbox (Prins & Kingdom, 2018) (psychometric function fits, model estimations, and confidence intervals) and R (version 4.3.2, (R Core Team, 2022) and RStudio (version 2023.12.0, (RStudio Team, 2019), with packages ez (Lawrence, 2016) and dplyr (Wickham et al., 2023) (repeated measures ANOVAs, post-hoc tests). The proportion of 'With the Head' responses were calculated for each gain value in each condition. Psychometric functions based on cumulative Gaussians were fit to the data using the PAL_PFML_Fit function from the Palamedes Toolbox (Prins & Kingdom, 2018). Biases were estimated at the point of 50% "With the Head" responses. Precisions were estimated as the inverse of the slope of the psychometric function. This corresponds to the standard deviation

of cumulative Gaussian fit to the data. Accordingly, higher standard deviations indicate lower precision. Lapse rates were a free parameter, constrained to a maximum of 0.02 (Prins, 2012). Example Psychometric function fits from a naïve subject in the Visual, Audio, and Audio-Visual No Jitter conditions can be seen in Figure 3.

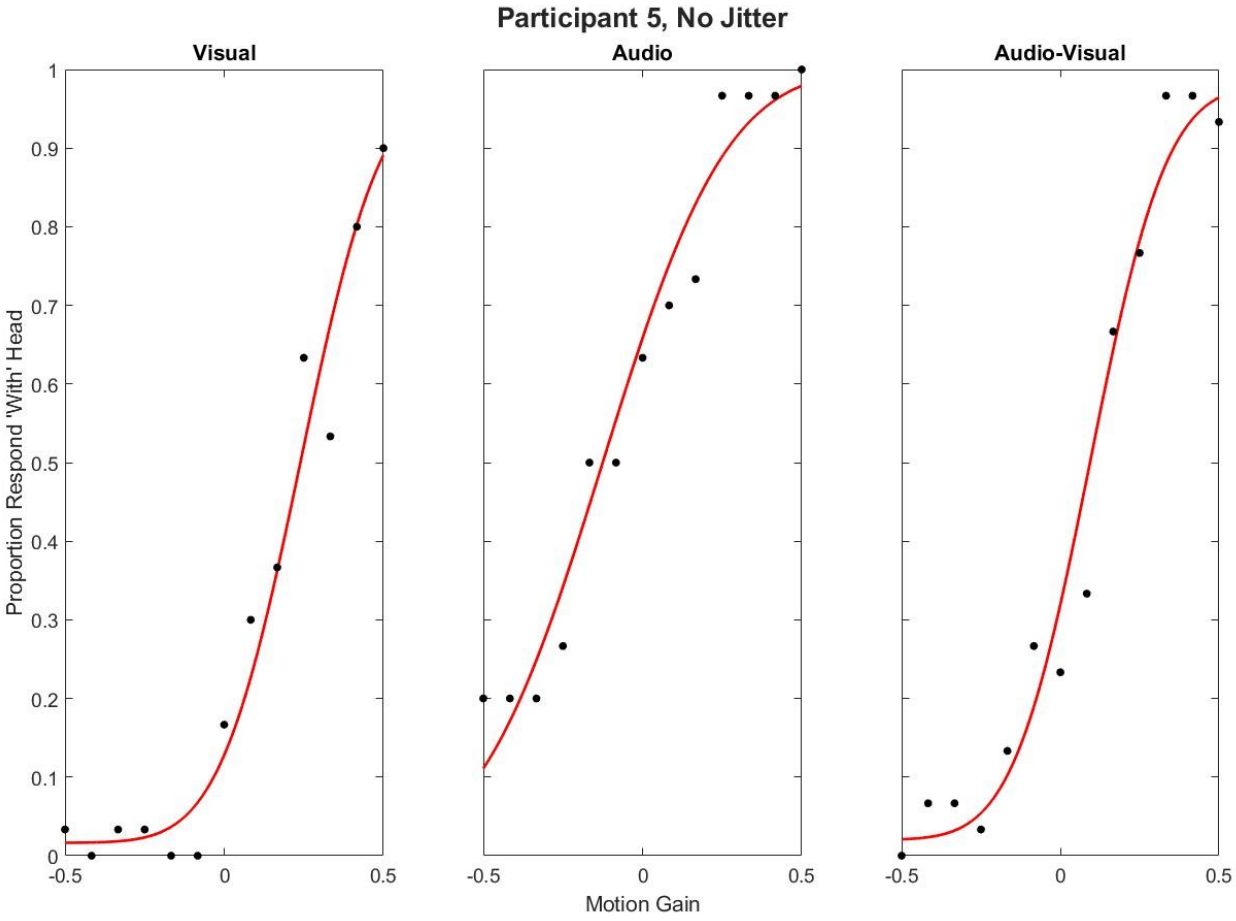

Figure 3. Example psychometric function fit from a naïve participant.

Table 2
Audio-Visual Conditions by Predicting Conditions

Audio-Visual Condition	Predicting Conditions
No Jitter	Visual – No positional jitter Audio – No positional jitter
Visual Jitter	Visual – Positional jitter Audio – No positional jitter
Audio Jitter	Visual – No positional jitter Audio – Positional jitter

Audio-Visual Jitter	Visual – Positional jitter Audio – Positional jitter

Models

Data in each of the four audio-visual conditions (No Jitter, Visual Jitter, Audio Jitter, Audio-Visual Jitter) were fit with three alternative models based on parameters obtained in the Audio and Visual conditions (see Table 2 for how the unimodal prediction conditions were coupled to appropriate Audio-Visual conditions). Model 1 was based on optimal integration of body-centred cues but ignoring the shared-noise defined by the self-movement signal common to both compensated cues. Model 2 included this shared noise and Model 3 was based on more standard optimal integration from auditory and visual image motion alone (i.e. self-movement signals ignored). The parameters for these models were fixed, including the shared noise in Model 2, which was measured in a separate experiment on the same participants (Appendix A). The only free parameters were those relating to the initial fitting of the psychometric functions to the raw data to obtain relevant PSEs and slope parameters needed to calculate the model predictions. Hence, there was no need to correct for the number of free parameters when comparing models because each had the same. To evaluate which model best predicted precisions and biases on a group level, we calculated the squared differences of the model predictions to the observed values. We then compared these squared differences using repeated measures ANOVAs, with Model and Condition as factors. Significant main effects and/or interactions were followed up with post-hoc Bonferroni-corrected pairwise t-tests where necessary. We also compared the difference between each prediction and measured parameter on an individual level using 95% confidence intervals constructed from 2,000 bootstrapped samples. Confidence intervals that did not contain 0 indicated that the prediction significantly differed from the actual parameter on an individual basis. Moreover, in a second analysis based on the same data using techniques in Haynes et al. (2024), we also accounted for the head-movement variability, which potentially affects the interpretation of the slopes of the psychometric functions in ways described below. Importantly,

this second analysis tests the models at the level of the psychometric function, rather than on the basis of psychometric function parameters from the single cue conditions. The root mean square error (RMSE) of each model fit to the data was calculated, and the differences in RMSE across Model and Condition were analysed using repeated measures ANOVAs as above. Thus, this second analysis provided further converging evidence to support our conclusions.

Model 1: Body-Centred Integration (BCI)

We used a standard cue integration approach (Ernst & Banks, 2002; Ernst & Bülthoff, 2004) to predict the audio-visual bias ($\hat{S}_{av}$) based on the weighted sum of the compensated audio ($\hat{S}_a$) and visual biases ($\hat{S}_v$):

$$\hat{S}_{av} = w_a\hat{s}_a + w_v\hat{s}_v \qquad (1)$$

Bias was defined as the PSE of the cumulative Gaussian fit to the psychophysical data, and reliability the inverse of its variance. Bias and reliability therefore correspond to the accuracy and variance of body-centred cues i.e. audio, visual or audio-visual image motion that has been transformed into body-centred coordinates using an estimate of head rotation (see Figure 1). The weightings were based on the reliabilities (rel_a, rel_v) of the predicting conditions listed in Table 2:

$$rel_a = \frac{1}{\sigma_a^2} \qquad rel_v = \frac{1}{\sigma_v^2} \qquad (2)$$

$$w_a = \frac{rel_a}{rel_a + rel_v} \qquad w_v = \frac{rel_v}{rel_a + rel_v} \qquad (3)$$

It's important to note that the standard cue integration approach being referred to here is typically applied to situations where bias has been introduced externally via small cue conflicts (Alais & Burr, 2004b; Ernst & Banks, 2002; Landy et al., 1995). The explicit assumption in those papers is that the underlying signals themselves (i.e. the 'estimators') are unbiased, or 'internally consistent' (Burge et al., 2010). This is not the case here: the

compensation process produces body-centred auditory and visual cues that are not necessarily unbiased i.e. the PSE describing these 'predicting' conditions is not necessarily at a motion gain of 0 (perfect compensation). Nevertheless, the predicted precision for the cue-combined condition remains unchanged (see Scarfe & Hibbard, 2011). We return to the assumption of unbiased estimators when developing predictions for uncompensated image-motion cues (Model 3).

The standard cue integration approach also allows us to predict the precision of the audio-visual cue (σ_{av}). This is based on the variance of the respective auditory and visual predicting conditions i.e. the reciprocal of the reliabilities defined in Eq. 2:

$$\sigma_{av} = \sqrt{\frac{\sigma_a^2 \sigma_v^2}{\sigma_a^2 + \sigma_v^2}} \qquad (4)$$

Model 2: Body-Centred Integration, adjusted for shared noise (BCI+)

In the BCI model, the audio and visual cues are transformed into body-centred coordinates using the same self-movement signal. The transformed cues are therefore correlated because they share a common source of noise (σ_{sm}). To account for this, the BCI model can be modified to include the correlation (ρ) between these cues (Oruç et al., 2003). This changes the way the reliabilities of auditory and visual cues are calculated:

$$rel_a = {}^1\!/_{\sigma_a^2} - (\rho \sqrt{{}^1\!/_{\sigma_a^2} \times {}^1\!/_{\sigma_v^2}})$$

$$rel_v = {}^1\!/_{\sigma_v^2} - (\rho \sqrt{{}^1\!/_{\sigma_a^2} \times {}^1\!/_{\sigma_v^2}}) \qquad (5)$$

The augmented reliabilities can then be used to predict the audio-visual bias using Eq. 1 and 3. The shared noise also changes the prediction for audio-visual precision, which can be calculated as follows:

$$rel_{av} = \frac{{}^1\!/_{\sigma_a^2} + {}^1\!/_{\sigma_v^2} - 2\rho \sqrt{{}^1\!/_{\sigma_a^2} \times {}^1\!/_{\sigma_v^2}}}{1-\rho^2} \qquad (6)$$

$$\sigma_{av} = \sqrt{\frac{1}{rel_{av}}} \quad (7)$$

One approach to evaluating BCI+ is to fit the equations to the data by allowing the correlation (ρ) to be a free parameter. Note, however, that the correlation is defined as the ratio of the shared noise to the product of the noise associated with individual cues:

$$\rho = \frac{\sigma_{SM}^2}{\sqrt{\sigma_a^2 \sigma_v^2}} \quad (8)$$

We already know the variance of the individual cues (σ_a^2 and σ_v^2) from the psychometric functions of the predicting conditions. To estimate the variance of the self-movement signal (σ_{SM}^2) for each of the participants we used the technique described by Haynes et al. (2024), described in Appendix A. With σ_{SM}^2 now known, the goodness-of-fit of BCI+ can be directly compared to BCI, without having to correct for differences in the number of free parameters used.

Model 3: Image-Centred Integration (ICI)

The above two models are based on auditory and visual motion that has been transformed into body-centred coordinates. As a comparison, we compared their predictions to a model based on uncompensated cues i.e. auditory and visual image motion alone. The ICI model assumes that self-movement signals are ignored. To do this, note that performance in the auditory and visual conditions is limited by two sources of noise. Assuming these are Gaussian, we can use the variance sum law to isolate the precision of the image signals by subtracting the variance of the self-movement signal (σ_{SM}^2) from the respective auditory and visual predicting conditions:

$$\sigma_{a\_Im} = \sqrt{\sigma_a^2 - \sigma_{SM}^2} \qquad \sigma_{v\_Im} = \sqrt{\sigma_v^2 - \sigma_{SM}^2} \quad (9)$$

We then used these to predict audio-visual precision using the equations defined in BCI.

As discussed above, models of cue combination typically assume that input signals like auditory and visual image motion are unbiased, or internally consistent (Burge et al, 2010). On that basis, the ICI model predicts no auditory-visual bias, or put another way, an audio-visual PSE of 0 (i.e. veridical). To reiterate, the predicted precision for the ICI model is unaffected by any bias related to the signal inputs (see Scarfe & Hibbard, 2011).

Across-Trial Noise Analysis

At first sight, manipulating stimulus motion based on a proportion of ongoing self-movement, or motion gain, seems to resolve the problem that head movements vary both within and across trials. However, as discussed in Haynes et al. (2024), the judgement made by the participant is based on inputs coded as motion amplitude, rather than motion gain. Variability in head movements at each level of stimulus gain also leads to unavoidable across-trial noise that is not accounted for by fitting a standard psychometric function. This additional across-trial noise can thus introduce surprising effects on the psychometric function, including a steeper slope than is fit by a standard cumulative Gaussian, and – at high head-movement variabilities – function asymptotes that deviate significantly from 0 and 100% that cannot be accounted for by the constrained lapse rates used in our standard analysis (Haynes et al., 2024). To address this issue, we therefore also fit the data with a custom psychometric analysis which captures this additional source of noise, and then re-fit the BCI, BCI+ and ICI models based on the result (see Appendix B for full details). The goodness-of-fit of each model to the audio-visual condition data was calculated using Root Mean Square Error (RMSE).

Statistics

Non-parametric bootstraps of 2000 samples were used to estimate standard errors around each psychometric function parameter using the PAL_PFML_BootstrapNonParametricMultiple function from the Palamedes toolbox in

MATLAB, in order to construct confidence intervals. Confidence intervals were used to test for significant differences between empirical psychometric functions and predicted psychometric functions across each model in individual subjects.

To construct confidence intervals for the BCI+ and ICI models, bootstrapped psychometric function parameters were obtained for each of the three repetitions of the self-movement paradigm (Appendix A). As for the empirical estimate, the variance sum law was used to extract the variance of self-motion signal (σ_{SM}^2) for each repetition, and the mean taken as the final measure. Due to occasional extremes in sampling during the bootstrapping procedure, it was sometimes possible to obtain negative self-movement variances. This occurred when the bootstrapped sample produced self-movement precision was greater than the precision of the audio and visual cues alone. These samples were excluded from calculation of the final bootstrapped self-movement variance parameter. Thus, it was possible for the final σ_{SM}^2 estimate to be based on the mean of one, two, or three repetitions. Of the 2,000 bootstrapped estimates, this occurred 17.79%, 39.3% and 39.2% of the time, respectively. Only 3.71% of the 2000 bootstrapped self-movement estimates were excluded entirely i.e. when all three repetitions resulted in negative self-movement variances.

To analyse which model best predicted the data on a group level, repeated measures ANOVAs with Model and Condition as factors were conducted on the squared errors from the model predictions and precisions and biases, as well as on the RMSE values from the Across-Trial Noise analysis. Significant effects were followed up with post-hoc Bonferroni-corrected t-tests.

Eye and Head Movement Analysis

Eye movements were analysed to ensure that participants maintained fixation on the fixation point. Eye movement analysis followed that of Haynes et al. (2024). Samples with less than 0.6 confidence (defined by PupilLabs software) were excluded, with gaps filled using linear interpolation (Halow et al., 2023). The entire waveform was excluded from analysis if

50% or fewer samples remained. A Gaussian filter (σ = 16 Hz frequency domain) was used to smooth remaining waveforms, and velocity, acceleration, and jerk were established by taking the 1st, 2nd and 3rd derivatives respectively. The initial and final 20 samples of each waveform were removed. Saccades (and 4 samples either side) were removed from the analysis, with saccades detected using Wyatt's jerk analysis (jerk threshold = 20,000°/s^2). Mean velocity and speed in the third head sweep were then calculated. For comparison, we calculated the eye velocity expected for compensatory vestibulo-ocular reflex (VOR) using the equation $E = -H(1 + R/D)$ (Leigh & Zee, 2015), where H is the average head velocity per condition across participants, R = 0.1 m (the approximate distance from the eye to centre of head rotation) and D = 1.2 m (the distance from participant to speakers/LED ring). Eye movements could not be recorded from one participant due to technical problems.

Recorded head movements were smoothed using MATLAB's lowpass filter with a passband of 8 Hz. The temporal derivative was taken, and the median velocity was calculated over 20-60% of the 3rd sweep length because we have previously shown this region-of-interest to produce a stable estimate (Haynes et al., 2024). The distribution of median velocities for each participant in each condition was then fit with a Gaussian to extract the mean and variance of head velocity.

Data Availability

Data and analysis codes are available on the Open Science Framework: https://osf.io/yj27m/

Results

Psychophysics

Precision

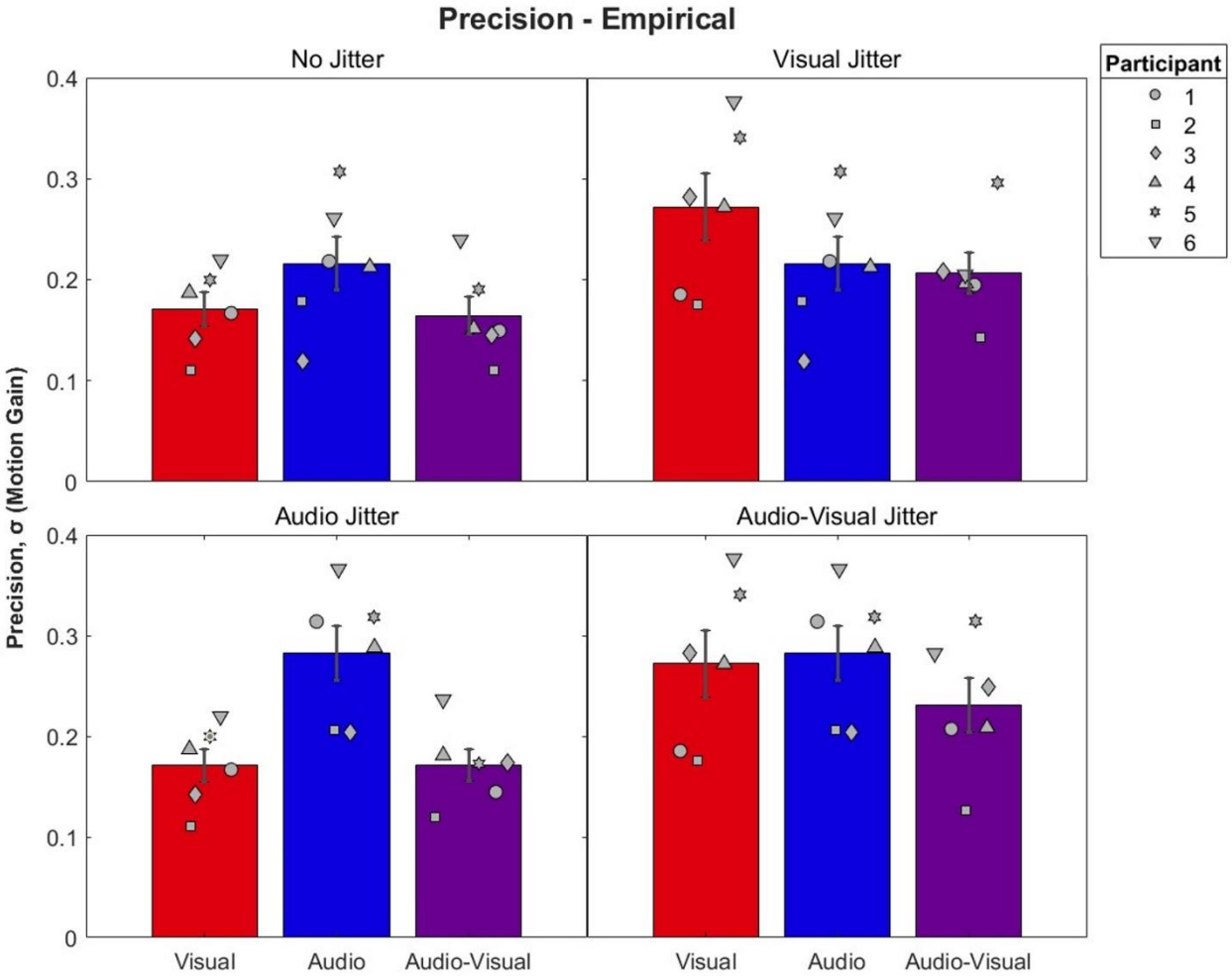

Figure 4. *Mean and individual precisions by condition. Audio-visual bars are presented alongside their predicting conditions, with visual and audio conditions indicated by red and blue respectively. Note that predicting conditions can appear more than once across panels (Table 2). Error bars represent ±1 standard error between-participants. Lower numerical values indicate smaller standard deviations, and thus greater precision. Participants 5 and 6 were naïve to the hypotheses of the study.*

Mean and individual precisions for each condition can be seen in Figure 4. As expected, the addition of positional jitter decreased the precision of the cue(s) to which it was applied, with the biggest decrease in precision when applied to both cues, indicated by larger standard deviations. Crucially, audio-visual precision was similar or better compared to the most precise

audio or visual cue in each condition, with the most benefit of integration observed when both visual and audio signals had similar levels reliability in the Audio-Visual Jitter condition.

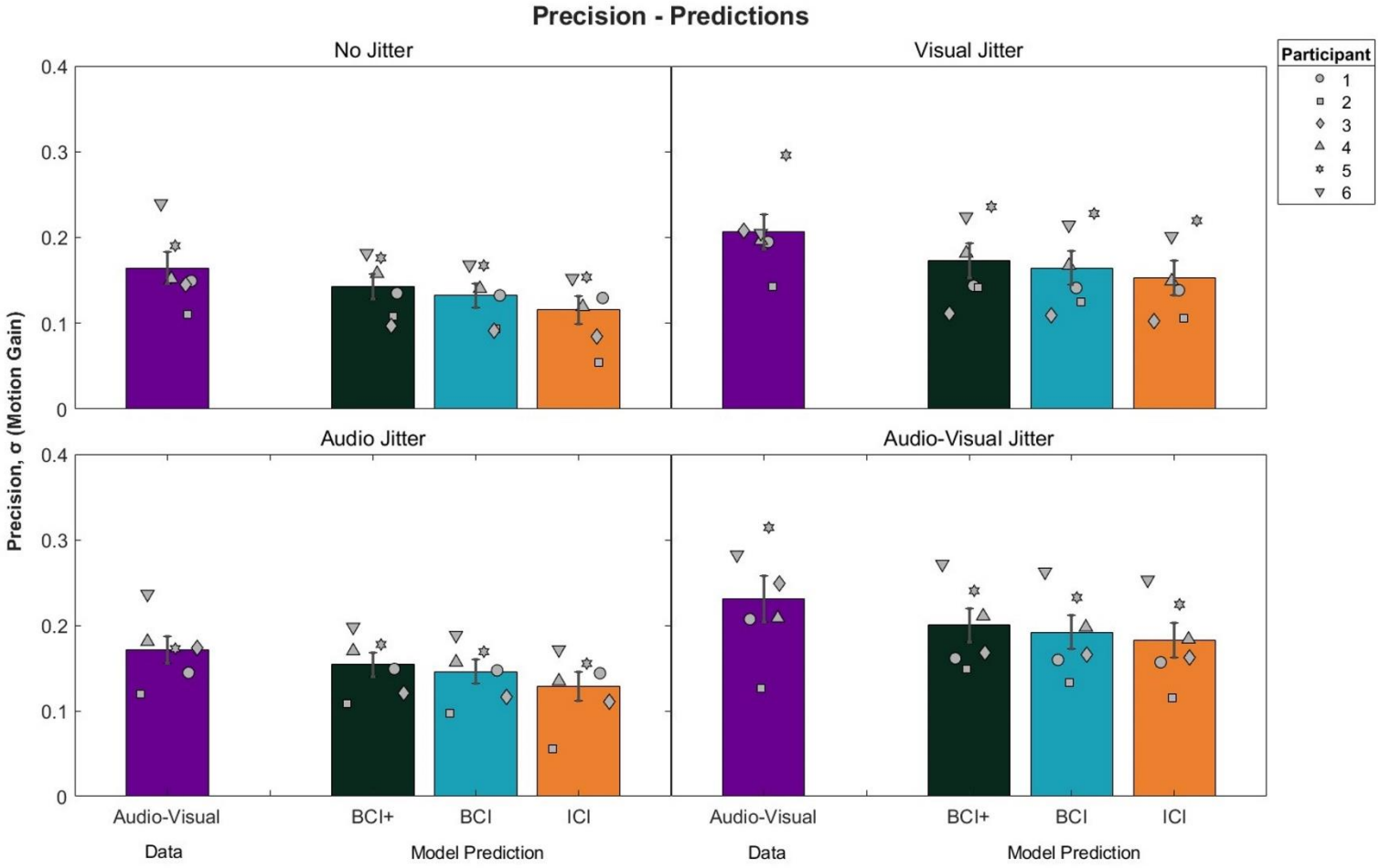

Figure 5*. Audio-visual precisions and model predictions by condition Error bars represent ±1 standard error between-participants. Lower numerical values indicate smaller standard deviations, and thus greater precision. Participants 5 and 6 were naïve to the hypotheses of the study.*

Model predictions compared to the audio-visual conditions can be seen in Figure 5. In general, the BCI+ model resulted in the closest prediction to the actual performance obtained for the audio-visual conditions. ICI predictions were the furthest from actual performance, suggesting that cue integration was based on some form of coordinate transformA 3x4 repeated measures ANOVA was conducted on the squared differences between model predictions and observed precisions (Table 3). A main effect of Model was found ($F(2, 10) = 26.16$, $p < .001$, $\eta^2_G = 0.05$). Bonferroni-corrected pairwise t-tests revealed that the BCI+ model

was significantly better at predicting precisions compared to the BCI model (average squared errors: BCI+ = 0.0017, BCI = 0.002, p = .001) and to the ICI model (average squared errors: ICI = 0.003, p < .001). The BCI model was also better at predicting precisions compared to the ICI model (p < .001). No other main effects or interactions were significant (Condition: F(3, 15) = 0.83, p = .50, Model*Condition: F(6,30) = 0.73, p = .63). It is also worth noting that all three models predicted greater precision than actually observed, an important point taken up in the Discussion.

Table 3

Mean (SD) squared errors for Precisions by condition and model

Condition	ICI	BCI	BCI+
No jitter	0.0029 (0.0026)	0.0015 (0.0020)	0.0010 (0.0015)
Visual jitter	0.0040 (0.0040)	0.0031 (0.0037)	0.0027 (0.0035)
Audio jitter	0.0024 (0.0020)	0.0011 (0.0013)	0.0008 (0.0011)
Audio-Visual jitter	0.0033 (0.0036)	0.0027 (0.0032)	0.0024 (0.0029)

On an individual level, confidence intervals revealed that the BCI model predicted precisions in 16/24 cases, while the ICI model predicted precisions in 12/24 cases. The BCI+ model again had the best performance, predicting individual precisions in 19/24 cases (Figure 6).

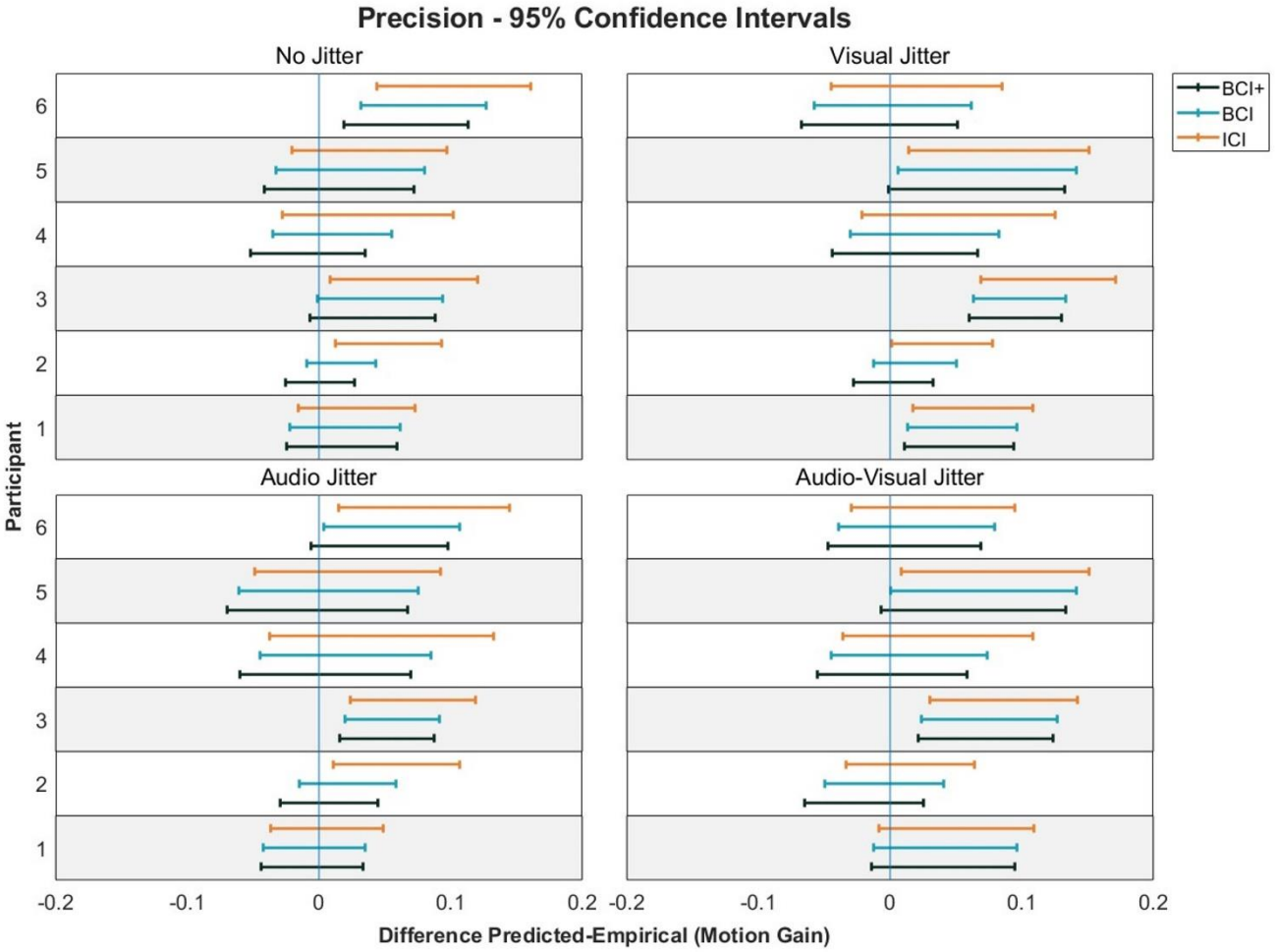

Figure 6. *95% Confidence intervals on the difference between predicted and empirical audio-visual Precisions. Confidence intervals that cross the 0 point indicate no significant difference between predicted and empirical precisions. Participants 5 and 6 were naïve to the hypotheses of the study.*

Bias

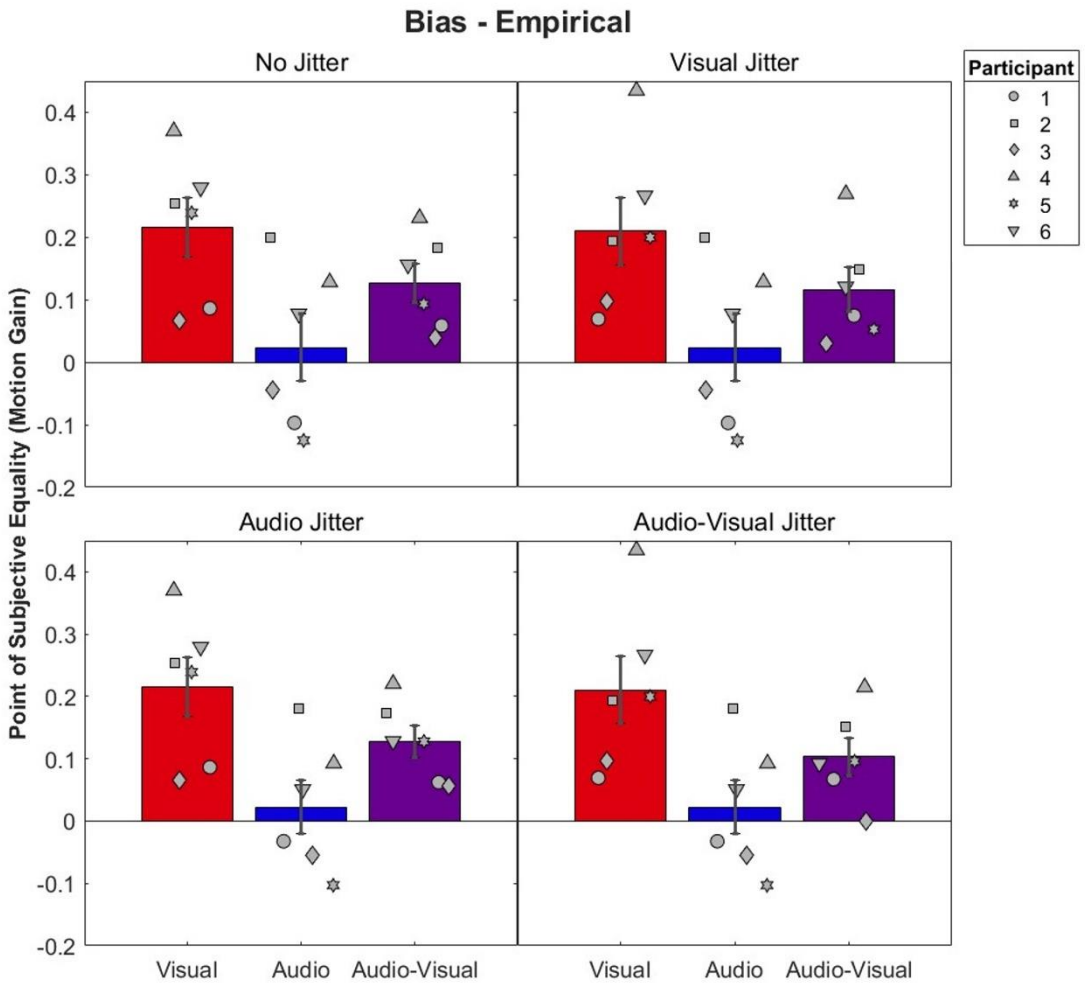

Figure 7. *Mean and individual biases by audio-visual condition. Audio-visual bars are presented alongside their predicting conditions, with visual and audio conditions indicated by red and blue respectively. Note that predicting conditions can appear more than once across panels (Table 2). Error bars represent ±1 standard error between-participants. Participants 5 and 6 were naïve to the hypotheses of the study.*

Mean biases for each condition can be seen in Figure 7. Biases were largest in all visual conditions, indicating that participants perceived a stationary visual stimulus to move in the opposite direction of the head movements (i.e., a Filehne Illusion, Filehne, 1922; Freeman, 2007; Haarmeier & Thier, 1996; Mack & Herman, 1973). As expected, audio-visual biases were in between visual and audio biases.

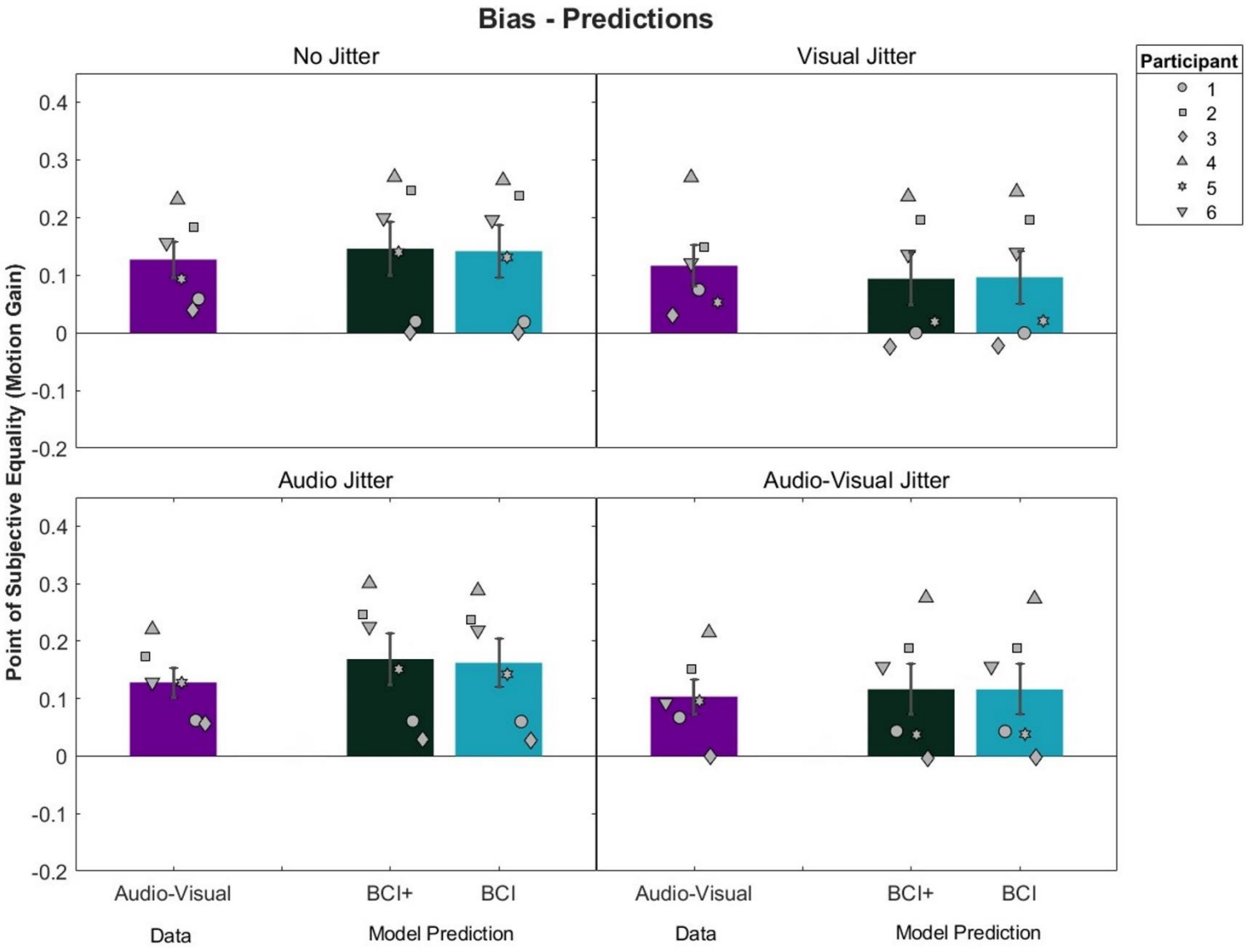

Figure 8*. Audio-visual precisions and model biases by condition. Error bars represent ±1 standard error between-participants. Participants 5 and 6 were naïve to the hypotheses of the study.*

Model predictions compared to the audio-visual conditions can be seen in Figure 8. Recall the ICI model, by definition, predicts a bias of 0 and so is not included in the figure. In general, both models predicted similar biases. The 2 x 4 repeated measures ANOVA on the squared differences between model predictions and observed biases was conducted (Table 4). This analysis revealed a significant main effect of Model ($F(1,5) = 7.63$, $p = .04$, $\eta^2_G = 0.006$), with lower squared errors for the BCI (mean = 0.022) vs BCI+ (0.0026) model. No main effect of Condition was found ($F(3,15) = 0.47$, $p = .71$). A significant interaction between Model

and Condition was found ($F(3, 15) = 3.97$, $p = .03$, $\eta^2_G = 0.005$), with lower squared errors for the No jitter and Audio jitter conditions for the BCI model, and similar squared errors across both models in the visual and audio-visual jitter conditions. However, Bonferroni-corrected pairwise t-tests revealed no significant differences across any model or condition.

Table 4

Mean and SD squared errors for Biases by condition and model

Condition	BCI	BCI+
No jitter	0.0017 (0.0007)	0.0021 (0.0010)
Visual jitter	0.0021 (0.0020)	0.0022 (0.0019)
Audio jitter	0.0030 (0.0032)	0.0037 (0.0039)
Audio-Visual jitter	0.0021 (0.0017)	0.0022 (0.0017)

On an individual level, both models predicted biases in 17/24 cases (Figure 9). There was no consistent pattern in whether models over- or underestimated biases.

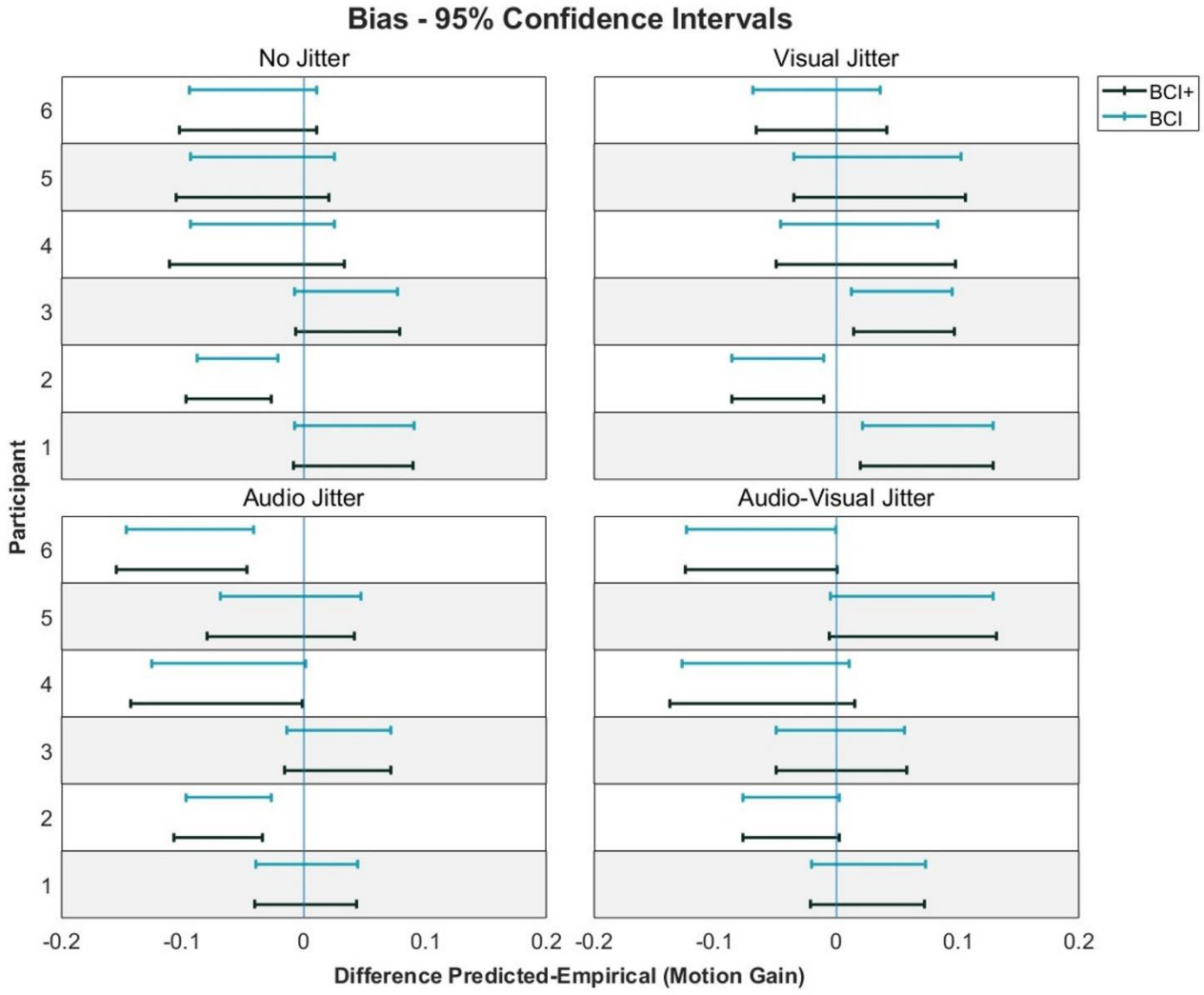

Figure 9. *95% Confidence intervals on the difference between predicted and empirical Audio-Visual Precisions. Participants 5 and 6 were naïve to the hypotheses of the study.*

Eye and Head movements

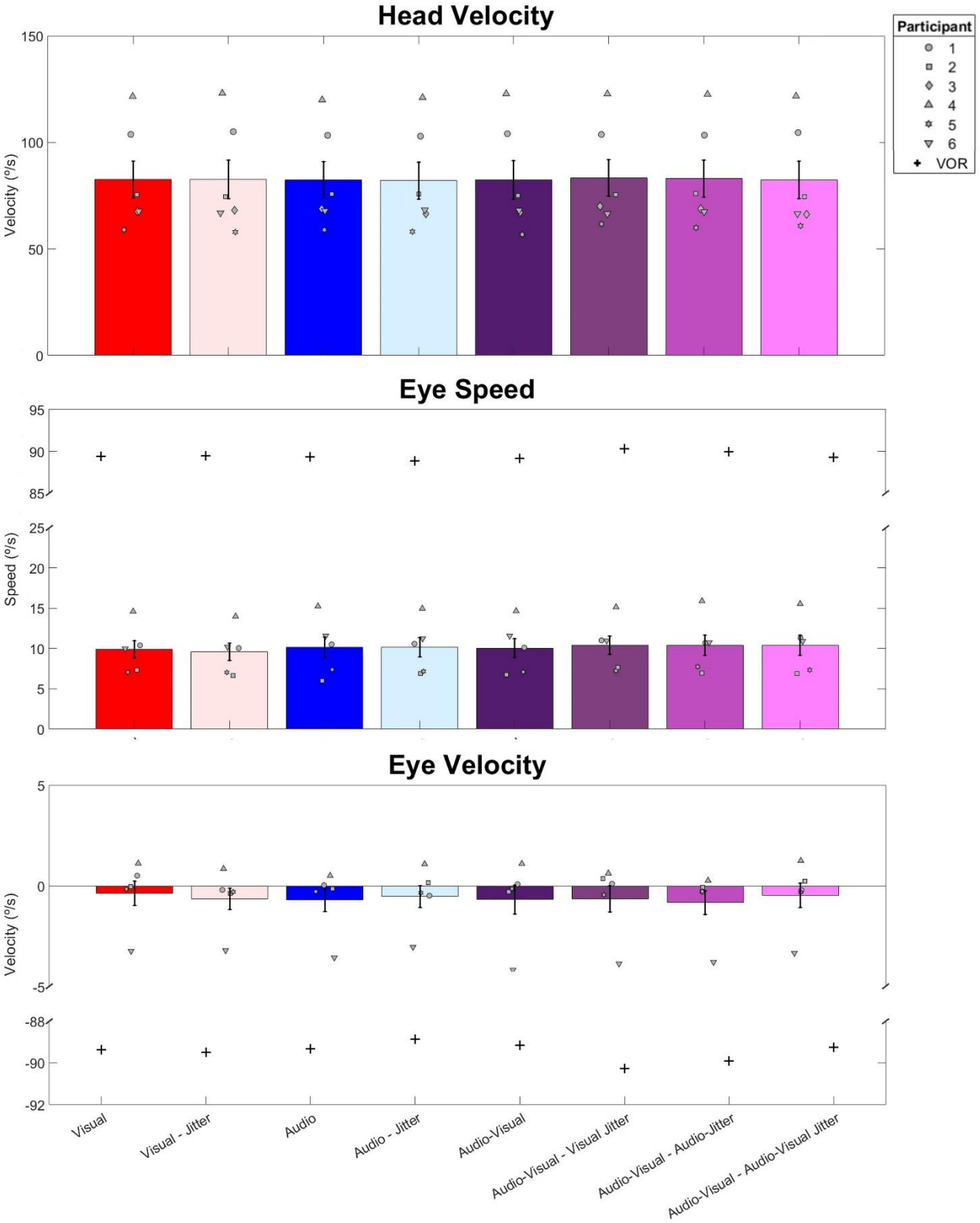

Figure 10. *Top: Absolute head movement velocities across all six participants in each condition. Middle: Eye movement speeds from 5/6 participants. Bottom: Eye movement velocities from 5/6 participants. Negative velocities indicate that the eyes move in the opposite direction to the head movement. Bars show means across participants, error bars represent ±1 standard error between-participants. Plus symbols indicate compensatory VOR. Participants 5 and 6 were naïve to the hypotheses of the study.*

As can be seen in Figure 10, head movement velocities within participants were similar across all 8 conditions of the main task, although speeds across participants varied. Importantly, eye movements were much smaller than would have been expected from the VOR (+ symbols in Figure 10) needed to perfectly compensate the observed head movements, suggesting participants were able to keep the eyes fixed in the head throughout the experiment.

Across-Trial Noise Analysis

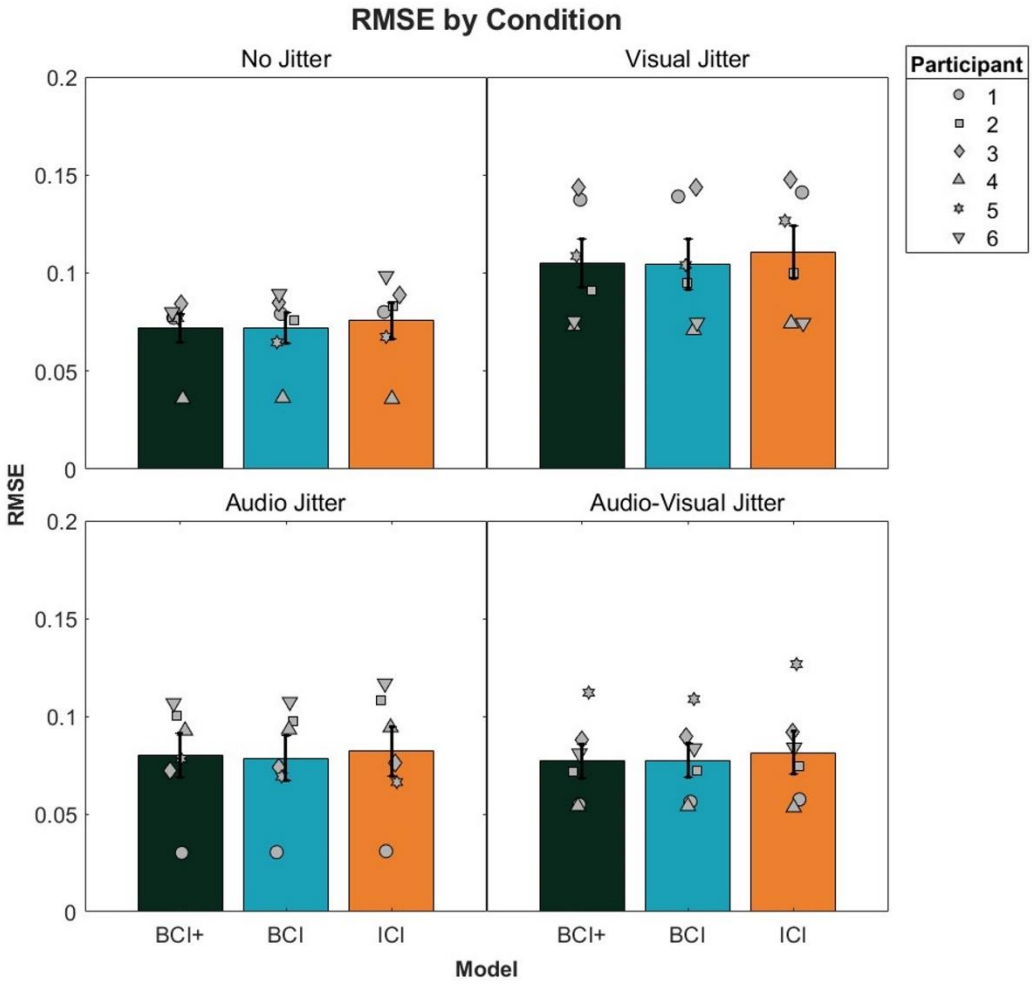

Figure 11. *RMSE for model fits to each audio-visual condition. Error bars represent ±1 standard error between-participants. Participants 5 and 6 were naïve to the hypotheses of the study.*

Mean and standard error RMSEs for each model and audio-visual condition can be seen in Figure 11. In general, the BCI and BCI+ models were similar across conditions, but both of these models were better than the Image-Only model (ICI) in all conditions.

A 3 x 4 repeated measures ANOVA with Model and Condition as factors was conducted on the RMSEs. The main effect of Model was significant ($F(2, 10) = 6.49$, $p = .02$, $\eta^2_G = .007$). Bonferroni-corrected paired t-tests revealed significant differences between the BCI+ and ICI models (average RMSE: BCI+ = 0.084, ICI = 0.087, $p = .03$), and the BCI and

ICI models (average RMSE: BCI = 0.083, p = .005). No significant difference was found between the BCI+ and BCI models. No other main effects or interactions were significant (Condition: $F(6, 30) = 1.95$, p = .16, Model*Condition: $F(6,24) = 0.24$, p = .68 [Greenhouse Geisser correction for sphericity]). On an individual level, the BCI+ model was the best-fitting model in 14/24 cases, although in general the difference in RMSE between the BCI and BCI+ models was small.

While this ‘across-trial noise’ analysis appears to show less difference between BCI and BCI+ models, compared to the more standard analysis examined earlier, it is important to note that these two approaches differ markedly in how the psychometric function parameter predictions are assessed. In particular, the more standard analysis considers predictions based on precision and bias separately, while the ‘across-trial noise’ analysis in Figure 11 groups these two parameters together. Moreover, conclusions are drawn about the latter on the basis of the goodness-of-fit of model-driven psychometric functions to the raw psychophysical data, as opposed to comparing empirical and model bias and precision in the case of the more standard model. Given that the more standard analysis showed minimal differences in bias predictions across models (e.g. Figure 8), we suspect that this is the reason why we find smaller differences between them when using the ‘across-trial noise’ analysis.

Overall, both approaches suggest that integration is based on compensated cues, given the better performance of the BCI and BCI+ models in comparison to the ICI model. Moreover, the more standard approach shows that audio-visual precision is best accounted for by the BCI+ model, both at a cohort and individual level.

Discussion

The integration of sensory signals is made complicated during self-movement because cues may be represented in different coordinate frames, and are subsequently affected differently by head and/or eye movements. We proposed that integration of audio and visual signals occurs following the transformation of these cues into a common coordinate frame. Participants completed a psychophysical multisensory integration task while making active yaw head rotations and fixating on a head-fixed target. Participants judged whether a visual, auditory, or audio-visual target, presented at high or low stimulus reliabilities, moved left or right. We found that performance in the audio-visual conditions was well-described by models based on the combination of 'compensated' audio and visual signals represented in a common coordinate frame. A model based on uncompensated image-based signals could not explain performance in the majority of audio-visual conditions in most participants. We also found that taking into account the inevitable shared noise between compensated cues accounted for the increase in audio-visual precision better than other models. Accordingly, our data show that it is likely that audio and visual cues are first transformed into common coordinates before these signals are integrated according to principles of optimal integration.

Previous studies have considered the integration of moving audio-visual targets in stationary observers and consistently found evidence for optimal integration (Alais & Burr, 2004a; Meyer & Wuerger, 2001; Soto-Faraco et al., 2004; Wuerger et al., 2010). For instance, variability in the estimate of moving object arrival times is reduced when both audio and visual cues are present, consistent with MLE (Wuerger et al., 2010). Similarly, detection thresholds for motion perception are significantly improved in the presence of audio-visual stimuli together, rather than individual modalities alone (Alais & Burr, 2004a). By contrast, few studies have considered the impact of self-motion on the process of integration, although several studies have demonstrated that self-movement can impact the perception of auditory and visual stimulus motion more generally. For example, rotation and translation of the body and/or head can impact the localisation of both auditory and visual targets (Carriot et al., 2011; Cooper et

al., 2008; Lackner & DiZio, 2010; Teramoto et al., 2014), and stationary stimuli are perceived as moving in the opposite direction to head and/or eye movements (Freeman, 2007; Freeman et al., 2017). This latter so-called Filehne illusion is also present in our current study, reflected in the biases obtained for both audio, visual, and audio-visual conditions. Curiously, while we find similar visual biases to those previously-reported for smooth eye pursuit (Filehne, 1922; Furman & Gur, 2012), auditory biases are much smaller than in an earlier study from our lab that used similar auditory stimuli (Freeman et al., 2017). The smaller biases may have arisen because here we presented the cues to be judge during a single 'sweep' of the head movement, instead of continuously as in Freeman et al. (2017). We note too that the Filehne illusion depends on basic stimulus properties such as spatial frequency (Freeman & Banks, 1998; Wertheim, 1987), which determine the size of the image-motion estimate to which the self-motion signal is compared. Hence the Filehne illusions found for the visual and auditory conditions will also depend on the specific auditory and visual stimuli used. As such, it is possible that we find a larger visual versus auditory Filehne illusion in the present study due to differences in stimulus parameters, such as the standard deviation of the stimuli. results build upon this existing literature, proposing a mechanism through which self-movement, visual motion and auditory motion are integrated to help us perceive movement.

In our study, participants maintain their gaze on a head-fixed target while making yaw head movements, meaning that both the eyes and ears move with respect to the external world. Accordingly, to successfully make directional judgements as in the present task, observers must account for this self-movement and put audio and visual cues into a common reference frame prior to integration. We suggest that the common coordinate frame is body-centred, given that this reference frame remains stationary during head and eye movements, and is parsimonious given the evidence both here and in other papers that both auditory and visual signals can be transformed into this reference frame (Furman & Gur, 2012; Goossens & van Opstal, 1999; Kopinska & Harris, 2003; Lewald et al., 1999). Yet it is possible that there is an alternative route to cue integration. In particular, previous research on auditory

localisation has suggested that auditory signals are first transformed into an eye-centred reference frame, and then integrated with visual signals before a transformation to body-centred coordinates at a later stage (Lee & Groh, 2012; Lewald et al., 2000). The present data cannot differentiate between these alternate transformation routes, given that eye- and head-centred reference frames coincide due to the type of head movement employed in the study. However, it is important to note that firstly, the eye-centred model has been used to explain the influence of eye gaze on auditory localisation alone, rather than audio-visual cue combination as explored here, and secondly, the eye-centred model is based on localisation of auditory cues in stationary observers, rather than in cases when the head and eyes move. . In addition, as we pointed out in the Introduction, there is good evidence that localisation and gaze orienting to auditory and visual targets is driven by body-centred not eye-centred coordinates (Goossens & van Opstal, 1999; Kopinska & Harris, 2003). Moreover, evidence suggests a host of coordinate frames are represented at different neural levels. For example, auditory and visual motion may be represented in eye-centred, head-centred or even in hybrid frames across regions including the superior and inferior colliculus, primate ventral intraparietal cortex (VIP), and V1 (Bulkin & Groh, 2006; Furman & Gur, 2012; Ilg et al., 2004; Ilg & Thier, 1996; Lee & Groh, 2012; Zhang et al., 2004). Future research is therefore necessary to delineate precisely which coordinate frame transforms are conducted during multisensory integration, and in what order. While we propose a body-centred model, many alternatives are possible given the diversity of neural representations and sensory tasks implicating optimal integration.

Outstanding questions remain regarding which sensory signals are used to compensate for self-movement in our paradigm. It is likely that self-movement compensation would include cues from the vestibular system, neck proprioception, efference copy, and signals from eye muscles (Dokka et al., 2015; Furman & Gur, 2012; Genzel et al., 2016). The role of the eye muscles and efference copy has been widely explored in relation to visual localisation during smooth pursuit eye movements (Bogadhi et al., 2013; Furman & Gur, 2012).

Similarly, visuo-vestibular integration is necessary to compensate for translational self-movement (Dokka et al., 2015), while proprioceptive and vestibular signals are required for auditory spatial updating (Genzel et al., 2016). In our research, we considered a unified 'self-movement' signal as the source for self-movement compensation, without distinguishing which inputs formed this signal. As such, future research should aim to more precisely define which sensory modalities form this self-movement signal, and under which circumstances they are combined.

Our results and analysis suggest that precision in the audio-visual conditions was not completely optimal. As expected, we observed the greatest increase in audio-visual precision when both audio and visual cues had similar reliabilities. However, when audio and visual reliabilities diverged, audio-visual precisions were close to the 'best' unimodal condition. In general, non-jittered visual signals were more precise than auditory signals. Accordingly, it may be difficult to establish whether participants used a "best cue" strategy, rather than an optimal integration strategy. However, the clearly increased precision in the audio-visual jittered condition would suggest that participants did indeed engage in (near) optimal integration. Given that all trial types were interleaved, it seems unlikely that participants would switch between these alternate strategies on a trial-by-trial basis. Nonetheless, to further assess whether and how much audio-visual integration deviates from optimality, and to better distinguish between a "best cue" vs optimal integration strategy, future research may be necessary – for example, through individually tailoring audio and visual precisions to more clearly explore the increased precision apparent in audio-visual conditions, or through introducing cue conflicts to examine whether predicted and measured audio and visual weightings diverge (Rohde et al., 2016).

Finally, the models presented here predicted greater precision than was actually observed, suggesting an additional unaccounted source of noise. While we emphasise the conversion of signals into a common *reference frame*, it is likely that further conversions are needed to transform audio and visual signals into common *units*. When observing moving

objects, visual cues are dominated by speed (Freeman et al., 2018; Reisbeck & Gegenfurtner, 1999), while auditory cues are dominated by displacement (Carlile & Best, 2002; Freeman et al., 2014). Thus, when combining hearing and vision, auditory and visual cues must also be transformed into common units (i.e., displacement to speed, or speed to displacement). This process likely adds noise (Haynes et al, 2024). As we predicted audio-visual conditions on the basis of separate audio and visual conditions, it is possible that this unit conversion noise is missing from our final predictions. After all, the need for common units is only necessary when hearing and vision are directly compared to each other, which in our experiments corresponds to the audio-visual conditions, not the predicting conditions. Furthermore, when the eye and/or head is moving, an additional unit conversion is needed between self-movement and image-motion signals. Vestibular and motor signals are likely to be encoded in terms of acceleration and speed units (Angelaki & Cullen, 2008; Cullen, 2019; Freeman et al., 2018), making the combination of self-movement and visual motion relatively straightforward. However, additional conversion is needed to combined speed-based self-movement cues with displacement-based auditory cues. Given that our predicting conditions involved the same self-movement as the combined audio-visual conditions, we have likely accounted for the unit conversion noise for each modality separately. However, it is possible that the *shared* noise from these conversions remains unaccounted for. Accordingly, future experiments and models are needed to resolve the unit conversion problem, over and above the reference frame issue we present here.

Conclusions

Overall, here we demonstrate that audio-visual motion perception during active self-movement is based on the combination of sensory signals which are transformed into a common coordinate frame. Importantly, our study investigated audio-visual integration during self-generated, active movements, expanding our knowledge of multisensory integration to more naturalistic task constraints. We propose that the common coordinate frame is body-centred, however alternatives may be possible based on the modalities and tasks involved in

any given multisensory scenario. Accordingly, this research opens a new avenue for future work to investigate integration during natural, active self-movement.

Acknowledgements

The work here was supported by a Leverhulme Trust project grant (RPG-2018-151). We would also like to thank Dr Richard Morey for deriving equation (8)

Appendix A: Measuring Self-Movement Variability

A separate paradigm was used to measure the precision of self-movement signals, based on Haynes et al. (2024). The precision of self-movement was required to calculate the shared noise between the audio and visual signals for the BCI+ model prediction (see Equation 8). All participants from the main paradigm completed this additional paradigm in a separate session after completing the main experiment. We measured speed discrimination in a separate 2-interval forced choice (2IFC) task with two phases: one containing a self-movement signal and one without this signal. Briefly, in Phase 1, the standard interval consists of the participant moving the head while head-fixed visual stimulus appears 'on the nose' (and consequently does not generate 'image' motion across the retina). The test interval consists of the same visual stimulus that is scaled by a motion gain, but presented while the head is stationary. The participant is then asked in which interval the stimulus appears to 'move more', and a psychometric function is obtained. In Phase 2, the visual stimuli from Phase 1 are replayed, but the head remains stationary throughout. Accordingly, we assume that the slope of psychometric function in Phase 1 is limited by two sources of noise, one corresponding to the self-movement signal and the other the image-motion signal, while Phase 2 is limited by noise from the image signal alone. Using the variance sum law, the precision of self-movement signals can therefore be estimated by subtracting half of the variance of Phase 2 from Phase 1.

Procedure

Participants were first trained to move their head at the average speed they moved in the main paradigm. A red light moved sinusoidally with the velocity and amplitude of each participants' head movements, measured in the main paradigm. Participants had to memorise the motion of this light, and then replicate it by moving their head with a red light yoked to their head movements. Trials were classed as correct if they fell within ±10% of the target speed. Participants repeated the training at least three times, until 75% of trials were correct. We have

shown previously that this type of training paradigm can be used to enable individual participants to reproduce a wide range of different average head speeds (Haynes et al., 2024).

The task consisted of a 2IFC, repeated in two phases. In Phase 1, Interval 1, participants made back-and-forth yaw head movements at the trained speed. On the third head sweep, a green target light was yoked to the head movement. In Phase 1, Interval 2, participants kept their head stationary while the target light was replayed and scaled to a proportion of the original head speed. Participants then reported whether the target moved faster in the first or second interval. A cumulative Gaussian psychometric function was fitted to response data of Phase 1, with the Point of Subjective Equality used to determine the speed of the targets in Phase 2. In Phase 2, both intervals from Phase 1 were replayed with the participant remaining stationary throughout. In Interval 1, target speed was presented at the PSE from Phase 1. Interval 2 targets were scaled to a proportion of the Interval 1 target speed. Participants again had to judge which interval contained the faster target. The entire procedure was repeated three times in total.

Stimuli and Equipment

Stimuli were presented with the same LED ring as the main paradigm. Eye and head movement were collected using the same equipment as the main paradigm.

The target light was a diffuse green LED blob spanning approximately 2.25°. The target light was yoked to the participants' head movements in Phase 1, Interval 1. In Phase 1, Interval 2, the target light moved on the same path determined by the Interval 1 head movements, but its velocity was scaled as a proportion of head speed, from 40-100% of head speed in seven steps.

In Phase 2 the targets moved on the same path as Phase 1, Interval 1, determined by the head movement. Target velocity in Interval 1 was scaled to match the perceived speed of the target, determined by the Point of Subjective Equality (PSE) from Phase 1. Interval 2 target velocity was the PSE ± 30% in seven steps.

In both phases, each target velocity was repeated 10 times, and the entire paradigm was repeated three times, giving a total of 210 trials per phase (7 target velocities * 10 presentations * 3 repetitions). The entire paradigm took approximately 1 hour to complete.

Data Analysis

Psychometric Analysis

Data were analysed in MATLAB r2022b. The proportion of trials in which the participant responded the target was faster were calculated for each target velocity. For each repetition of Phase 2, target velocities were normalised to 0±30% of head speed by subtracting the PSE. Cumulative Gaussian psychometric functions were fitted using the Palamedes Toolbox, using the PAL_PMFL_FIT function. PSE, slopes, and lapse rates were free parameters, with lapse rates constrained to values between 0-0.02 (Prins, 2012). The PSE was defined as the 50% point of the psychometric function. The precision was defined as the inverse of the psychometric function slope (i.e., the standard deviation of the cumulative Gaussian fit to the data), such that larger numerical values indicated poorer precision.

The variance of the self-movement signal was calculated from the precision values obtained from Phase 1 (*Ph1*) and 2 (*Ph2*):

$$\sigma_{SM}^2 = \sigma_{Ph1}^2 - \frac{\sigma_{Ph2}^2}{2}$$

Note that the precision values of Phase 2 are halved, as both intervals contain image noise, whereas Phase 1 only contained image noise during the test interval. The self-movement variance was calculated for each repetition, and the square-root of the average variance used to estimate self-movement precision for the BCI+ model.

Estimates of the variability around parameter estimates were calculated as described for the main experiment, using 2,000 nonparametric bootstrapped samples with the function PAL_PFML_BootstrapNonParametricMultiple.

Note that this analysis is consistent with the standard psychophysical approach used in the main text, whereas Haynes et al. (2024) used an analysis consistent with the Across-Trial Noise analysis outlined in Appendix B. As noted in the main text, there is little difference between these two approaches.

Head and Eye Movement Analysis

Head and eye movements were analysed in the same way as the main experiment. Eye movement data for one participant was not recorded due to technical problems.

Results

Table 1

Self-Movement Variability Results

Participant	Phase 1 PSE (Motion gain)	Self-Movement Precision (Motion gain)	Head Movement Velocity (°/s)	Eye Movement Velocity (°/s)
1	0.64	0.038	123.23	-0.49
2	0.73	0.094	64.16	-0.20
3	0.78	0.046	57.49	N/A
4	0.72	0.104	88.78	-0.01
5	0.60	0.086	113.99	-2.75
6	0.69	0.099	71.30	0.20

Results can be seen in Table 1. Phase 1 PSEs for all participants were < 1, indicating that stimuli pursued by the head were perceived as slower than those that were not pursed. I.e., a PSE of 0.7 indicates that a stimulus moving past a stationary participant has to be slowed by 30% to be perceived as moving at the same speed as during the pursuit interval. This effect thus resembles a classic Aubert-Fleischl effect (Aubert, 1886; Dichgans et al., 1975;

Garzorz et al., 2018), driven by head movements rather than eye pursuit. Similar findings are reported by Haynes et al. (2024).

Two participants had negative self-movement variance estimations on at least one repetition of the paradigm. These repetitions were excluded from the final average estimate of self-movement variances. The square root of the precisions reported in Table 1 (i.e., the self-movement variance) were used to generate model predictions in the main experiment.

Head movement velocities were similar to those obtained in the main experiment, with the exception of participant 5 whose head speeds were consistently faster compared to the main experiment. Eye movement velocities were negligible, indicating that participants were successfully able to track the head-fixed fixation point.

Appendix B: Across-Trial Noise Analysis

Unimodal conditions

On each trial, auditory or visual stimuli move across the speaker/ LED array. Stimulus motion (M) is made a fixed proportion (g) of the recorded head rotation (H) in real time:

$$M(t) = gH(t) \quad (1)$$

We refer to g as the 'motion gain'. When g = 1, the stimulus moves at the same speed and direction as the head rotation i.e. on the nose. When g = 0 , the stimulus is stationary. When g = -1, the stimulus moves at the same speed but in the opposite direction to the head.

In both the unimodal and bimodal conditions, the task for observer is to judge whether M moved to the left or right of the body during the head rotation made in the 3rd sweep. For hearing, M is the sum of image motion (I) and head rotation (H). In our experiment, this is also the case for vision because we inhibit eye rotations in the skull by providing a head-centred fixation target (akin to a scratch on a pair of glasses, albeit at a more comfortable viewing distance!). Eye movement analysis shown in Figure 10 indicates that they are able to do this very well, which confirms the finding of Haynes et al. (2024). Hence for both modalities:

$$M = I + H \quad (2)$$

To recover the body-centred motion M, the observer must estimate I and H from internal signals i and h. We assume both are corrupted by fixed Gaussian noise with 0 mean. Using $\mathrm{N}(\mu, \sigma)$ to denote a normal distribution with mean μ and standard deviation σ, then across trials:

$$i = \mu_i + \mathrm{N}(0, \sigma_i) \quad (3)$$

$$h = \mu_h + \mathrm{N}(0, \sigma_h) \quad (4)$$

Head movements will vary across trials, which means h must vary with this too. The mean μ_h is therefore a random variable. This is also the case for i because stimulus motion

is yoked to the head movement as defined by Eqn (1). Assuming H is also Gaussian distributed across trials, and noting that $I = (g-1)H$ from Eqns (1) and (2):

$$i = b_i(\mathrm{g}-1)\mathrm{N}(\mu_H, \sigma_H) + \mathrm{N}(0, \sigma_i) \quad (5)$$

$$h = \mathrm{N}(\mu_H, \sigma_H) + \mathrm{N}(0, \sigma_h) \quad (6)$$

where b_i is a bias term that sets the gain of the image-movement signal relative to its input i.e. $i = b_i I$. Note the bias term also captures the relative difference in accuracy between the head movement signal and image-movement signal. Although more general accounts include a bias term to h (see, for instance, Freeman & Banks, 1998), if we assume linearity as we do here, perceived motion would be determined by the ratio of these two bias terms. This is equivalent to the single parameter b_i (see Freeman, 2001, for a comparison between linear and non-linear accounts).

Combining (5) and (6) with (2), perceived motion (m_i) is given by:

$$m_i = (b_i(\mathrm{g}-1)+1)\mathrm{N}(\mu_H, \sigma_H) + \mathrm{N}(0, \sigma_h) + \mathrm{N}(0, \sigma_i) \quad (7)$$

In the experiment, motion gain g is varied across trials and observers judge whether m_i was to the left or right. The resulting psychometric function describes the probability of judgements in a particular direction (e.g. rightward) as a function of motion gain. Following standard signal detection theory (e.g. Jones, 2016), m_i is an internal decision variable. Hence the choice 'appeared to move rightward' corresponds to $m_i > 0$. From signal detection theory we define:

$$d = \frac{\mu_{m_i}}{\sigma_{m_i}} \quad (8)$$

such that the probability of choosing rightward is given by:

$$P = \frac{\lambda}{2} + (1-\lambda)\Phi\left(\frac{d}{\sqrt{2}}\right) \quad (9)$$

where λ is the lapse rate and Φ is the cumulative distribution function of the standard normal distribution.

By inspection, from Eqn(7):

$$\mu_{m_i} = (b_i(\mathrm{g} - 1) + 1)\mu_H \quad (10)$$

Variances sum, so again from Eqn (7):

$$\sigma_{m_i} = \sqrt{(b_i(\mathrm{g} - 1) + 1)^2\sigma_H^2 + \sigma_h^2 + \sigma_i^2} \quad (11)$$

The point of subjective equality (PSE) occurs when $\mu_{m_i} = 0$. At this point $b_i = 1/(1 - g)$. If there is no bias (i.e. b_i = 1), then the PSE occurs when g = 0 i.e. the stimulus is stationary in the speaker/LED ring. This makes sense because no bias means that sensed image motion i for a stationary stimulus is equal and opposite to the sensed head movement h (recall that for our visual stimuli, we provide a head-centred fixation point to inhibit eye rotation in the skull). If b_i > 1, then the PSE occurs when g > 1. This describes the head-movement equivalent of a Filehne illusion (Filehne, 1922; Freeman, 2007; Haarmeier & Thier, 1996; Mack & Herman, 1973), in which stationary objects appear to move opposite to a smooth eye pursuit. In order to null the Filehne illusion, the stimulus therefore needs to move in the same direction as the pursuit, equivalent to a positive motion gain. For b_i < 1, the opposite is true.

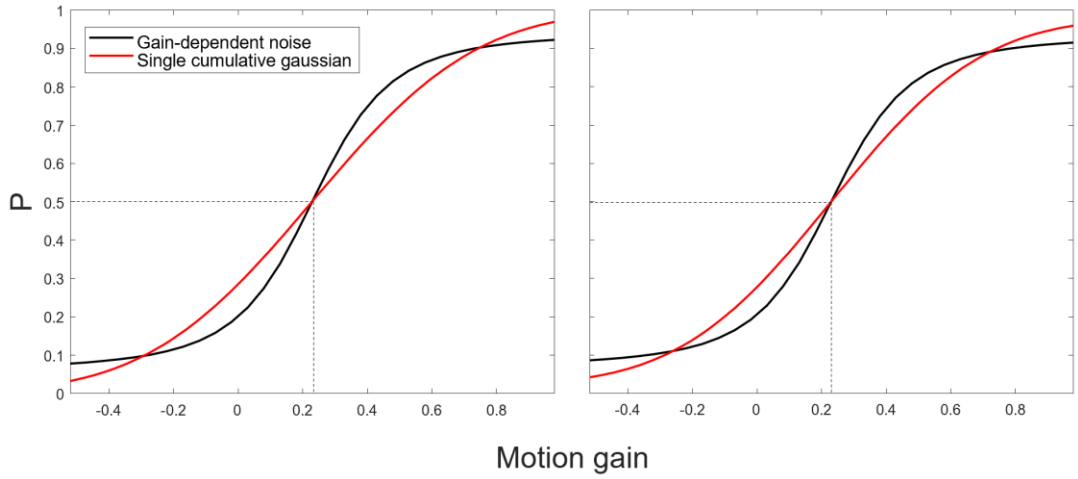

Figure 1: Left: The black curve shows a psychometric function based on gain-dependent noise with $\{\mu_H, \sigma_H^2, \sigma_h^2, \sigma_i^2, b_i, \lambda\} = \{10, 40, 6, 3, 1.3, 0\}$ (note that these parameters do not reflect the typical values used in this experiment, but were deliberately chosen to demonstrate the difference between the standard approach and the present approach). The red curve is the best fitting cumulative Gaussian as determined by the Palamedes toolbox, again with $\lambda = 0$. Right: This time with lapse-rate $\lambda = 0.02$ for the gain-dependent noise psychometric function, and constrained to vary no larger than 0.02 for the single cumulative Gaussian (Prins, 2012).

If the head movement did not vary across trials (σ_H^2 = 0), then the sum $\sigma_h^2 + \sigma_i^2$ in Eqn (11) could easily be recovered from the best-fitting cumulative Gaussian as per standard fitting of psychometric functions. However, $\sigma_H^2 \neq 0$. Variable head movements make the recovery of $\sigma_h^2 + \sigma_i^2$ more complicated because they act as an external source of 'gain-dependent' noise that varies with motion gain across the psychometric function. Fitting a single cumulative Gaussian is an approximation at best, as demonstrated in Figure 1. The black curves show example psychometric functions based on the formulae above and parameter values given in the legend; the red curves the best-fitting single cumulative Gaussian. The difference between the two panels is whether lapse rates are included or not. Including gain-dependent noise has two main effects: (1) the asymptotes of the psychometric function move away from P= 0 and 1; (2) the slope becomes steeper and is not well fit by a single cumulative Gaussian. The degree to which the gain-dependent noise causes substantial departures from the standard fit

depends on the relationship between the values of $\mu_H, \sigma_H^2, \sigma_h^2, \sigma_i^2, b_i$ and whether lapse-rate is allowed to vary in the standard fit.

Fitting procedure

To tackle these issues, we fit psychometric functions to our unimodal condition data based on the formulae above, using the measured head movements to estimate the mean and standard deviation of H. The latter were obtained by fitting a Gaussian distribution to the histogram of these movements for the trials making up that psychometric functions. We fixed the variance of the head-movement signal σ_h^2 in a separate experiment (see Appendix A) because this allowed us to estimate the correlation between cues when evaluating a correlated cue-combination model as discussed below. We note that for the empirical parameter values intrinsic to our observers we did not find much difference between fitting a gain-dependent noise psychometric function and a standard single cumulative Gaussian. One likely explanation for this similarity was that the head movements were relatively consistent (σ_H^2 low) given the repetitive nature of the task. But also, we allowed lapse rate to be a (constrained) free parameter when fitting a single cumulative Gaussian ($\lambda \leq 0.02$). As shown in Figure 1B, lapse rate can mimic the asymptotic behaviour of the gain-dependent psychometric function, albeit for the wrong reasons.

Bimodal conditions

As explained in the main text, quantitative predictions for bimodal performance start with the idea that cue combination is a weighted average of individual cues, where the weights are the reciprocal of each cue's precision. When precision is expressed as the variance, the result is a maximum likelihood estimate (see Ernst & Banks, 2002). During self-motion, cue combination is made more complicated because different modalities reside in different coordinate frames: cues must be 'promoted' into a common reference frame (Landy et al.,

1995). Here we hypothesise that for vision and hearing the common reference frame is body-centred. In the case of our experiment, this means that both modalities use a head rotation signal to interpret image motion, and that combination occurs beyond this 'compensation' process. In other words, the perceptual system combines body-centred cues. To investigate, we develop two models, one where body-centred cues are combined using standard MLE principles, and one in which the shared head-rotation signal is taken into account. The first model treats the body-centred cues as independent, whereas the second treats them as partially correlated and follows the logic detailed in Oruç et al. (2003). In both cases, the gain-dependent noise presents an obstacle because, as discussed above, the resulting psychometric functions on which the weights are based are not a single cumulative Gaussian. For this reason, we follow the logic above and define the models at the level of the psychometric function as follows.

Combining uncorrelated body-centred cues

From Eqn (11), the variance of the body-centred audio and visual cues is given by:

$$\sigma_{i_{bc}}^2 = (b_i(\mathrm{g}-1)+1)^2\sigma_H^2 + \sigma_h^2 + \sigma_i^2 \quad (12)$$

for i = a and v. Defining reliability as $r_{i_{bc}} = 1/\sigma_{i_{bc}}^2$, then the weights are:

$$w_{a_{bc}} = r_{a_{bc}}/\left(r_{a_{bc}} + r_{v_{bc}}\right) \quad (13)$$

$$w_{v_{bc}} = r_{v_{bc}}/\left(r_{a_{bc}} + r_{v_{bc}}\right) \quad (14)$$

Following the logic of the single-cue condition, we need first to determine the decision variable $d_{av_{bc}} = \mu_{av_{bc}}/\sigma_{av_{bc}}$ in order to construct the psychometric function. This time, however, the decision variable will be based on the sum of the weighted body-centred distributions for audio and visual cues:

$$d_{av_{bc}} = w_{a_{bc}}\mathrm{N}\left(\mu_{a_{bc}}, \sigma_{a_{bc}}\right) + w_{v_{bc}}\mathrm{N}\left(\mu_{v_{bc}}, \sigma_{v_{bc}}\right) \quad (15)$$

Given that Eqn (10) defines the mean for a single cue, and Eqn (12) defines its variance, then:

$$\mu_{av_{bc}} = \mu_H \,.\, [w_{a_{bc}}(b_a(\mathrm{g}-1)+1) + w_{v_{bc}}(b_v(\mathrm{g}-1)+1)] \quad (16)$$

$$\sigma_{av_{bc}} = \sqrt{w_{a_{bc}}^2\sigma_{a_{bc}}^2 + w_{v_{bc}}^2\sigma_{v_{bc}}^2} \quad (17)$$

The psychometric function is then derived using Eqn (9).

Combining correlated body-centred cues

Following Oruç et al. (2003), the reliabilities of correlated cues must be corrected for the shared noise. Specifically:

$$r'_{a_{bc}} = r_{a_{bc}} - \rho\sqrt{r_{a_{bc}} + r_{v_{bc}}} \quad (18)$$

$$r'_{v_{bc}} = r_{v_{bc}} - \rho\sqrt{r_{a_{bc}} + r_{v_{bc}}} \quad (19)$$

These define a new wet of weights $w'_{a_{bc}}$ and $w'_{v_{bc}}$ using the same logic as Eqns (13) and (14). The decision variable $d_{av_{bc}}$ then has mean and standard deviation:

$$\mu'_{av_{bc}} = \mu_H \,.\, [w'_{a_{bc}}(b_a(\mathrm{g}-1)+1) + w'_{v_{bc}}(b_v(\mathrm{g}-1)+1)] \quad (20)$$

$$\sigma'_{av_{bc}} = \sqrt{w'^2_{a_{bc}}\sigma_{a_{bc}}^2 + w'^2_{v_{bc}}\sigma_{v_{bc}}^2 - 2\rho w'_{a_{bc}}w'_{v_{bc}} \Big/ \sqrt{r_{a_{bc}} + r_{v_{bc}}}} \quad (21)$$

where Eqn (21) is based on Eqn (6) of Oruc et al.

Fitting procedure and model evaluation

The parameters defining the psychometric functions for the correlated and uncorrelated cue combination models are all fixed by the unimodal single-cue conditions, apart from the correlation ρ and the ubiquitous lapse rate λ. The correlation could be made free to

vary when fitting the correlated cue combination model to the bimodal data. However, we have recently developed a technique to measure the variance of the head movement signal σ_h^2, which therefore allows us to fix the correlation as described below. Models are then evaluated by comparing goodness-of fit measures at the level of the psychometric functions, as opposed to the normal route which is based on PSEs and precision measures (e.g. thresholds, slopes). To reiterate, the reason this standard procedure could fail is because the use of a single cumulative Gaussian with an accompanying lapse rate is at best an approximation.

To obtain the correlation, recall from Eqn (7) that the body-centred audio and visual cues (m_a and m_v) share exact copies of the head-movement signal (h) but scaled copies of actual head movements (H). The scale is determined by their biases and the motion gain. Noting that the correlation between m_a and m_v is their covariance divided the square-root of the product of their variances, it can be shown that:

$$\rho = \left(k_a k_v \sigma_H^2 + \sigma_h^2\right) \Big/ \sqrt{(\sigma_a^2 + k_a^2\sigma_H^2 + \sigma_h^2)(\sigma_v^2 + k_v^2\sigma_H^2 + \sigma_h^2)} \qquad (22)$$

where $k_i = b_i(\mathrm{g} - 1) + 1$ for $i = a$ and v .

References

Alais, D., & Burr, D. (2004a). No direction-specific bimodal facilitation for audiovisual motion detection. *Cognitive Brain Research*, *19*(2), 185–194. https://doi.org/10.1016/j.cogbrainres.2003.11.011

Alais, D., & Burr, D. (2004b). The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Current Biology*, *14*(3), 257–262. https://doi.org/10.1016/j.cub.2004.01.029

Andersen, R. A., Snyder, L. H., Li, C.-S., & Stricanne, B. (1993). Coordinate transformations in the representation of spatial information. *Current Opinion in Neurobiology*, *3*(2), 171–176. https://doi.org/10.1016/0959-4388(93)90206-E

Angelaki, D. E., & Cullen, K. E. (2008). Vestibular System: The Many Facets of a Multimodal Sense. *Annual Review of Neuroscience*, *31*(1), 125–150. https://doi.org/10.1146/annurev.neuro.31.060407.125555

Aubert, H. (1886). Die Bewegungsempfindung. *Pflüger, Archiv für die Gesammte Physiologie des Menschen und der Thiere*, *39*(1), 347–370. https://doi.org/10.1007/BF01612166

Bentvelzen, A., Leung, J., & Alais, D. (2009). Discriminating Audiovisual Speed: Optimal Integration of Speed Defaults to Probability Summation When Component Reliabilities Diverge. *Perception*, *38*(7), 966–987. https://doi.org/10.1068/p6261

Bogadhi, A. R., Montagnini, A., & Masson, G. S. (2013). Dynamic interaction between retinal and extraretinal signals in motion integration for smooth pursuit. *Journal of Vision*, *13*(13), 5–5. https://doi.org/10.1167/13.13.5

Bolognini, N., Leor, F., Passamonti, C., Stein, B. E., & Làdavas, E. (2007). Multisensory-Mediated Auditory Localization. *Perception*, *36*(10), 1477–1485. https://doi.org/10.1068/p5846

Bulkin, D. A., & Groh, J. M. (2006). Seeing sounds: Visual and auditory interactions in the brain. *Current Opinion in Neurobiology*, *16*(4), 415–419. https://doi.org/10.1016/j.conb.2006.06.008

Burge, J., Girshick, A. R., & Banks, M. S. (2010). Visual–Haptic Adaptation Is Determined by Relative Reliability. *Journal of Neuroscience*, *30*(22), 7714–7721. https://doi.org/10.1523/JNEUROSCI.6427-09.2010

Burns, J. K., & Blohm, G. (2010). Multi-Sensory Weights Depend on Contextual Noise in Reference Frame Transformations. *Frontiers in Human Neuroscience*, *4*. https://doi.org/10.3389/fnhum.2010.00221

Carlile, S., & Best, V. (2002). Discrimination of sound source velocity in human listeners. *The Journal of the Acoustical Society of America*, *111*(2), 1026–1035. https://doi.org/10.1121/1.1436067

Carriot, J., Bryan, A., DiZio, P., & Lackner, J. R. (2011). The oculogyral illusion: Retinal and oculomotor factors. *Experimental Brain Research*, *209*(3), 415–423. https://doi.org/10.1007/s00221-011-2567-5

Cohen, Y. E., & Andersen, R. A. (2002). A common reference frame for movement plans in the posterior parietal cortex. *Nature Reviews Neuroscience*, *3*(7), 553–562. https://doi.org/10.1038/nrn873

Collins, T., Heed, T., & Röder, B. (2010). Eye-movement-driven changes in the perception of auditory space. *Attention, Perception, & Psychophysics*, *72*(3), 736–746. https://doi.org/10.3758/APP.72.3.736

Cooper, J., Carlile, S., & Alais, D. (2008). Distortions of auditory space during rapid head turns. *Experimental Brain Research*, *191*(2), 209–219. https://doi.org/10.1007/s00221-008-1516-4

Cullen, K. E. (2019). Vestibular processing during natural self-motion: Implications for perception and action. *Nature Reviews Neuroscience*, *20*(6), 346–363. https://doi.org/10.1038/s41583-019-0153-1

Dichgans, J., Wist, E., Diener, H. C., & Brandt, Th. (1975). The Aubert-Fleischl phenomenon: A temporal frequency effect on perceived velocity in afferent motion perception. *Experimental Brain Research*, *23*(5). https://doi.org/10.1007/BF00234920

Dokka, K., MacNeilage, P. R., DeAngelis, G. C., & Angelaki, D. E. (2015). Multisensory Self-Motion Compensation During Object Trajectory Judgments. *Cerebral Cortex*, *25*(3), 619–630. https://doi.org/10.1093/cercor/bht247

Ernst, M. O., & Banks, M. S. (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 Sciences*, *8*(4), 162–169. https://doi.org/10.1016/j.tics.2004.02.002

Filehne, W. (1922). Über das optische Wahrnehmen von Bewegungen. *Zeitschrift Für Sinnesphysiology*, *53*, 134.

Freeman, T. C. A. (2001). Transducer models of head-centred motion perception. *Vision Research*, *41*(21), 2741–2755. https://doi.org/10.1016/S0042-6989(01)00159-6

Freeman, T. C. A. (2007). Simultaneous adaptation of retinal and extra-retinal motion signals. *Vision Research*, *47*(27), 3373–3384. https://doi.org/10.1016/j.visres.2007.10.002

Freeman, T. C. A., & Banks, M. S. (1998). Perceived head-centric speed is affected by both extra-retinal and retinal errors. *Vision Research*, *38*(7), 941–945. https://doi.org/10.1016/S0042-6989(97)00395-7

Freeman, T. C. A., Cucu, M. O., & Smith, L. (2018). A preference for visual speed during smooth pursuit eye movement. *Journal of Experimental Psychology. Human Perception and Performance*, *44*(10), 1629–1636. https://doi.org/10.1037/xhp0000551

Freeman, T. C. A., Culling, J. F., Akeroyd, M. A., & Brimijoin, W. O. (2017). Auditory compensation for head rotation is incomplete. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(2), 371–380. https://doi.org/10.1037/xhp0000321

Freeman, T. C. A., Leung, J., Wufong, E., Orchard-Mills, E., Carlile, S., & Alais, D. (2014). Discrimination Contours for Moving Sounds Reveal Duration and Distance Cues Dominate Auditory Speed Perception. *PLOS ONE*, *9*(7), e102864. https://doi.org/10.1371/journal.pone.0102864

Furman, M., & Gur, M. (2012). And yet it moves: Perceptual illusions and neural mechanisms of pursuit compensation during smooth pursuit eye movements. *Neuroscience & Biobehavioral Reviews*, *36*(1), 143–151. https://doi.org/10.1016/j.neubiorev.2011.05.005

Garzorz, I. T., Freeman, T. C. A., Ernst, M. O., & MacNeilage, P. R. (2018). Insufficient compensation for self-motion during perception of object speed: The vestibular Aubert-Fleischl phenomenon. *Journal of Vision*, *18*(13), 9. https://doi.org/10.1167/18.13.9

Genzel, D., Firzlaff, U., Wiegrebe, L., & MacNeilage, P. R. (2016). Dependence of auditory spatial updating on vestibular, proprioceptive, and efference copy signals. *Journal of Neurophysiology*, *116*(2), 765–775. https://doi.org/10.1152/jn.00052.2016

Girshick, A. R., Landy, M. S., & Simoncelli, E. P. (2011). Cardinal rules: Visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, *14*(7), 926–932. https://doi.org/10.1038/nn.2831

Goossens, H. H. L. M., & van Opstal, A. J. (1999). Influence of Head Position on the Spatial Representation of Acoustic Targets. *Journal of Neurophysiology*, *81*(6), 2720–2736. https://doi.org/10.1152/jn.1999.81.6.2720

Gu, Y., Angelaki, D. E., & DeAngelis, G. C. (2008). Neural correlates of multisensory cue integration in macaque MSTd. *Nature Neuroscience*, *11*(10), 1201–1210. https://doi.org/10.1038/nn.2191

Haarmeier, T., & Thier, P. (1996). Modification of the filehne illusion by conditioning visual stimuli. *Vision Research*, *36*(5), 741–750. https://doi.org/10.1016/0042-6989(95)00154-9

Hairston, W. D., Laurienti, P. J., Mishra, G., Burdette, J. H., & Wallace, M. T. (2003). Multisensory enhancement of localization under conditions of induced myopia. *Experimental Brain Research*, *152*(3), 404–408. https://doi.org/10.1007/s00221-003-1646-7

Halow, S., Liu, J., Folmer, E., & MacNeilage, P. R. (2023). *Motor Signals Mediate Stationarity Perception*. https://doi.org/10.1163/22134808-bja10111

Harris, L. R. (1994). Visual Motion Caused by Movements of the Eye, Head and Body. In A. T. Smith & R. J. Snowden (Eds.), *Visual detection of motion* (pp. 397–435). Academic Press.

Haynes, J. D., Gallagher, M., Culling, J. F., & Freeman, T. C. A. (2024). The precision of signals encoding active self-movement. *Journal of Neurophysiology*. https://doi.org/10.1152/jn.00370.2023

Ilg, U. J., Schumann, S., & Thier, P. (2004). Posterior Parietal Cortex Neurons Encode Target Motion in World-Centered Coordinates. *Neuron*, *43*(1), 145–151. https://doi.org/10.1016/j.neuron.2004.06.006

Ilg, U. J., & Thier, P. (1996). Inability of rhesus monkey area V1 to discriminate between self-induced and externally induced retinal image slip. *The European Journal of Neuroscience*, *8*(6). https://doi.org/10.1111/j.1460-9568.1996.tb01283.x

Jones, P. R. (2016). A tutorial on cue combination and Signal Detection Theory: Using changes in sensitivity to evaluate how observers integrate sensory information. *Journal of Mathematical Psychology*, *73*, 117–139. https://doi.org/10.1016/j.jmp.2016.04.006

Kopinska, A., & Harris, L. R. (2003). Spatial representation in body coordinates: Evidence from errors in remembering positions of visual and auditory targets after active eye, head, and body movements. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *57*(1), 23–37. https://doi.org/10.1037/h0087410

Lackner, J. R., & DiZio, P. (2010). Audiogravic and oculogravic illusions represent a unified spatial remapping. *Experimental Brain Research*, *202*(2), 513–518. https://doi.org/10.1007/s00221-009-2149-y

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research*, *35*(3), 389–412. https://doi.org/10.1016/0042-6989(94)00176-M

Lawrence, M. A. (2016). *ez: Easy Analysis and Visualization of Factorial Experiments* (Version 4.4-0) [Computer software]. https://cran.r-project.org/web/packages/ez/index.html

Lee, J., & Groh, J. M. (2012). Auditory signals evolve from hybrid- to eye-centered coordinates in the primate superior colliculus. *Journal of Neurophysiology*, *108*(1), 227–242. https://doi.org/10.1152/jn.00706.2011

Leigh, R. J., & Zee, D. S. (2015). *The Neurology of Eye Movements*. Oxford University Press.

Lewald, J., Dörrscheidt, G. J., & Ehrenstein, W. H. (2000). Sound localization with eccentric head position. *Behavioural Brain Research*, *108*(2), 105–125. https://doi.org/10.1016/S0166-4328(99)00141-2

Lewald, J., & Karnath, H.-O. (2000). Vestibular Influence on Human Auditory Space Perception. *Journal of Neurophysiology*, *84*(2), 1107–1111. https://doi.org/10.1152/jn.2000.84.2.1107

Lewald, J., Karnath, H.-O., & Ehrenstein, W. H. (1999). Neck-proprioceptive influence on auditory lateralization. *Experimental Brain Research*, *125*(4), 389–396. https://doi.org/10.1007/s002210050695

Mack, A., & Herman, E. (1973). Position Constancy during Pursuit Eye Movement: An Investigation of the Filehne Illusion. *Quarterly Journal of Experimental Psychology*, *25*(1), 71–84. https://doi.org/10.1080/14640747308400324

MATLAB version 9.12.0.1884302 (R2022a). (2022). The Mathworks, Inc.

Meyer, G. F., & Wuerger, S. M. (2001). Cross-modal integration of auditory and visual motion signals: *Neuroreport*, *12*(11), 2557–2560. https://doi.org/10.1097/00001756-200108080-00053

Newsome, W. T., Wurtz, R. H., & Komatsu, H. (1988). Relation of cortical areas MT and MST to pursuit eye movements. II. Differentiation of retinal from extraretinal inputs. *Journal of Neurophysiology*, *60*(2), 604–620. https://doi.org/10.1152/jn.1988.60.2.604

Ono, S., & Mustari, M. J. (2012). Role of MSTd Extraretinal Signals in Smooth Pursuit Adaptation. *Cerebral Cortex*, *22*(5), 1139–1147. https://doi.org/10.1093/cercor/bhr188

Oruç, İ., Maloney, L. T., & Landy, M. S. (2003). Weighted linear cue combination with possibly correlated error. *Vision Research*, *43*(23), 2451–2468. https://doi.org/10.1016/S0042-6989(03)00435-8

Prins, N. (2012). The psychometric function: The lapse rate revisited. *Journal of Vision*, *12*(6), 25. https://doi.org/10.1167/12.6.25

Prins, N., & Kingdom, F. A. A. (2018). Applying the Model-Comparison Approach to Test Specific Research Hypotheses in Psychophysical Research Using the Palamedes Toolbox. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01250

R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria. https://www.R-project.org/

Reisbeck, T. E., & Gegenfurtner, K. R. (1999). Velocity tuned mechanisms in human motion processing. *Vision Research*, *39*(19), 3267–3286. https://doi.org/10.1016/S0042-6989(99)00017-6

Rohde, M., van Dam, L. C. J., & Ernst, M. O. (2016). Statistically Optimal Multisensory Cue Integration: A Practical Tutorial. *Multisensory Research*, *29*(4–5), 279–317. https://doi.org/10.1163/22134808-00002510

RStudio Team. (2019). *RStudio: Integrated Development for R.* [Computer software]. RStudio Inc.

Scarfe, P., & Hibbard, P. B. (2011). Statistically optimal integration of biased sensory estimates. *Journal of Vision*, *11*(7), 12–12. https://doi.org/10.1167/11.7.12

Sober, S. J., & Sabes, P. N. (2003). Multisensory Integration during Motor Planning. *Journal of Neuroscience*, *23*(18), 6982–6992. https://doi.org/10.1523/JNEUROSCI.23-18-06982.2003

Soto-Faraco, S., Lyons, J., Gazzaniga, M., Spence, C., & Kingstone, A. (2002). The ventriloquist in motion: Illusory capture of dynamic information across sensory modalities. *Cognitive Brain Research*, *14*(1), 139–146. https://doi.org/10.1016/S0926-6410(02)00068-X

Soto-Faraco, S., Spence, C., Lloyd, D., & Kingstone, A. (2004). Moving Multisensory Research Along: Motion Perception Across Sensory Modalities. *Current Directions in Psychological Science*, *13*(1). https://doi.org/10.1111/j.0963-7214.2004.01301008.x

Sperry, R. W. (1950). Neural basis of the spontaneous optokinetic response produced by visual inversion. *Journal of Comparative and Physiological Psychology*, *43*(6), 482–489. https://doi.org/10.1037/h0055479

Teramoto, W., Cui, Z., Sakamoto, S., & Gyoba, J. (2014). Distortion of auditory space during visually induced self-motion in depth. *Frontiers in Psychology*, *5*. https://www.frontiersin.org/articles/10.3389/fpsyg.2014.00848

Vliegen, J., Van Grootel, T. J., & Van Opstal, A. J. (2004). Dynamic Sound Localization during Rapid Eye-Head Gaze Shifts. *The Journal of Neuroscience*, *24*(42), 9291–9302. https://doi.org/10.1523/JNEUROSCI.2671-04.2004

von Holst, E., & Mittelstaedt, H. (1950). Das Reafferenzprinzip. *Naturwissenschaften*, *37*(20), 464–476. https://doi.org/10.1007/BF00622503

Wallach, H. (1987). Perceiving a Stable Envrionment When One Moves. *Annual Review of Psychology*, *38*, 1–27.

Wertheim, A. H. (1987). Retinal and Extraretinal Information in Movement Perception: How to Invert the Filehne Illusion. *Perception*, *16*(3), 299–308. https://doi.org/10.1068/p160299

Wertheim, A. H. (1994). Motion perception during selfmotion: The direct versus inferential controversy revisited. *Behavioral and Brain Sciences*, *17*(2), 293–311. https://doi.org/10.1017/S0140525X00034646

Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D., Software, P., & PBC. (2023). *dplyr: A Grammar of Data Manipulation* (Version 1.1.4) [Computer software]. https://cran.r-project.org/web/packages/dplyr/index.html

Wuerger, S., Meyer, G., Hofbauer, M., Zetzsche, C., & Schill, K. (2010). Motion extrapolation of auditory–visual targets. *Information Fusion*, *11*(1), 45–50. https://doi.org/10.1016/j.inffus.2009.04.005

Zhang, T., Heuer, H. W., & Britten, K. H. (2004). Parietal Area VIP Neuronal Responses to Heading Stimuli Are Encoded in Head-Centered Coordinates. *Neuron*, *42*(6), 993–1001. https://doi.org/10.1016/j.neuron.2004.06.008