Identifying audio-visual benefits in a manual localization task

Master Thesis Aerospace Engineering

Tessa Mennink

Identifying audiovisual benefits in a manual localization task

Master Thesis Aerospace Engineering

by

Tessa Mennink

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday August 29, 2023 at 13:00.

Student number:	4547462	
Project duration:	August 1, 2022 – August 29, 2023	
Thesis committee:	Prof. dr. ir. M. Mulder,	TU Delft, chair
	Dr. ir. D.M. Pool,	TU Delft, supervisor
	Dr. P. Bremen,	Erasmus Medical Center, supervisor
	Dr. ir. D. Dirkx,	TU Delft

Front page image obtained from
https://www.freshconsulting.com/insights/blog/human-machine-interaction-how-ai-is-revolutionizing-tech/,
accessed on 19/08/2023

TU Delft

Contents

List of Figures

List of Tables

List of Abbreviations

A	Audio
AV	Audiovisual
CDF	Cumulative Distribution Function
HRTF	Head-Related Transfer Function
HSE	Head Shadow Effect
ILDs	Interaural Level Differences
IQR	interquartile range
ITDs	Interaural Time Differences
MAE	Mean Absolute Error
MLE	Maximum Likelihood Estimation
RT	Reaction Time
V	Visual

Acknowledgements

I hereby present my final thesis project for the completion of my master's degree at the Faculty of Aerospace Engineering, Delft University of Technology. As I approach the end of my time as a student at the Faculty of Aerospace Engineering, a blend of emotions courses through me. There is a sense of sadness as this chapter comes to an end. The faculty has not only been a source of academic enrichment but I have also taken great pleasure in engaging with various other aspects of campus life. However, I am excited to start a new adventure and am curious about the new challenges to come.

During my thesis, I got the opportunity to perform my research in collaboration with the Neuroscience department at the Erasmus Medical Center. Neuroscience has always been an interest of mine and I enjoyed the opportunity of applying my skills as an engineer to this new research field. Moreover, I discovered that the fields of human-machine interaction and behavioral neuroscience have more in common than one might think at first glance. In my research, I tried to combine knowledge gained in both fields, hopefully being an encouragement for more combined work in the future. Before I encourage you to dive into the details of my work I would like to express my gratitude to some individuals who played a crucial role in this project.

Firstly, I would like to thank Peter for sharing part of his elaborate knowledge with me. Given my aerospace background, my familiarity with your field was limited, yet you consistently dedicated time to answer my questions. Your extensive knowledge, eye for detail, and commitment to explaining the measured data repeatedly astonished me. Moreover, I would like to thank Johan for the help you offered in structuring my ideas. Moreover, I appreciated the personal attention and specifically your positivism in the end phase of my thesis. Lastly, I would like to thank Daan for offering me all the freedom to define my own research whilst also helping me structure and define boundaries. I valued your genuine interest in me as a person and in my work, but also the short communication line we had resulting in quick problem-solving. Furthermore, I want to thank all my supervisors for the flexibility you offered me during the final phase, which enabled me to complete this research. Recognizing that there were likely simpler approaches, I value your efforts even more.

I would also like to express my gratitude to my friends, with whom I shared the journey of studying, working, and creating cherished memories throughout my time as a student in Delft. Moreover, I want to thank Mathijs, for his unwavering support and listening ear during difficult moments, allowing me to vent and gain perspective. Lastly, but by no means least, I would like to express my appreciation to my family. Throughout my entire student journey, your presence, unwavering motivation, and nurturing environment provided me with the strength to keep moving forward.

Tessa Mennink
Rotterdam, August 2023

I

Scientific Paper

Identifying audiovisual benefits in discrete head and arm localization tasks within an echoic environment

*

Tessa Mennink (MSc student)
Supervisors: Dr. Ir. D.M. Pool*, Dr. Ir. J.J.M. Pel † and Dr. P. Bremen †
* *Dept. of Control & Simulation, Faculty of Aerospace Engineering, Delft University of Technology, Delft, the Netherlands*
† *Dept. of Neuroscience, Erasmus MC, Rotterdam, the Netherlands*

***Abstract*—Although the roles of visual and haptic cues in motor tasks have been well studied, the benefits of audio cues in complex motor tasks have been underexplored. Audiovisual cues reduce reaction times and decrease the variance of endpoint responses in simple head-orienting localization tasks. The aim of this research is to explore the potential of audiovisual stimuli for motor tasks performed in a real-life environment. Participants performed simple head and arm localization tasks in an echoic room. Comparing audiovisual stimuli to unimodal stimuli, a decrease in reaction times was observed. For head responses, reaction times were reduced on average by 58 ms for visual stimuli and 35 ms for auditory stimuli. For arm responses, reaction times showed a decrease of 30 ms on average for both unimodal conditions. Likewise, we noted decreased endpoint variance in audiovisual responses in comparison to unimodal auditory responses. This reduction was observed in both head responses ($50^{\circ 2}$) and arm responses ($44^{\circ 2}$). Moreover, we observed a modality-effector dependency in the weighting of auditory and visual signals in the audiovisual responses. These results show the potential of audiovisual stimuli for real-life applications in motor tasks.**

***Index Terms*—Manual control, audiovisual integration, sound localization, human-machine interaction, audiovisual benefits**

Nomenclature

α	Target angle
α_{max}	Maximum target angle
β	Bias
γ	Gain
$\hat{S}_{AV}$	Audiovisual percept
$\hat{S}_{A}$	Auditory percept
$\hat{S}_{V}$	Visual percept
μ	Mean
σ^2	Variance
$c_{(1,2)}$	Input/output variables linear fit
F_{AV}	CDF audiovisual responses
F_{A}	CDF visual responses
F_{race}	Race model prediction
F_{V}	CDF auditory responses
I_{L_α}	Scale factor left speaker for target α
I_{R_α}	Scale factor right speaker for target α
$I_{rel\alpha}$	Relative position vector
L_{AV}	Visual response characteristics
L_{A}	Auditory response characteristics
L_{V}	Audiovisual response characteristics
R^2	Coefficient of determination
$w_{A_{MLE}}$	Visual MLE weight
$w_{V_{EMP}}$	Visual empirical weight
$w_{V_{MLE}}$	Auditory MLE weight
I_{rel_α}	Relative volume for specific target α
t	Time

I. Introduction

In our natural environment, our sensory systems continuously encounter noisy and uncertain signals, presenting a challenge for the brain to construct an accurate representation of the world. While the roles of visual and haptic cues in motor tasks have been extensively studied [1]–[3], the potential advantages of audio cues in complex motor tasks have received limited attention, despite their relevance in human-machine interaction scenarios, as the interaction between audio and visual processing can lead to enhanced performance in simple localization tasks [4]–[6]. Earlier studies have demonstrated that incorporating additional auditory cues into continuous tracking tasks, which are predominantly influenced by visual stimuli [7], can lead to a reduction in response latency and an improvement in tracking accuracy [8], [9]. More recent research has extended its focus to real-world scenarios, including applications such as assisting in blind driving, facilitating rehabilitation therapy, and enhancing performance in complex motor tasks [10]–[13]. In the field of aerospace engineering, the inclusion of supplementary auditory cues enhances situational awareness, reduces cognitive load, and reduces reaction time [14]–[16]. Directional cues have been shown to be useful in providing supplementary spatial localization information. This benefit is especially obvious in fighter jet pilots, for whom visual cues can lack reliability and contribute to spatial disorientation [14], [17]. These are all promising outcomes, however, understanding the fundamental mechanisms responsible for these performance enhancements necessitate further research. In this research, we tested whether simulated sound sources result in accurate sound localization. Moreover, we researched audiovisual benefits during both head and arm localization tasks, within an environment where sound wave reflections were present, i.e. an echoic space.

A. *Principles of spatial hearing*

Whereas the visual system employs a straightforward one-to-one mapping for object localization, the auditory system relies on input from both ears to create a spatial percept of the sound source [18]. Humans exhibit the capacity to precisely and reliably determine the location of sounds emanating from a stationary source [18]–[20].

Spatial information for sound localization is derived from the spectral, temporal, and intensity characteristics of the incoming sound signal [18], [19], [21]. In the context of sound localization, two distinct directional aspects are taken into account: the azimuth angle and elevation angle [18]. The azimuth angle pertains to the horizontal plane passing through the head (left/right orientation), while the elevation angle corresponds to the vertical plane (up/down orientation) [18]. Sound localization in the horizontal direction (azimuth) relies primarily on interaural time differences (ITDs) and interaural level differences (ILDs). ITDs arise due to the microsecond-level time gap between the arrival of a sound at each ear for locations away from the centerline [18]. Moreover, the presence of the head introduces an acoustic shadow, referred to as the Head Shadow Effect (HSE), resulting in ILDs, wherein the sound's intensity diminishes at the ear farther from the sound source in comparison to the ear closer to it [18], [22]. For frequencies exceeding approximately 2-3 kHz, where sound wavelengths are shorter than the dimensions of the head, sound localization is primarily influenced by interaural level differences (ILDs), whilst for sounds below 1.5 kHz ITDs are dominant [19], [23]. Even when conflicting ITDs are introduced, sounds within the 4-16 kHz range can be accurately localized using ILDs [24]. This principle was subsequently verified in the context of broadband noise [25].

For vertical localization (elevation), additional spatial cues are obtained caused by sound wave reflection and diffraction on the body, particularly the pinnae – a part of the outer ear [19]. The pinnae function as filters that selectively enhance or reduce frequencies based on the elevation angle, resulting in a distinct filter for each elevation angle [18], [19]. This filter is also known as the head-related transfer function (HRTF) [26]. Due to the similar shapes of the pinnae for both ears, elevation localization primarily relies on a single ear (monaural) approach [18].

The majority of sound localization experiments are conducted under what is termed the "free-field condition". In this environment, sound propagation is uniform, making it optimal for conducting experiments involving acoustic manipulation [27]. A space that attains the "free-field" condition is termed an anechoic room. However, in real-world scenarios, such as an aircraft cockpit, sound waves interact with surfaces like walls, ceilings, and floors, resulting in reflections and the formation of a reverberant field [27]. In the context of the present study, this condition will be termed an "echoic room". Additionally, real-world scenarios introduce specific constraints, like a limited number of available sound sources. In this research, we aimed to investigate the potential of audio-visual benefits in a real-world environment hence experiments were performed in an echoic room. By linearly adjusting the volume of two speakers, ILDs were simulated for the desired angle of the sound source [28].

B. *Audiovisual benefits*

Extensive research has been performed on the neural pathways and principles underlying audiovisual benefits for simple motor responses [4]–[6]. Sensory processing involves two types of stimuli: unimodal (single sensory input) and multimodal (inputs from different sensory channels). Meredith and Stein studied audiovisual benefits by examining neural firing rates as a measure of multisensory integration [6], [29], [30]. They introduced the principle of superadditivity, indicating that multisensory signals generate responses greater than the sum of individual unisensory responses [6]. Multisensory integration occurs when stimulus properties facilitate the merging of information from two modalities, resulting in the perception of a unified object. Conversely, if stimuli properties favor the separation of sensory information into distinct objects, response enhancement is minimal or absent. The percept of a unified object is influenced by temporal and spatial proximity of events [29]. The strength of stimuli also determines the degree of response enhancement, where multimodal stimuli presented near the threshold value show the most enhancement [6].

If the brain assumes a common source of the sensory events, humans weigh both unimodal percepts based on their reliability to create the multimodal percept [31], [32]. Reliability is inversely related to a signal's variance and higher weights are assigned to the most reliable stimulus. Reliability-weighted multisensory integration results in the most precise and statistically optimal perceptual estimate, known as the maximum likelihood estimate (MLE) [31], [32] . Consequently, this integration enhances performance in spatial discrimination tasks [31]. For the common source assumption, perception plays a vital role, as stimuli do not need to be spatially close in physical terms, but must be perceived as such by the brain to result in multisensory benefit [33], [34]. To demonstrate multisensory integration, it is essential to explore various experimental conditions, including situations where response enhancement is diminished due to spatial and temporal misalignment of signals [35], [36].

C. *Sensory motor processing*

Localization tasks in humans can be described as a simple input-output model including three main steps; the processing of the sensory cues, the spatial perception created from this, and the motor response executed to a specific location [18]. The inputs to the system are the stimuli, e.g. a sound or a light flash. The outputs of the system are motor responses, e.g. head or eye movements or a simple button press. The type of motor output (eye/head/arm) is referred to as the effector.

Measuring multisensory processing in humans can be challenging due to the complex nature of the mechanisms involved. Nevertheless, several methods have been developed to study the interplay between sensory processing and motor output generation, as discussed in a review by Stevenson [37]. One common approach involves measuring natural gaze-orienting behavior in a localization task [38]. Gaze, representing the eye-in-space position, is recorded as participants perform coordinated eye-head movements toward target stimuli. Parameters like response latency and accuracy can be derived from the movement trajectory, serving as metrics to characterize human localization behavior [37].

The use of gaze for characterizing sensorimotor processing is motivated by the initial encoding of visual information in an oculocentric, eye-centered reference frame, while auditory information is encoded in a craniocentric, head-centered reference frame [18]. These differences arise from the reconstruction of sound source location using binaural and monaural cues related to the distance between ears and the filter characteristics of the torso, head, and pinnae [18]. Generally, two types of reference frames are considered: egocentric (relative to the observer's body) and allocentric (relative to objects in the world regardless of the observer's position) [18]. Achieving alignment between the egocentric, including the craniocentric and oculocentric reference frames, and the allocentric reference frame is crucial for programming goal-oriented eye-head movements towards various stimuli, despite the introduction of latency and noise due to the coordinate transformations [18].

In research focusing on smooth pursuit tasks, it was discovered that humans cannot utilize smooth pursuit eye movements to track moving sounds, as they do with moving visual targets [39]. Nevertheless, continuous tracking of moving sounds is achievable through head movements [20]. When comparing auditory and visual stimuli, it was observed that auditory stimuli trigger lower eye-head latency than visual stimuli [38]. These findings indicate that employing a modality-matched effector, e.g. using eye movements for visual targets or head movements for auditory targets, is more favorable for goal-directed orienting. In terms of velocity, head movements exhibit lower peak velocities compared to eye movements, regardless of the stimulus modality [38]. Additionally, consistent peak velocities are observed for head movements with different unimodal stimuli [38]. This suggests that, for head movements, motor commands and response properties, such as peak velocity, are independent of the target's modality and are solely associated with the specific effector used.

Studies centered on localization tasks for arm movements often involve pointing tasks, wherein participants are asked to indicate visual or auditory targets through pointing gestures [40], [41]. However, there exists research that delves into localization tasks utilizing a joystick, implying potential advantages in terms of reaction time [4]. Arm positioning is encoded within body-centric coordinates, necessitating the transformation of targets initially encoded in oculocentric or craniocentric coordinates into the appropriate reference frame for the programming of accurate orienting responses. Moreover, the motor commands necessary for executing arm movements are notably more complex when compared to eye or head movements, which can consequently lead to prolonged reaction times and increased response variability [42]. Despite the mechanisms underlying the creation of motor plans for eye, head, and arm movements remaining somewhat uncertain, it has been established that accuracy in arm pointing is diminished under head-restrained conditions [43], [44]. This outcome is likely attributed to the inhibition of the natural coordination among eye, head, and arm movements [44]. Additionally, in tasks involving eye-hand coordination, the eye has been observed to take precedence over the hand in movement initiation [45]. These observations align with the modality-matched effector hypotheses presented.

To explore the potential of audiovisual benefits in motor tasks applicable to real-life scenarios, we performed a localization task within an echoic room utilizing simulated sound sources. This study includes both a head-localization and an arm-localization tasks, enabling us to assess effector-related dependencies in localization performance. In the arm-localization task, participants maneuvered a joystick towards specified target locations. This decision is supported by the common practice of using manual control tasks to assess motor performance [1], [3], [46]. Additionally, by measuring head responses, we can attain a more comprehensive understanding of the influence of the echoic room, given that prior head-localization experiments were conducted in an anechoic room [28]. With the goal of gaining a deeper understanding of the practical opportunities presented by audiovisual stimuli, we utilized two sound sources to simulate multiple target locations, a configuration that could be feasible in environments such as cars and cockpits. A more elaborated explanation of the setup is provided in section II.

II. Materials and Methods

A. Participants

A total of 12 individuals took part in our experiment. Their ages ranged from 22 to 43 years (median = 25 years), including 4 female and 8 male participants, and all were right-handed. All participants provided written informed consents, and the conducted experiments received approval from the ethical committee of Delft University of Technology (Application number: 3222).

B. Apparatus

All experiments took place at the human-machine interaction lab facilities at the Aerospace Engineering faculty of the Delft University of Technology. This facility is regularly used to research manual control tracking tasks for automotive as well as aerospace applications. All experiments were conducted in a darkened room (4.3 m x 4.6 m x 3 m). The setup consisted of a visual projection system, a sound stimulation system, an inertial measurement unit (Movella Xsens Dot; serial: 40195bea809f0098) for measuring head movements, and a PC joystick (Thrustmaster TCA sidestick Airbus edition joystick). All systems were controlled and synchronized via a Linux-based PC system (3 Dell desktops, Ubuntu 18.04.2 LTSS) running the Delft University Environment for Communication and Activation (DUECA) [47]. DUECA is a module-based real-time software that accounts for software timing latency between individual modules (projector, sound, head tracker, joystick) and allows for easy distribution of modules across multiple computers. Visual stimuli were projected with a BenQ TH690ST projector on a white screen that had a width of 3.5 m and a height of 2.2 m. Sound stimuli were presented using two speakers of a 5.1 surround system (Logitech Z906). Participants were seated behind a wooden car bonnet (height: 1.70 m, depth: 1 m) on an adjustable chair (maximum height: 1.35 m, minimum height: 1.10 m) at a distance of 2.75 m from the screen. The bonnet was covered with a felt blanket to limit acoustical reflections. With the aid of the chair, the participant's interaural axis was aligned with the center of the projection screen. Maximum viewing angles were ±33° in azimuth and ±22° in elevation. A top view of the setup used for the experiments including the two speakers is presented in fig. 1.

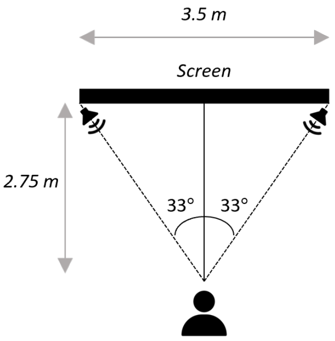


Figure 1: Top view of the setup used for the experiments including the speakers being located at the maximum view angles (±33°).

C. Stimuli

Target locations ranged from -33° to +33° with a step size of 4.7° in the azimuth direction. All targets were presented at an elevation of 0°. The selection of the step size was based on similar experiments performed in an anechoic room [28]. Negative/positive angles indicate locations to the left/right of the center line, respectively.

1) Auditory stimuli: We generated high-pass filtered Gaussian white noise (4-20 kHz) using custom-written Matlab (R2020b, The Mathworks) functions. To simulate the head shadow effect, we manipulated the relative sound levels for the left and right speaker [28]. First, a relative position vector was defined to specify the azimuth angle of the target, α (°) with respect to the maximum angle α_{max} (°).

$$I_{\mathrm{rel}_\alpha} = \frac{\alpha}{\alpha_{\mathrm{max}}} \tag{1}$$

Where α is always smaller than or equal to the maximum deviation angle from the center line, defined as 33° in all experiments.

$$\alpha \leq \alpha_{\mathrm{max}} \tag{2}$$

Resulting in a scaling vector I_{rel_α} which varies between [-1, +1]. Scaling factors for the sound intensities for the left (I_{L_α}) and right (I_{R_α}) speaker for each α are then calculated using

$$\begin{aligned} I_{L_\alpha} &= 0.5 \cdot (1 - I_{rel_\alpha}) \\ I_{R_\alpha} &= 0.5 \cdot (1 + I_{rel_\alpha}) \end{aligned} \tag{3}$$

2) Visual stimuli: Visual stimuli were white (RGB = [1.0 1.0 1.0]) and gray (RGB = [0.33 0.33 0.33]) circles subtending 0.64° and 0.2° of the participant's field of view for, respectively, the central fixation spot and the visual targets. All stimuli were projected on a black background (RGB = ([0.0 0.0 0.0]).

3) Audiovisual stimuli: For audiovisual stimuli, we both presented auditory and visual stimuli. All audiovisual stimuli were congruent, i.e., co-located in space and presented at the same time.

D. Experiment Procedure

We used two tasks to study sensorimotor processing. For the head-localization task, participants aligned their gaze with the perceived target location. For the arm-localization task, we required participants to align a visual marker, controlled via a joystick, with the perceived target location. Participants were instructed to move as accurately and quickly as possible to the perceived location. This deliberate contradictory instruction aims to prevent participants from focusing solely on one specific movement characteristic.

1) Head-localization task: To measure head movements, participants wore a head-band mounted head-tracker (Movella Xsens Dot). The tracker was positioned on the forehead facing the projection screen (fig. 2). Participants also wore glasses with a mounted laser pointer that projected a red beam onto a small, frame-attached disk (diameter: 1 cm) at 38 cm in front of the participant's nose (fig. 2). During the experiments, participants fixated their eyes on the laser dot projected on the disk and aligned this dot with the perceived target location. This protocol guaranteed that the eye-in-head position remained constant during stimulus presentation and endpoint alignment, thereby aligning head orientation with gaze direction.

In order to calibrate the head tracker, each participant performed a calibration experiment at the beginning of an experimental session. Firstly, a fixation point was presented at 0° azimuth and -5° elevation. Subsequently, visual stimuli were displayed at target locations with azimuth angles of approximately 0°, ±11°, ±22°, and ±33°, along with corresponding elevation angles of 0°, ±10°, and ±20° for each target location. These stimuli were displayed for a period of 5 seconds, during which participants were instructed to align a projected laser – and consequently, their head orientation – with the presented dot. This sequence was repeated for all specified target positions.

2) Arm-localization task: Arm responses were measured using a gain-controlled joystick positioned in front of the participant behind the bonnet. A deflection of the joystick

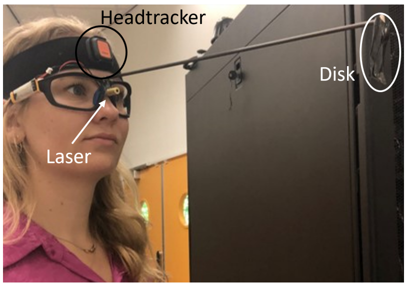


Figure 2: Close up of participant in a head-localization task. The participant is equipped with a head tracking device and specialized glasses containing a laser apparatus affixed to it. This laser projects onto a black surface situated 38 cm ahead of the participant. It should be emphasized that the glasses are exclusively worn during head-localization task.

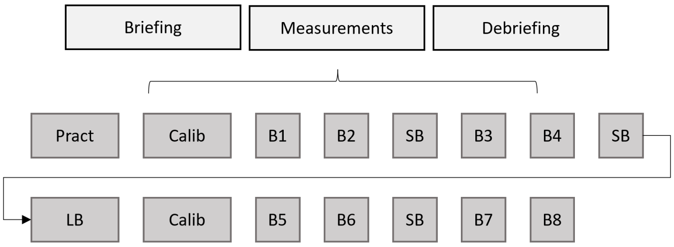


Figure 3: Schematic overview of the experiment procedure for the final experiments. Each session started with a practice block (Pract), after which the calibration block (Calib) started both with a duration of 5 min. Between blocks, participants had a break of 2 min (SB), after 4 blocks participants had a break of 10 min (LB). During each block (B), either head- or arm-localization tasks were performed whilst audio, visual, and audiovisual trials were mixed within the blocks.

directly related to a proportional position change of the projected output. The localization output of the joystick was visualized with a red (RGB = [1.0 0.0 0.0]) circle subtending 0.32° of the participant's field of view. Participants could move the joystick in both azimuth and elevation. Participants were allowed to move their heads freely and no specific instructions regarding head or eye movements were given. We measured the participant's head movement with the same head tracker as during the head-localization task.

3) General Procedure: The general experiment procedure is depicted in fig. 3. The experiment began with a short briefing, followed by several blocks of measurements, and ended with a short debriefing. To familiarize participants with the two tasks and to verify that the stimuli could be heard and seen by the participants, we included practice runs for both arm and head localization tasks. These practice runs comprised 30 trials each and lasted around 5 minutes. Following this, a 5-minute head tracker calibration experiment was conducted. Each block, containing either head or arm responses, encompassed a total of 135 trials. This configuration resulted in a duration of approximately 15 minutes for each block. Following the completion of two experimental blocks, participants were given a brief 2-minute break. Subsequently, after completing four blocks, participants were provided with a longer 10-minute break. The arrangement of blocks was randomized among participants for each effector. Appendix D contains the specific block sequence for each participant. Within each block, the three stimulus types were randomly mixed. Every block contained 3 repetitions of each stimulus condition and target location. This resulted in 12 repetitions for each target location per stimulus type and each effector. The entire experiment had a duration of around 3 hours.

4) General Trial Structure: Figure 4 provides an illustration of an arm trial, where the joystick output is depicted by a red dot. Head trials closely resemble arm trials, with the only distinction being the absence of the displayed red joystick output. Each trial started with a white fixation circle at 0° azimuth and -5° elevation, disappearing after 3 seconds. Following a randomized delay ranging from 0 to 1000 ms (in increments of 250 ms), the audio, visual, or audiovisual stimulus was presented for 100 ms. Randomization was introduced to prevent the predictability of stimulus presentation. Depending on the specific variable delay interval, the response interval ranged from 1000 to 1900 ms, resulting in an overall trial duration of 5000 ms.

E. Data preprocessing

The original time traces for the head, sampled at 60 Hz, and the arm, sampled at 100 Hz, were subjected to interpolation to achieve a higher sampling rate (1000 Hz). This higher rate was chosen to facilitate accurate determination of onset and offset times of the movements. Modified Akima Interpolation, implemented through the "interp1" function in Matlab, was employed for the interpolation.

The unprocessed data derived from the head tracker was initially represented in Euler angles. Employing a coordinate transformation matrix, these angles were subsequently converted into horizontal (azimuth) and vertical (elevation) angles in a fixed reference frame [48], to avoid singularity problems. After this, head response endpoints were calibrated using the data obtained in the calibration experiment. The endpoints for each target location were established by calculating the average measured angle during the interval spanning from 2 to 3 seconds after the onset of the stimulus. Combinations of converted elevation and azimuth angles from the head tracker and known corresponding visual target locations were used to train a two-layer neural network. The networks were trained by the Bayesian regularization implementation of the backpropagation algorithm (MatLab; Neural Networks Toolbox) to avoid overfitting [49]. The networks also accounted for small inhomogeneities in the measured data. The trained networks were then used to calibrate the experimental data per participant. Moreover, in every head trial, we accounted for the temporal disparities between the head tracker and the DUECA software by implementing manual corrections. These corrections ranged from 0 to 340 ms. The calibration as well as the time corrections are elaborately explained in Appendix A.

The only modification applied to the time traces of arm responses involved transforming the positional data, which was originally in terms of pixel units, into angles through the utilization of geometric relationships.

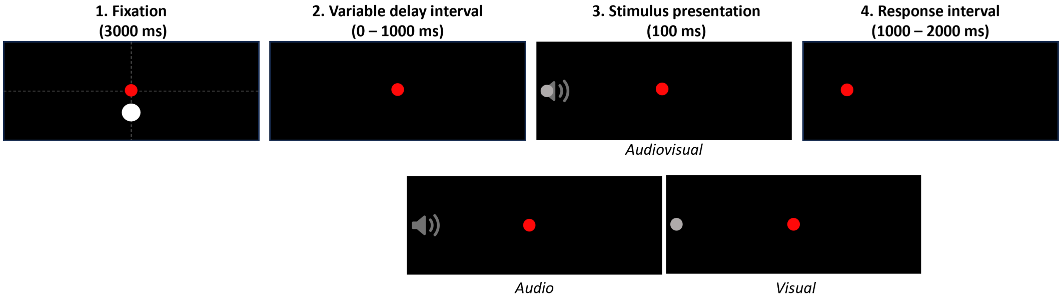


Figure 4: The trial structure for arm trials. Note that head trials are similar only the red joystick output is not shown. 1) Fixation dot is positioned at an elevation of -5° and azimuth of 0° and shown at the start of each trial. Note that for arm trials the fixation of the eyes (white dot) is not aligned with the initial position of the joystick (red) in elevation 2) Random delay is introduced ranging between 0-1000 ms in steps of 250 ms 3) Stimulus (audio, visual or audiovisual) presentation in which the audio and visual stimuli are always presented at the same time and location (congruent) for audiovisual trials. 4) Depending on the delay interval, the response interval varies between 1000 - 1900 ms resulting in a total trial time of 5000 ms

1) Detection of Arm and Head Movements: Here, we call goal-directed arm and head movements ‘saccades’ in analogy with goal-directed eye movements. A custom-written Matlab script was used to automatically detect saccades in the calibrated data by using two preset velocity criteria. The initial criterion is the minimum velocity necessary for categorizing the motion as a saccade, i.e. if the velocity criterion is not met, the movement is not detected as a saccade. The second velocity criterion establishes the start and end points of a saccade.

Figure 5 illustrates a head trial, presenting the positional trace (fig. 5A) and the velocity trace (fig. 5B). The shaded gray region depicts the saccade (Sac) selected, starting at the saccade onset (Sac Onset) and ending at the saccade offset (Sac Offset). In fig. 5A, the target’s (Tar) azimuth (Az) and elevation (El) are represented by thick solid horizontal black and red lines, while the azimuth and elevation of the fixation (Fix) are indicated by dashed lines. Response traces (Trace) for azimuth and elevation are shown using black and red lines, respectively. The thin grey line represents the stimulus trace (Stim) with a stimulus duration (Stim Dur) of 100 ms. The reaction time (RT) is the difference between the stimulus onset (Stim Onset) and the saccade onset. Saccades characterized by reaction times below 100 ms were eliminated from consideration due to their nature as anticipatory responses, as were saccades exhibiting reaction times surpassing 1000 ms. This temporal range is denoted as the response window. The azimuth and elevation error were determined by the difference between the target and the response at the saccade offset. It is noteworthy that multiple saccades could be identified within a given trial. Reaction times and endpoints were determined based on the attributes of the initial saccade occurring within the defined response interval.

In fig. 5B, the solid blue line represents the velocity trace. The initial velocity criterion of $40°/s$ is depicted as a solid horizontal line, while the second criterion $(10°/s)$ is represented by a horizontal dashed line. In the context of arm movements, the first velocity criterion was set at $20°/s$, and the second criterion was set at $1°/s$.

F. Data Models

1) Race Model: A commonly known model to quantify audiovisual benefits for reaction times is the statistical facilitation method also referred to as the race model [45], [50], [51]. This method relies on the principle that the benefit of providing two stimuli results from the so-called race between different unimodal stimuli and that the first neuronal representation arriving at a decision stage elicits a reaction. The race model posits that there is no interaction during the processing of the unimodal stimuli until they are merged at the decision stage.

Through a straightforward statistical analysis, and by treating the reaction time distributions of unimodal stimuli as independent variables, it becomes evident that the introduction of multimodal stimuli leads to consistently faster reaction times compared to unimodal stimuli, even in the absence of any interaction. [37], [52].

The race models can be computed using eq. (4). In this equation, $F_A(t)$ and $F_V(t)$ are the probabilities that the response started before time t for unimodal audio and visual stimuli, respectively. F_{race} is the resulting Cumulative Distribution Function (CDF) of the race model.

$$F_{race}(t) = F_A(t) + F_V(t) - F_A(t) \cdot F_V(t) \tag{4}$$

The race model assumes no interaction between unimodal stimuli processing, hence, violation of the race model implies that the benefit cannot be attributed solely to one of the unimodal responses and multisensory interaction is taking place. The difference between the multisensory response CDF and the race model is referred to as the Miller inequality [50]. Positive values indicate multisensory integration, however, negative values are more difficult to interpret and may for instance be due to cognitive factors influencing the response [50].

2) Maximum Likelihood Estimation (MLE) Model: Currently, one approach in multisensory research revolves around the common source assumption [31]. That is, the brain assumes that all sensory information originates from one source. Accordingly, sensory information from distinct modalities

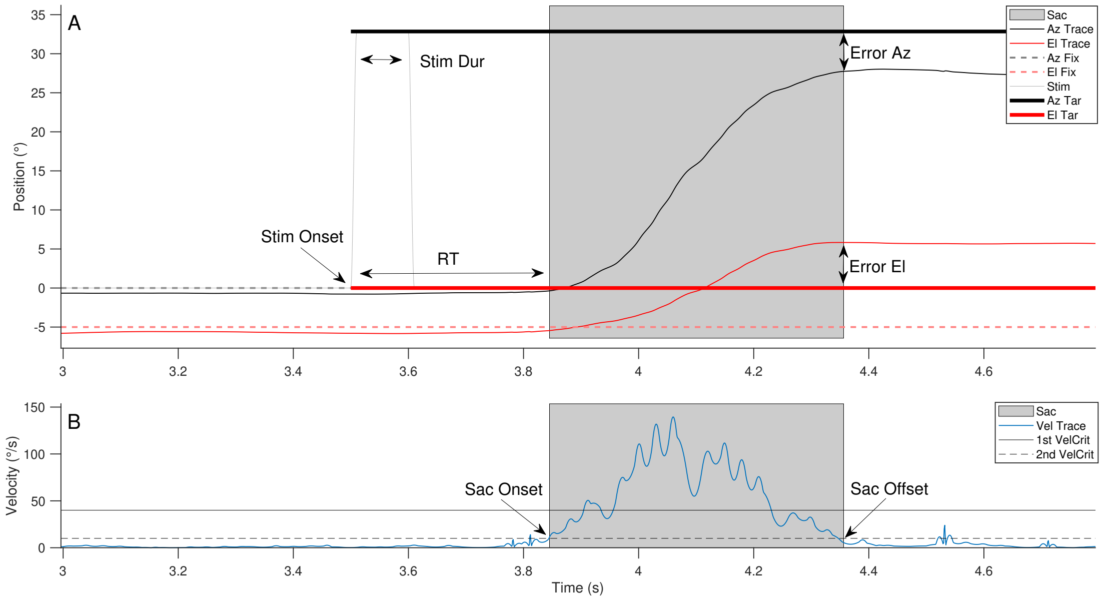


Figure 5: Illustration of a head trial, presenting the positional trajectory (fig. A) and the velocity trajectory (fig. B). In fig. A, the target's azimuth (Az Tar) and elevation (El Tar) are denoted by black and red horizontal lines, while the azimuth (Az Fix) and elevation of the fixation (El Fix) are represented by dashed lines. Responses for azimuth (Az Trace) and elevation (El Trace) are indicated with black and red lines respectively. The shaded gray region depicts the detected saccade (Sac), starting and ending at the second velocity threshold ($10°/s$). The thin grey line shows the stimulus trace (Stim), with a stimulus duration (Stim Dur) of 100 ms. In fig. 5B, the initial velocity threshold ($40°/s$) is presented as a solid line, while the subsequent threshold ($10°/s$) is shown as a dashed line. The velocity trace is depicted in blue. Saccade processing resulted in the following parameters: reaction time (RT) (i.e. time between stimulus onset (Stim Onset) and saccade onset (Sac Onset)) and saccade endpoints (i.e. horizontal and vertical location at saccade offset (Sac Offset). The error is defined as the difference between the target location and the saccade endpoint for azimuth (Az) and elevation (El).

should be combined. This assumption holds when the two signals are exhibited either without any contradiction or with a minor cue-related discrepancy, such as a spatial difference of 6° [31].

In cases where the common source assumption is valid, the Maximum Likelihood Estimation (MLE) model formulates predictions of the participants' spatial estimates. These result from the fusion of the auditory and visual-spatial estimates. These predictions specifically relate to the utilization of sensory weights during the process of integration formulated in eq. (5). In which $\hat{S}_{AV}$ represents the estimate of the object's location. Whilst, $\hat{S}_A$ and $\hat{S}_V$ present the perceived locations based on auditory and visual stimuli, respectively. The weights $w_{A_{MLE}}$ and $w_{V_{MLE}}$ are calculated using the endpoint variance of auditory (σ_A^2) and visual (σ_V^2) responses.

$$\hat{S}_{AV} = w_{A_{MLE}}\hat{S}_A + w_{V_{MLE}}\hat{S}_V \tag{5}$$

with

$$w_{A_{MLE}} = \frac{\frac{1}{{\sigma_A}^2}}{\frac{1}{{\sigma_A}^2} + \frac{1}{{\sigma_V}^2}}$$

and

$$w_{V_{MLE}} \frac{\frac{1}{{\sigma_V}^2}}{\frac{1}{{\sigma_A}^2} + \frac{1}{{\sigma_V}^2}}$$

3) Empirical weights: To determine the effectiveness of the Maximum Likelihood Estimation (MLE) method in representing the data, we calculated weights based on the audio, visual and audiovisual responses eq. (5). In this context, L_A, L_V, and L_{AV} denote response characteristics, i.e. reaction time and endpoint locations, for auditory, visual, and audiovisual responses. For each target location, responses were randomly matched. Weights were found by minimizing the mean-squared error of eq. (6) with the Matlab routine fminbnd (Levenberg-Marquardt).

$$L_{AV} = (1 - w_{EMP_V})L_A + w_{EMP_V}L_V \tag{6}$$

4) Linear regression: To compare between two conditions a simple linear regression model can be used:

$$c_2 = \gamma \cdot c_1 + \beta \tag{7}$$

Responses for one effector/modality, condition 1 (c_1), are directly compared to another effector/modality, condition (c_2). Linear regression parameters γ and β are solved by minimizing the mean squared error [53]. Correlation between the two conditions is assessed with the correlation coefficient R^2 and the error is defined as the value of the residuals between the data and the linear fit. The 95% confidence intervals are calculated using a bootstrapping approach, wherein 75% of the data was resampled iteratively for 1000 cycles.

G. *Hypotheses*

The hypotheses are categorized into two primary topics. First, the hypothesis related to sound localization is presented, followed by two hypotheses concerning multisensory benefits.

I **Sound localization:**

- Discrete sound-source locations can be adequately simulated using a two-speaker setup (azimuth +/-33°). By imposing volume changes that mimic ILDs at the listener's ears, a resolution of 4.7° can be achieved in an echoic room. This resolution is equal to the distance between adjacent target locations.

II **Multisensory benefits**

a) Compared to unimodal stimuli, faster reaction times and decreased localization variance are expected for multimodal stimuli. The spatial percepts of the auditory and visual stimuli are expected to align sufficiently, resulting in audiovisual interactions.

b) Audiovisual benefit is independent of the effector. The reduction in reaction time and endpoint variance with audiovisual stimuli compared to unimodal stimuli is expected to be the same for head and arm movements.

III. Results

A. *Sound localization*

To evaluate sound-localization behavior, we first evaluated the endpoint locations for each target location in azimuth and elevation. In fig. 6 the responses for all head and arm trials of participant 7 are depicted as an example. To provide a clearer overview, we also displayed the mean of the responses for each target location using squares, while the standard deviation in azimuth and elevation is depicted using thick horizontal and vertical lines. By evaluating responses for both head and arm responses, we observed a distinct clustering of average responses for target locations exceeding 9° (azimuth), concentrating towards more peripheral locations. Interestingly, despite all target locations being presented at an elevation of 0°, the participant had a pronounced elevation response, particularly around the centerline (azimuth 0°). Additionally, the participant's elevation responses were slightly higher and more variable in elevation for azimuth locations close to 0°.

Next, we evaluated the ability to accurately localize target azimuth by calculating the mean error between the target location and mean response endpoints (fig. 7). Data from individual participants are shown in grey, the mean across all participants is shown with black squares, and the standard deviation across participants per target location is shown with grey vertical lines. For positive target locations, negative errors indicate that the response is more toward the centerline while a positive error indicates that the response was more peripheral than the target location. For negative target locations, the opposite is true.

Though responses could vary across participants, a clear pattern is visible for both head (fig. 7A) and arm (fig. 7B) responses. For target locations up to about ±23°, participants overshot, and for greater angles undershot the target locations (±28°, ±33°). Leading to the clustering of audio responses at an approximate angle of 25°. The largest errors were observed for the most peripheral locations and azimuth location ±10°. Interestingly, azimuth errors were smallest around the centerline and at about ±23°. The individual stimulus-response plots for all participants are given in Appendix B. Overall, azimuth localization behavior seemed to be fair given the echoic environment compared to experiments performed in anechoic rooms [28].

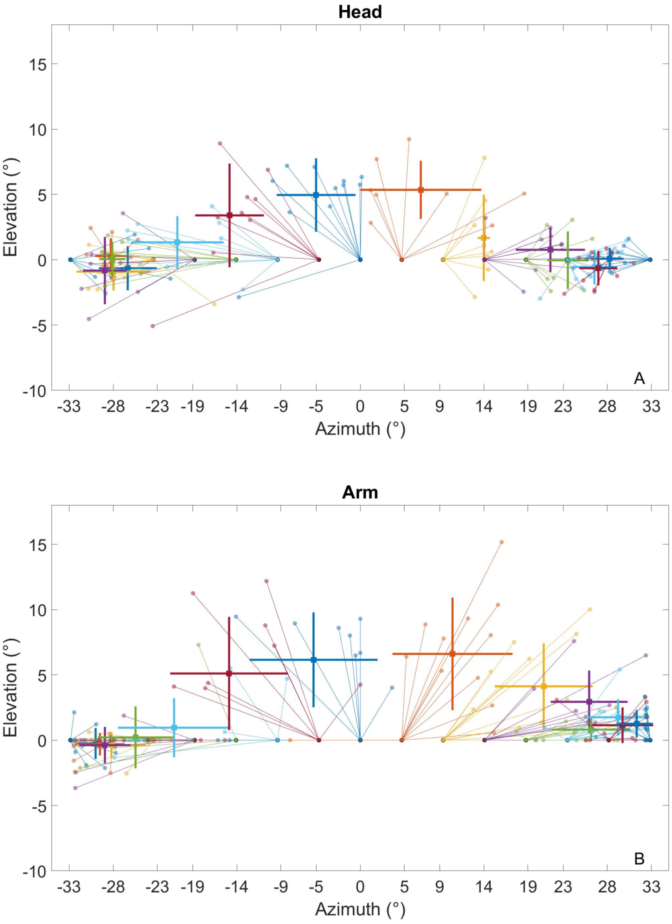


Figure 6: Head (A) and arm (B) auditory response endpoints to sounds at simulated azimuth locations for participant 7. The large and small circles indicate target locations and response locations for individual trials, respectively. Additionally, target and response locations are connected by thin lines in matching colors. Average response endpoints are indicated with squares and standard deviation in azimuth and elevation are depicted as thick horizontal and vertical lines, respectively.

To further quantify localization behavior we calculated the Mean Absolute Error (MAE) per azimuth target location. To ensure that participants could discriminate between adjacent target locations, the MAE must be smaller than the angular separation between two adjacent targets.

Across all participants, the MAE for head responses was found to be 4.9° [σ = 1.2°], while for arm responses, it was found to be 5.6° [σ = 2°]. Both MAE values were larger than the difference of 4.7° between adjacent target locations. This suggests that participants were unable to discriminate between adjacent locations. No clear pattern emerged in the MAE values across different target locations for both head and arm responses. However, it is worth noting that participant 2 exhibited the highest resolution among all participants, achieving an MAE of 3.3° for head responses and 2.8° for arm responses, respectively. In contrast, participant 8 demonstrated the lowest resolution for both head and arm responses, with MAE values of 8.2° and 10.1°, respectively.

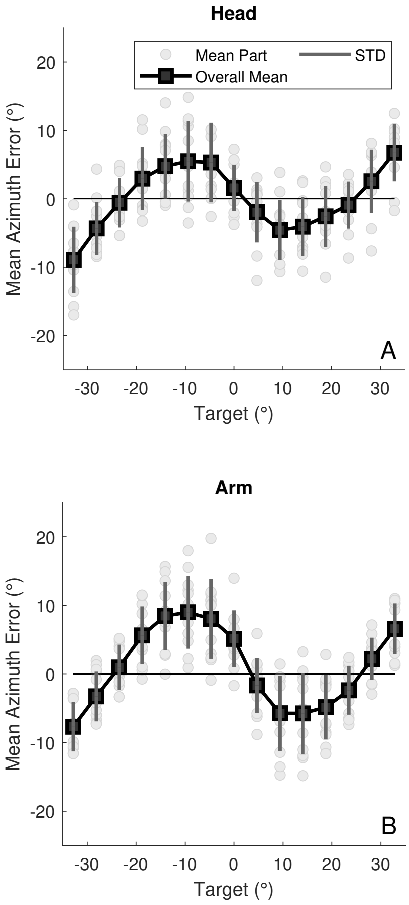


Figure 7: Mean error per participant (part) and standard deviation (STD) per target location were plotted with grey dots and dotted lines, respectively. The pattern of positive and negative errors results in responses clustered around 25°.

To visualize elevation response patterns across participants, we show the mean error in elevation responses per target location for head responses (fig. 8A) and arm responses (fig. 8B). Conventions are as in fig. 7. Similarly, as observed for participant 7 in fig. 6, all participants displayed distinct elevation responses near the centerline, reaching up to 8°, even though the simulated sounds lacked an elevation component. Towards peripheral target locations elevation response errors gradually decrease to approximately 0°.

Furthermore, the greatest variability in the mean responses was observed near the centerline, diminishing as the locations moved toward the periphery. It should be noted that the aural axis of our participants was vertically not precisely aligned with the speakers in our experimental setup, resulting in variable elevation perception among participants. Subsequently, it is plausible that slight disparities in speaker volume output contributed to a decorrelation between the two sound waveforms. Decorrelation is known to lead to ‘fuzzy’ or ‘cloudy’ percepts, leading to higher errors in the elevation percept [54]. In line with this, our participants reported that they identified some sounds that could be well localized whilst others were more diffuse. Moreover, they reported perceiving three clusters of sound locations: one situated centrally and two positioned at the outermost periphery.

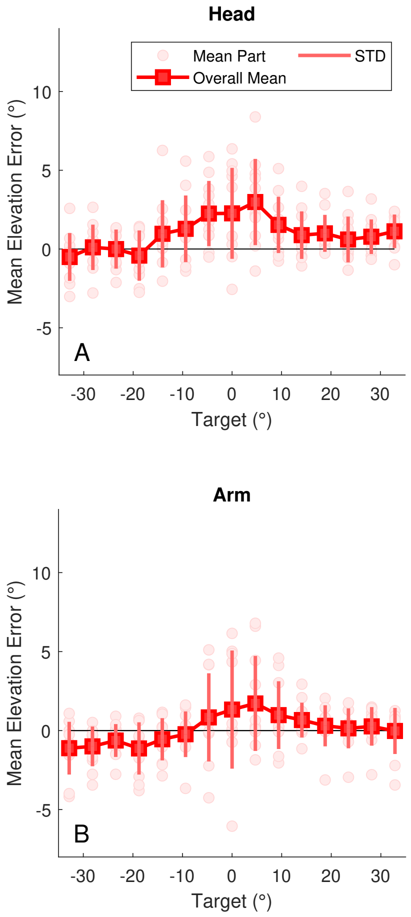


Figure 8: Elevation localization error as a function of target azimuth location. Data from individual participants are denoted as light red circles. The mean per target location and standard deviation (STD) across participants are shown with red squares and vertical lines, respectively. Positive errors indicate more upward responses relative to the target (0°).

B. *Audiovisual benefits*

After evaluating sound localization behavior, we proceeded to explore the interaction between audio and visual stimuli, focusing on the participants’ reaction times and the accuracy of their endpoint localization. Figure 9 illustrates the two-dimensional distributions, showing the relationship between reaction times and localization errors for responses to auditory (blue), visual (red), and congruent audiovisual stimuli (black) for a single participant. Each dot corresponds to a response, while squares indicate the mean localization error and the median reaction time. We also utilized ellipses to represent the standard deviation and interquartile range (IQR) for both the localization error, and the reaction time, respectively. Because reaction times are not normally distributed, we choose to summarize reaction times distributions by calculating their median and interquartile range (75th-25th percentile). In Appendix B plots for all participants can be found.

In terms of head responses, the most substantial average errors were noted for auditory responses, whereas the smallest errors occurred with visual responses (A/V/AV:-2.6°/-0.1°/1°). Conversely, concerning arm responses, the largest average error was observed in audiovisual responses, while the smallest average error was recorded for auditory stimuli (A/V/AV: 0.1°/-0.4°/0.7°). For auditory responses, the variance was larger when head responses were the effector (A/V/AV: 8.7°/5.4°/4.8°) compared to arm responses as the effector (A/V/AV: 9.6°/5.2°/4.3°). For both head and arm responses, the variance in localization error is slightly smaller for audiovisual responses compared to visual responses. Analyzing the reaction times for head responses, we found that audiovisual responses exhibited the smallest median reaction time (320 ms), while visual responses (377 ms) had the largest median reaction time, and auditory-evoked reactions fell in between the two (344 ms). For arm responses, audiovisual reaction times (370 ms) were faster than those elicited by auditory (400 ms) and visual (401 ms) stimuli, which showed similar reaction times. Furthermore, the analysis revealed that the IQR of reaction times in head-localization tasks was largest for auditory responses (167 ms) and smallest for audiovisual responses (70 ms), with the IQR of visually-evoked reaction times falling in between the two (109 ms). Similarly, this was observed for arm responses (A/V/AV: 99/59/60 ms). Note that this discrepancy in reaction time variability between head and arm responses is unexpected. Due to the complexity of the arm movement, one would expect a larger spread in reaction times for arm responses compared to head responses. Most likely, this problem is caused by underlying timing issues in the measurements of head responses, extensively discussed in Appendix A. In order to extend this analysis to all participants, we will begin by addressing the localization error, followed by a discussion of the reaction times.

1) Response endpoints: First, we compared localization variability for head (fig. 10A) and arm (fig. 10B) responses between multimodal, i.e., audiovisual, and unimodal, i.e., auditory and visual, responses. To visualize the variance of audio (A) and visual (V) responses, we used blue squares and red dots, respectively. The graph includes a black unity line indicating a one-on-one relationship between unimodal and multimodal variance. Points situated below the black line indicate a greater variance of unimodal responses compared to the multimodal variance, while points above the black line signify the opposite scenario. Visual variance tends to cluster around the unity line, indicating similarity to the multimodal variance. Out of the 12 subjects, 5 showed increased variance for multimodal responses in head movements when compared to the variance of unimodal visual responses. Similarly, for arm movements, 9 out of the 12 subjects displayed higher variances in response to multimodal stimuli as opposed to unimodal visual stimuli. Moreover, the visual variances are located to the left of the unimodal audio variances, indicating a lower variance for visual responses in general. This is in line with the lower average variances observed for visual responses ($\sigma^2_{head} = 25^{\circ 2}$, $\sigma^2_{arm} = 14^{\circ 2}$) compared to auditory responses ($\sigma^2_{head} = 74^{\circ 2}$, $\sigma^2_{arm} = 64^{\circ 2}$). Furthermore, for head responses, the audiovisual variance ($\sigma^2_{head} = 24^{\circ 2}$) is equivalent to the visual variances, whereas, for arm movements, the audiovisual variance ($\sigma^2_{arm} = 20^{\circ 2}$) is larger than the visual variances. The variance of auditory responses exhibits notable differences among participants, as evidenced by the large spread along the x-axis.

To assess the effector dependency of the reduced variability for multimodal stimuli, we defined the audiovisual benefit as the difference between the variance observed in unimodal responses and multimodal responses. A negative audiovisual benefit signifies a larger variance in multimodal responses, whereas a positive audiovisual benefit indicates a decrease in variance for multimodal responses. To directly compare the effector dependency, we plotted the audiovisual benefit of head responses against the audiovisual benefits of arm responses for auditory-evoked (A, blue) and visually-evoked (V, red) responses (fig. 11). The unity line is indicated with a solid black line. Points located below the unity line indicate a larger audio-visual benefit for head responses, whereas points above the line suggest a larger benefit for arm responses. Responses with small or negligible audiovisual benefits are indicated in the lower left corner. To assess the correlation between head and arm benefits, we fitted a line through the data. The blue and red lines indicate the linear fit through, respectively, the auditory and visual responses, with the specifics of these lines given as a text inset in the top left corner of the graph. If a one-on-one relation exists, the best linear fit is described by a gain (γ) close to one, a bias (β) close to zero, and a R^2 coefficient close to one.

Our analysis revealed that for the comparison between arm and head responses, audiovisual benefits compared to unimodal visual responses were poorly correlated (R^2 = 0.50). The relatively low coefficient of determination (R^2) suggests a weak correlation, underscoring the difference in audiovisual benefit between arm and head responses for unimodal visual stimuli. In comparison to unimodal audio responses, audiovisual benefits exhibited a stronger correlation (R^2 = 0.77). Nevertheless, the elevated gain (γ) signifies that a direct one-to-one relationship between the audiovisual benefit in unimodal audio responses, for head and arm responses, is not present. Furthermore, for visual responses, 9 out of 12 participants exhibited negative audiovisual benefits with arm responses. In contrast, for head responses, 5 out of 12 participants displayed a positive audiovisual benefit, indicating that the addition of audio stimuli had a more pronounced effect on head responses. Moreover, the audiovisual benefit was large for auditory responses as responses are situated above the unity line for all but three participants. This analysis suggests that audiovisual responses were dominated by the visual stimulus, and that audiovisual benefit was effector dependent.

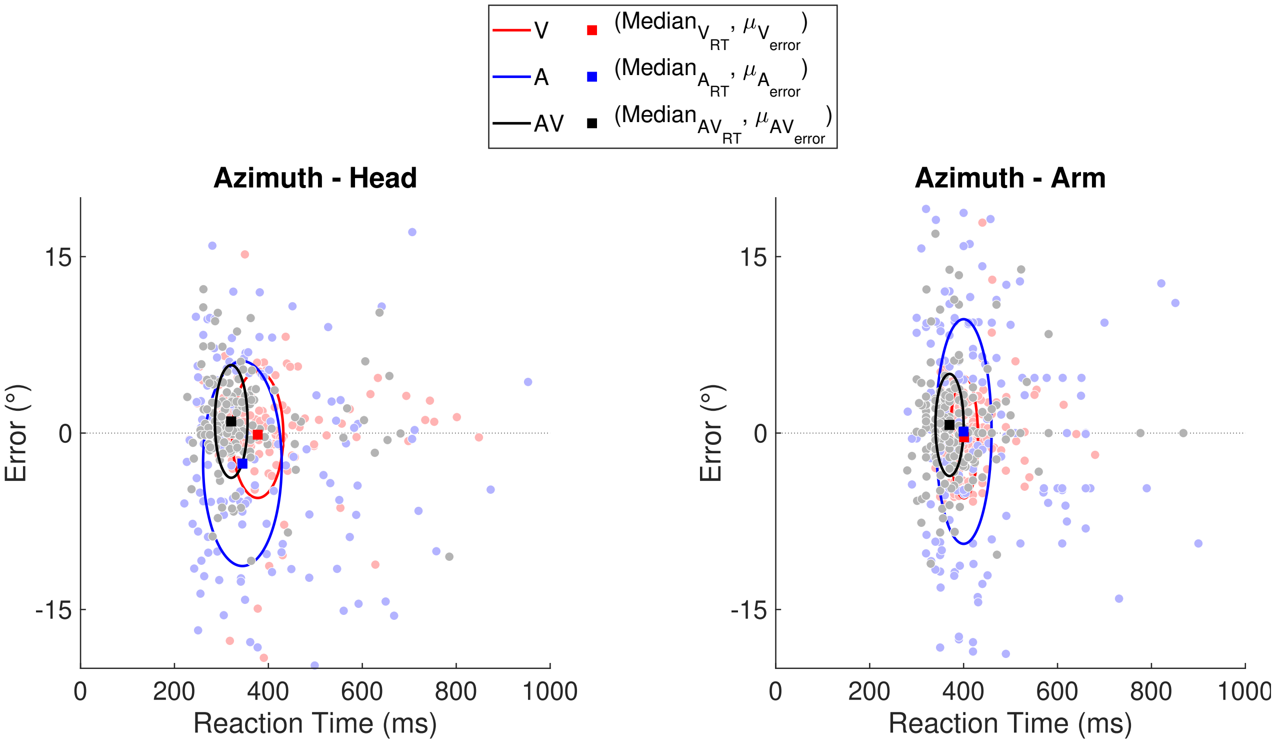


Figure 9: Localization error as a function of reaction time for auditory (A, blue), visual (V, red), and audiovisual (AV, black) responses of participant 11. Squares indicate the median reaction time (x-axis) and mean localization error (y-axis). Ellipses depict the interquartile range (75th-25th percentile) of the reaction times (x-axis) and the standard deviation of the errors (y-axis). Note the difference in ellipse sizes for different modalities

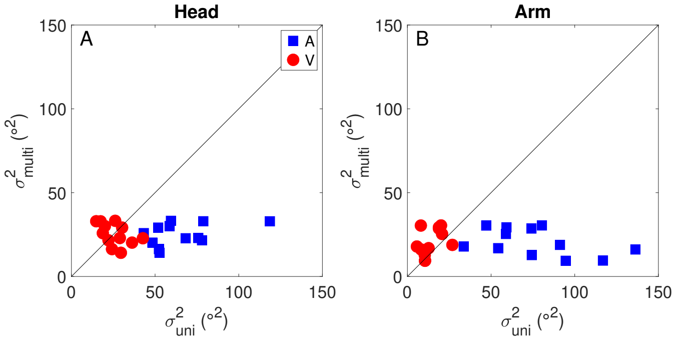


Figure 10: Comparison between multimodal (σ_{multi}) and unimodal (σ_{uni}), i.e., auditory (A, blue) and visual (V, red), localization variances for head (A) and arm (B) responses of all participants. The unity line is indicated with a black line.

2) Reaction times: Following the discussions of the response endpoints, we analyzed the reaction times across various modalities. In fig. 12, we present the median reaction times of responses to both unimodal and multimodal stimuli per participant pooled across locations. The figure conventions are the same as the previous figures. Correlations between audiovisual median reaction times and auditory (A, blue) and visual (V, red) for head and arm responses are depicted in fig. 12A and fig. 12B, respectively. We performed linear regression analyses to establish the best linear fit for the visual and auditory reaction times. The resulting fits are represented by the solid red line (visual) and the blue line (auditory) whilst the confidence intervals are shown as red and blue shaded areas. The specifics of these lines, including the gain (γ), bias (β), and coefficient of determination (R^2), are indicated in the upper left corner of the graph. To aid in interpretation, we included a black unity line representing a one-to-one relationship between unimodal and multimodal reaction times. Points below this line signify instances where unimodal responses were slower compared to multimodal responses.

The majority of unimodal reaction times for both head and arm responses were indeed slower than their audiovisual counterparts, as expected based on Hypothesis IIa. The linear fit for both auditory and visual reaction times was below the unity line, indicating that unimodal responses are indeed slower than multimodal responses.

To assess the statistical significance of the disparity between the fitted lines, we computed the confidence intervals for the linear fit. Both for arm and head responses, the confidence intervals for auditory and visual median reaction times overlap, suggesting no statistically significant difference between the modalities. Additionally, the one-on-one relationship with the multisensory reaction times, depicted in black, only overlaps slightly with the confidence interval of the unimodal auditory reaction times for head responses. For arm responses, unimodal reaction times are significantly slower compared to multimodal reaction times, as shown by the plotted confidence interval. When comparing the average median reaction times across all participants, it was evident that audiovisual reaction times (390 ms) were faster compared to both

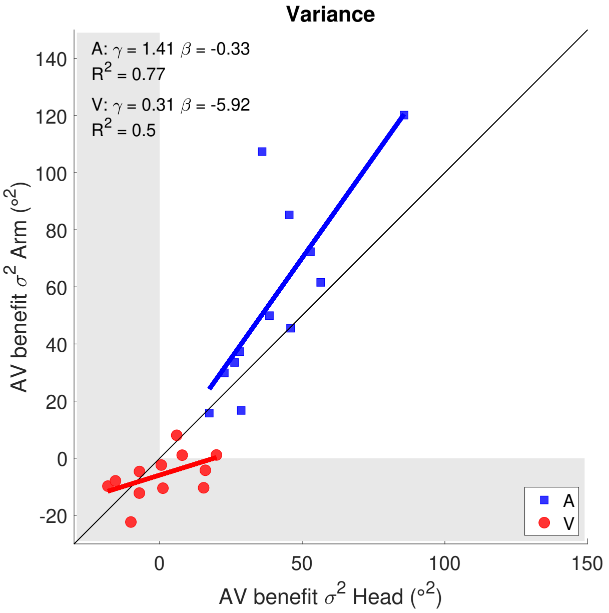


Figure 11: Comparison of unimodal responses, audio (A, blue) and visual (V, red), with respect to multimodal response pooled over all target locations per participant. The unity line is indicated with a solid black line, and linear regression lines for auditory and visual benefits with blue and red lines, respectively. Regression line parameters are given in the text inset, with γ and β indicating, respectively, the gain and bias. Points within the shaded grey region do not demonstrate audiovisual (AV) benefits for arm and/or head responses. Note that the variance for visual responses is significantly lower compared to auditory variance.

auditory (428 ms) and visual (429 ms) reaction times. Moreover, for head responses, only visual unimodal responses are significantly faster compared to audiovisual responses. Nonetheless, the average median reaction times for audiovisual responses (380 ms) were still faster than those for auditory (415 ms) and visual (438 ms) responses.

Similarly, as for the endpoint variances, we analyzed the effector dependency of audiovisual benefit defined as the difference in median reaction time between unimodal responses and multimodal responses (fig. 13). A negative audiovisual benefit signifies an increase in reaction time for multimodal responses, whereas a positive audiovisual benefit indicates a decrease in reaction time for multimodal responses. We correlated the audiovisual reaction time benefit of arm responses with that of head responses for auditory (A, blue) and visual (V, red) reaction times. The unity line is indicated with a black line and best linear fits with blue (audio) and red (visual) lines. In all but three participants audiovisual reaction time benefits relative to unimodal visual reaction times were larger with head responses compared to arm responses. The correlation between arm responses and head responses was weak (R^2 = 0.29) indicating that reaction time benefits were effector dependent. For the reaction time benefit between audiovisual and unimodal auditory reaction times, the correlation between arm and head responses was higher (R^2 = 0.74) indicating that the difference in benefit between effectors was less for the unimodal auditory condition. Overall, this result seems to suggest that the audiovisual benefit, for reaction times, can be effector dependent. This is in line with the effector-dependent benefit observed for the variability in the endpoints.

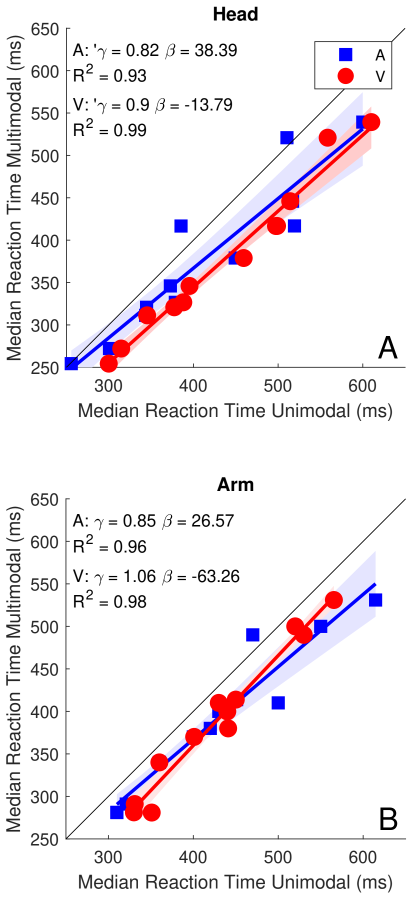


Figure 12: Multimodal median reaction times as a function of auditory (A, blue) and visual (V, red) unimodal reaction times. The resulting linear fit is represented by the solid red line (visual) and the blue line (auditory) whilst the confidence intervals are shown as red and blue shaded areas. The black line indicates the unity line. Note that the majority of both auditory and visual reaction times are below the unity line, indicating that unimodal reaction times were slower than multimodal reaction times.

We applied the race model to investigate potential multisensory integration. For the race model, the underlying assumption is that if audiovisual reaction times are faster than what the model predicts, audiovisual integration occurs. Conversely, if audiovisual reaction times are slower than predicted, it suggests the possible occurrence of negative interactions between the modalities.

An illustrative example of the race model is shown for one participant, involving both arm and head responses, in fig. 14. For head (fig. 14A) and arm (fig. 14B) responses the cumulative distribution function (CDF) for visual (V) reaction times is represented by a red line, auditory (A) reaction times by a blue line, and audiovisual (AV) reaction times by a black line. The race model, defined by eq. (4), is displayed in grey. The shaded grey area reflects the violation of the race model and the overall violation is indicated in the legend. In fig.

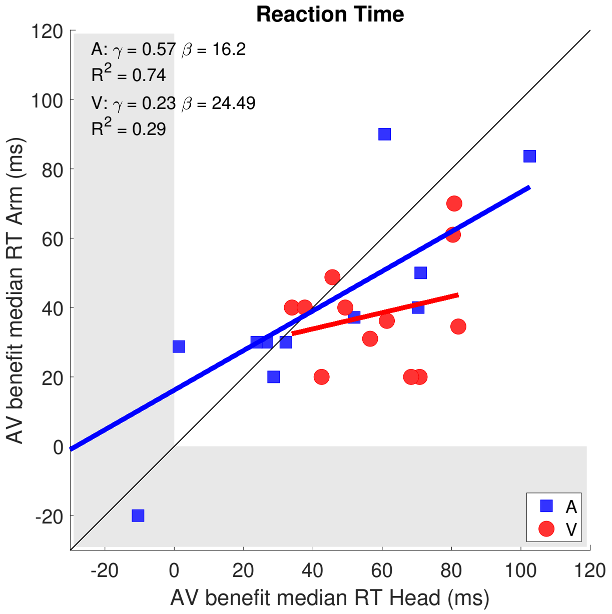


Figure 13: Comparison of unimodal responses, audio (A, blue) and visual (V, red), concerning multimodal response pooled over all target locations per participant. The unity line is indicated with a solid black line, and linear regression lines for auditory and visual benefits with blue and red lines, respectively. Regression line parameters are given in the top left corner, with γ and β indicating, respectively, the gain and bias. Points within the shaded grey region do not demonstrate audiovisual (AV) benefits for arm and/or head responses.

14C and fig. 14D the violation of the race model, which is defined as the difference between the CDF of the measured audiovisual responses (F_{AV}) and the race model (F_{race}), is shown.

Upon closer examination of the reaction times for head responses (fig. 14A and fig. 14C), we observe a clear negative violation up until 450 ms, implying that in trials with such fast reaction times, multisensory integration occurred. For reaction times larger than 450 ms, a positive violation is observed, with the largest violation occurring around 500 ms. Similarly, for arm responses (fig. 14B and fig. 14D), we observe a negative violation up until 580 ms, followed by a positive violation. Notably, when evaluating the unimodal responses for arm responses (fig. 14B), we observe that the CDF for visual responses is positioned to the left of the unimodal audio CDF, indicating that for participant 10, visual arm responses were faster compared to unimodal audio responses. The median visual reaction time was 52 ms faster compared to auditory reaction times. A similar trend is observed for head responses, for reaction times up to 500 ms, although to a lesser extent, with both CDFs being more intertwined (fig.14).

The violation of the race model exhibited substantial variability across participants, as demonstrated by three illustrative examples in fig. 15 for both head (fig. 15A) and arm (fig. 15B) responses. Participant 6 (blue) and participant 9 (orange) displayed a consistent positive violation across all reaction times for head responses. This indicates that the race model predicted faster reaction times compared to the actual measured responses for all conditions. Participant 9 also exhibited a purely positive violation for arm responses, while participant 6 (blue) demonstrated a purely negative violation suggestive of audiovisual integration for arm responses. Moreover, participant 7 (red) demonstrated an interesting pattern of race model violations. For head responses, a negative violation was observed for reaction times smaller than 300 ms, whereas a positive violation emerged for reaction times larger than 300 ms. Intriguingly, for arm responses, an even larger negative violation was evident for participant 7 for reaction times smaller than 400 ms. These observations suggest that the multisensory integration processes and the influence of the race model vary significantly across individuals and specific conditions. Furthermore, the average across all participants is shown in black. Notably, on average, there is a negative violation observed for reaction times below 390 ms in arm movements. For head movements, on average, only positive violations are present.

3) Maximum Likelihood Estimation: To investigate the individual contributions of the two modalities to audiovisual localization behavior, we utilized a maximum likelihood estimation (MLE). This model assumes that the multisensory percept arises by weighing the unimodal responses based on signal reliability. In order to validate the accuracy of the MLE-derived weights, we conducted a comparison with empirically derived weights. Note that MLE-derived weights are based on endpoint variances whilst the empirically derived weights are based on a fit of eq. (6) to the endpoint locations (section II). In fig. 16, we present the median visual weights for each target location across all participants for head responses (fig. 16A) and arm responses (fig. 16B). The MLE weights and empirical (emp) weights are represented as grey and red squares, respectively, and are connected by black lines for visual clarity. A weight of 1 denotes an audiovisual percept driven purely by visual input, while a weight of 0 signifies an audiovisual response solely influenced by the auditory input. Weights around 0.5 indicate an equal weighting of auditory and visual inputs. Note that the weights for arm responses are absent for the target location azimuth of 0°. This occurred because participants were not able to see the visual target due to the projection of the joystick.

For the empirical weights (red) obtained with head responses (fig. 16A), we observe an inverted U-shaped curve with near-zero visual weights for a target location of ±33°. From target locations ±28° to the centerline, visual weights increased nearly linear reaching a maximum of about 1.0 at the 0-degree target location. This observation is in congruence with the fact that the fovea, i.e., the area of sharpest vision, was aligned with the centerline at the beginning of the trial. Similarly, auditory weights are expected to be higher for more peripheral locations, where retinal resolution is lower. That is, vision dominates in central locations and audition dominates in peripheral locations.

The trend observed with the empirical weights (red) was consistent with the MLE weights (grey). In fact, the empirical weights and MLE weights were highly correlated for both head (R^2 = 0.98) and arm responses (R^2 =0.9), indicating the validity of the MLE model for our audiovisual data.

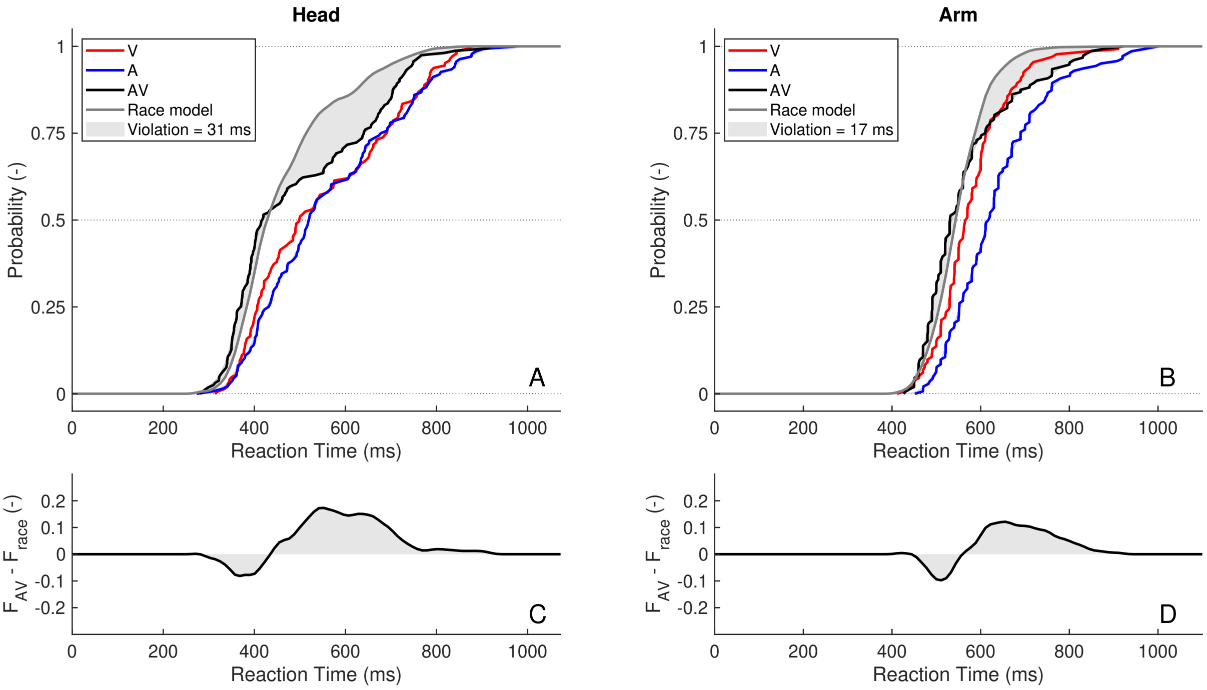


Figure 14: Cumulative distribution functions (CDFs) depict reaction times for participant 10 in response to visual (V), audio (A), and audiovisual (AV) stimuli (fig. A and fig. B). The grey representation represents the race model, with the shaded area indicating the total violation. Figures C and D provide a detailed view of the violation by illustrating the difference between the race model (F_{Race}) and the CDF of measured audiovisual reaction times (F_{AV}). This analysis offers insights into multisensory integration processes, for a single participant, revealing deviations between the predicted race model and actual audiovisual response distributions.

Interestingly, weights for the most peripheral location tested differed clearly between head responses (0-0.2) and arm responses (0.5-0.6), indicating an effector dependency for audiovisual interaction.

Likewise, we conducted location-dependent unimodal weighting for the measured reaction times. Given that the MLE model is related to the perceived location of an object and is not applicable to reaction times, we exclusively applied the empirical model given by eq. (6). The median visual weights for unimodal reaction times within the audiovisual percept are illustrated for both head responses (fig. 17A) and arm responses (fig. 17B). Again, note that the weights for arm responses are absent for the target location azimuth of 0°, as participants were not able to see the visual target due to the projection of the joystick. A visual weight of 1 signifies that multimodal reaction times are solely influenced by visual reaction times, whereas a visual weight of 0 signifies an audiovisual reaction time derived from auditory reaction times.

For head responses, the visual weights exhibit slight fluctuations around 0.5, indicating an equal weighting of visual and auditory reaction times. The comparatively lower visual weights in contrast to the visual weights for perceived locations align with the ”best-of-both-worlds” hypothesis, as auditory reaction times are faster than visual reaction times.

In the context of arm responses, we can identify three distinct regions. Firstly, in the case of the most peripheral target locations (±33°), we observe notably low visual weights (approximately 0.2), indicating a predominant influence of auditory processing on the audiovisual reaction times. In the second region, spanning from target locations ± 13° to ± 28°, visual weights of approximately 0.6 are apparent, reflecting a balanced weighting of auditory and visual reaction times. Lastly, we note high visual weights for target locations situated closer to the centerline (approximately 0.9). The decreasing visual weights for more peripheral locations align with the concept that audio stimuli are most salient at the extreme periphery, while visual acuity is more pronounced for central locations. Surprisingly, we did not observe this for head movement, possibly due to timing inconsistencies of the head tracker, extensively explained in Appendix A.

IV. Discussion

This paper aimed to assess the potential audiovisual benefit when performing localization tasks in an echoic room. We first assessed the ability to localize simulated sound sources in an echoic room. Following this, we conducted a comparison between responses to audiovisual stimuli and responses to individual unimodal stimuli (both auditory and visual) to quantitatively measure potential audiovisual benefits. These benefits were defined by reductions in reaction times and decreased variance when contrasted with unimodal stimuli. By assessing both head and arm responses, we were able to analyze the influence of the effector on the observed audiovisual benefits.

A. Sound Localization

Initially, we examined the sound localization behavior in an echoic room, employing two speakers to simulate sound sources positioned horizontally. Hypothesis I posited that a

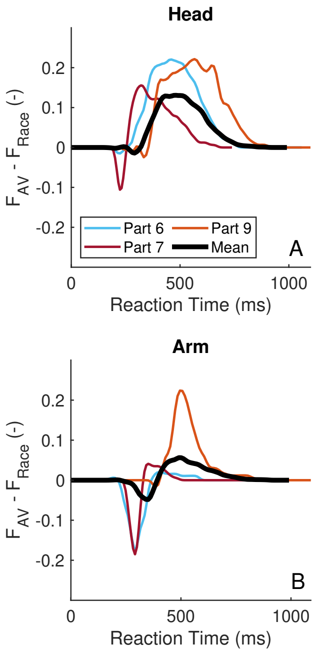


Figure 15: Difference between CDF as predicted by the race model (F_{race}) and the distribution of the measured audiovisual reaction times (F_{AV}) for three participants. The overall mean is shown in black. Note the differences between the participants for the CDFs as well as differences between the effectors.

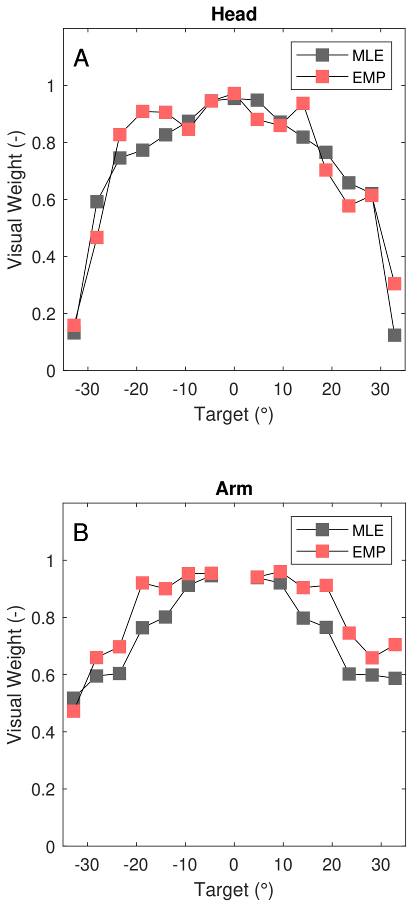


Figure 16: Visual (V) weights, derived through both the Maximum Likelihood Estimation (MLE) model (depicted in grey) and empirical (EMP) weights (illustrated in red), for endpoint responses. A visual weight approaching 1.0 signifies the dominance of visual perception in audiovisual responses. It is noteworthy that distinct weights are observed for each effector.

resolution of 4.7° could be attained under these experimental circumstances. Based on our findings, we reject hypothesis I, as the Mean Absolute Error observed across all participants for both head responses (4.9°) and arm responses (5.6°) exceeded the angle between neighboring target positions (4.7°).

For peripheral locations (±33° and ±28°) participants consistently undershot the target location whilst for the other locations they overshot it (fig. 7). This resulted in responses clustered around 25° for target locations between ±20° and ±33°. This clustering of responses can, in hindsight, be attributed to the approach used for simulating ILDs. When assessing the volume differences between the right and left ears across a 90-degree range, these differences exhibit a sigmoidal curve [22]. At angles of ±90° (locations aligned with the aural axis), the maximum volume difference is obtained. For angles in close proximity to 90°, human perception struggles to distinguish between these locations [22]. Notably, a linear relationship between the left and right ear volumes is discernible for target locations within ±33° from the centerline. In our experiment, the maximum volume difference was achieved at 33°, resulting in a compressed sigmoid response pattern and the clustering of target locations around the angle corresponding to the maximum volume difference.

This clustering phenomenon aligns with the feedback obtained from participants, who reported that they could clearly distinguish the difference between left and right in terms of volume, but in terms of localization, they perceived the sound as centered or leaning toward the sides.

While the clustering phenomenon is a byproduct of our experimental setup, we have identified a distinctive pattern of both overshooting and undershooting in head responses and arm responses. The observed tendency for locations to shift towards more peripheral positions could potentially be attributed to subtle temporal disparities between the speakers. These timing differences might lead to the auditory signals from the speakers being perceived not as a single sound source, but rather as distinct sources. Consequently, participants might have responded to the source that appeared louder in such instances [28], [54]. Furthermore, it's worth noting that the gaze-orienting system becomes underdamped for larger eccentricities, creating an overshoot in orienting responses [55].

Given this insight, one might anticipate that all peripheral target locations would be drawn toward the sides. However, intriguingly, for peripheral angles (±33° and ±28°), participants exhibited an undershooting tendency for target locations. This could potentially be associated with our experimental setup, wherein the maximum achievable angle corresponded to the outer boundaries of the screen. Resulting in participants being unable to overshoot for the most peripheral target locations. In the case of arm responses, there was a physical constraint in place — specifically, the joystick's projection was confined

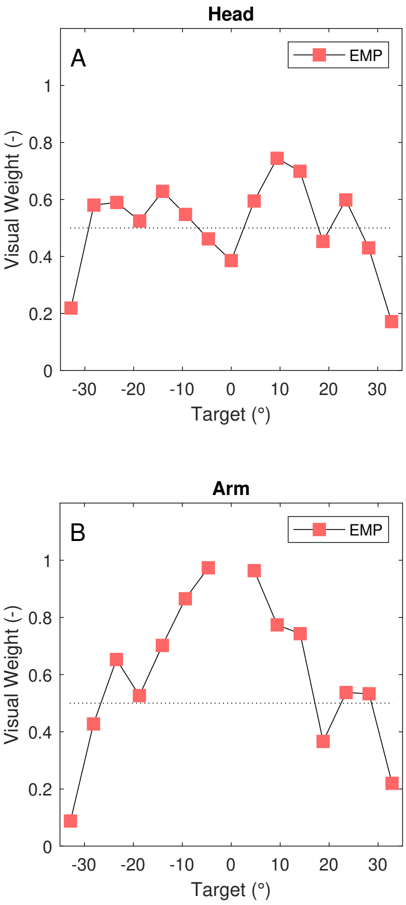


Figure 17: Visual (V) weights, derived using the empirical (EMP) model (illustrated in red) for reaction times. A visual weight approaching 1.0 signifies the dominance of visual reaction time in the audiovisual responses. Note the difference in weights between the effectors.

within the screen's boundaries. In contrast, head responses were not participant to such limitations. Nevertheless, participants were instructed to localize the sounds on the screen directly in front of them, effectively imposing the screen's edges as the maximum attainable angle. When coupled with the challenge of distinguishing between target locations within this range, this feature of the setup could potentially explain the observed undershooting tendencies.

In terms of elevation, the average response was clearly nonzero around the centerline (fig. 8), with pronounced variation across participants. This variability may arise from differences in the positioning of the aural axis in relation to the speaker setup. We used a height-adjustable chair to align participants via a projected laser with a central fixation target. Nonetheless, the natural posture of each individual may have impacted the execution of this alignment; for instance, certain participants exhibited a natural head position tilted forward. Interestingly, this prominent elevation perception was predominantly evident for locations near the centerline. Although elevation localization primarily relies on monaural cues, the perception of elevation from both ears must somehow be weighted [56]. Along the centerline, both sound sources exert substantial influence, making subtle timing and volume disparities more pronounced. This dynamic generates a diffuse percept characterized by distinct and variable elevation perceptions [54].

Overall, the slightly higher error observed in arm responses (MAE_{arm} = 5.6°, MAE_{head} = 4.9°) could be attributed to the intrinsic characteristics of the arm as an effector. The complexity of arm responses might result in less precise motions. Additionally, this discrepancy could arise from controlled element dynamics, where the higher gain setting led to more pronounced overshoot corrections, elaborately discussed in Appendix C. That is, participants had to learn the joystick dynamics whereas they have life-long experience moving their head. Moreover, as we will discuss in section IV-B, the necessary coordinate transformations differ between the head and arm with regard to localizing sounds, which introduces additional error sources.

For future research, it is advisable to improve the linear scaling of ILDs. If speakers remain in the same position, this adjustment would ensure that the maximum level difference is presented at a 90-degree angle, and volumes are scaled correspondingly for targets between ±33°. Alternatively, speakers could be positioned in line with the aural axis of the participant (at a 90° angle). Moreover, participants should have the ability to exceed the maximum target angle. This adjustment would facilitate a more in-depth exploration of our observations related to undershooting.

With regard to future experiments using this setup for the simulation of moving sounds, we did not explicitly test whether participants would be able to perceive a continuous sound trajectory. While localization behavior was fair across participants (fig. 7) the observed mean error suggests that participants may not necessarily perceive a continuous linear trajectory, but may instead perceive a wavy trajectory with significant elevation perception. However, it is also possible that the brain relies on predictive modeling to anticipate sound trajectory, as has been demonstrated for visual pursuit [57]. In that case, the brain may smooth out the perceived trajectory allowing for good tracking of the moving sound. Future experiments will need to address this open question.

Should it not be possible to achieve the perception of a smooth sound trajectory with the current setup, an alternative method to convey auditory stimuli is by employing headphones to replicate acoustic cues. Virtual acoustic space (VAS) stimulation is founded on the principle that the auditory system primarily relies on the acoustic pressure at the eardrums for spatial perception and linear filters can be utilized to replicate the sound wave-to-eardrum transformation [26]. However, VAS stimulation proves especially effective when non-acoustic signals, such as eye and head movements, can be effectively excluded from the perceptual task or are compensated for [18]. In an airplane cockpit, headphones could serve as an alternative means to enhance sound localization [58]. Systems that combine and align visual stimuli via a head-up display, sound via headphones, and head and eye tracking, are worth exploring and may provide the means to facilitate time-sensitive decisions or focus attention in aviation [14], [16], [42].

Lastly, it is important to acknowledge that for facilitating audiovisual interaction, the spatial percepts of the visual and auditory components of a stimulus should ideally fall within a 6° of each other [31]. This implies that our localization setup provides an acceptable level of accuracy for facilitating audiovisual interactions. We will address this question in the next section.

B. Multisensory benefits

For head-localization tasks in anechoic rooms, reduced endpoint variance and reduced reaction time have been observed for audiovisual responses compared to unimodal audio and visual responses [18]. This is based on the best-of-both-worlds principle, where the enhanced precision and reduced variability of visual responses complement the rapid reaction times characteristic of audio responses in audiovisual responses [18], [31].

Here we tested if these observations also hold with simulated spatial cues, i.e., ILDs, in an echoic room. As stated in Hypothesis IIa, we expected to observe reduced endpoint variability and reduced reaction times in the responses to multimodal stimuli as opposed to unimodal stimuli. In terms of endpoint variances, clear benefits were observed in unimodal audio responses compared to multimodal responses, both for head ($\sigma^2_{head} = 50^{\circ 2}$) and arm responses ($\sigma^2_{arm} = 44^{\circ 2}$). When comparing unimodal visual responses to multimodal responses, variances were equivalent for head movements, however, for arm movements, the visual variance was $6^{\circ 2}$ lower compared to the variance of audiovisual responses.

These differences in benefit for auditory and visual responses are in congruence with the best-of-both-worlds principle. The variability of endpoints for visual responses is smaller compared to the variability of endpoints for auditory responses (fig. 10). Therefore, any benefit of combining auditory and visual stimuli will be smaller with unimodal visual stimuli and larger with unimodal auditory stimuli (fig. 11). These results seem to indicate that audiovisual responses were dominated by the visual stimulus. We will return to this below when discussing the results of the unimodal weights.

Moreover, it is also well known that audiovisual interaction can lead to "negative benefit", i.e., an increase in variance and reaction time compared to unimodal responses [36], [37]. This phenomenon arises when incongruent multimodal stimuli, either spatially or temporally misaligned, are presented. As for some participants we observed errors larger than 6° (fig. 7), being the maximum misalignment for audiovisual interactions [31], we cannot rule out that for some participants auditory and visual stimuli were insufficiently aligned. As a result, some participants, showed higher variances for audiovisual responses compared to unimodal visual responses, indicating that adding auditory stimuli increased variance. These findings are consistent with experiments involving spatially misaligned targets conducted in anechoic environments [31].

Next, we address the comparison between unimodal and multimodal reaction times. It has been established that responses to unimodal audio stimuli are faster than responses to unimodal visual stimuli [5], [18]. Regarding reaction times for head responses, we observed a significant difference between unimodal visual reaction times and multimodal reaction times. Although for most participants (9/12) audio unimodal reaction times were faster than multimodal reaction times, the difference for all participants was not significant. Moreover, there is a decrease in reaction time observed when analyzing the average median reaction times. When comparing reactions to auditory stimuli, audiovisual stimuli led to a reduction in reaction times by 58 ms, whereas for visual stimuli, reaction times were shortened by 35 ms. When considering arm responses, both unimodal audio and visual reaction times were slower compared to reactions in the multimodal setting, resulting in an average median reaction time that was 30 ms slower. These findings are in congruence with Hypothesis II and the literature [18]. The similarity between reaction times for auditory and audiovisual stimuli in head-orienting responses, suggests that reaction times may have been dominated by the auditory stimuli in the audiovisual responses. This would be in congruence with the best-of-both-worlds principle. That is, audiovisual reaction times were dominated by audition, and as discussed above, endpoint variability by vision.

In order to assess if the speed-up of reaction times with audiovisual stimuli is due to audiovisual integration as opposed to statistical facilitation, we compared audiovisual responses with race model predictions [50], [59]. The presumption is that if audiovisual reaction times are faster than the model prediction, the reaction times are caused by processes other than simple statistical facilitation. If audiovisual reaction times are slower than the prediction, negative interactions between the modalities may take place, caused by other cognitive processes such as being distracted by either of the two modalities. Note that the comparison with the race model, therefore, does not attempt to describe and explain the mechanisms underlying audiovisual integration. Based on the spatial misalignment discussed, remarkably, a majority of participants showed a negative violation of the race model for head responses (6 out of 12) and arm responses (8 out of 12), implying the occurrence of audiovisual integration. However, this violation was only observed for a subset of reaction times. It is important to note that for this analysis we pooled reaction times across all target locations whilst reaction times may be location-dependent. Due to the limited number of trials per target location (<10), we were unable to undertake this analysis on a target-by-target basis. Given that auditory localization behavior differed clearly per target location (fig. 7 and fig. 8), we would expect to find location-dependent audiovisual benefits. Future studies should increase the number of repetitions per target location to add statistical significance to this analysis.

In conclusion, we partially confirm Hypothesis IIa. We found that reaction times were lower for multimodal responses compared to unimodal responses. Additionally, there was an average decrease in variance for multimodal responses compared to unimodal auditory responses. Head responses showed an average variance reduction of ($40^{\circ 2}$) and arm responses displayed a reduction of ($54^{\circ 2}$). However, no significant advantage was observed for all participants in terms of reduced variance for unimodal visual responses, leading to hypothesis IIa being only partly accepted.

Moreover, we evaluated the effector dependencies of these audiovisual interactions. Hypothesis IIb stated that audiovisual interactions are independent of the effector, meaning that potential reductions in reaction time and variance are similar for head responses and arm responses. We applied linear regression analysis to evaluate and contrast the audiovisual benefits of head-orienting and arm-orienting responses. The audiovisual benefit is defined as the difference between multimodal and unimodal response in terms of endpoint variances (fig. 11) and median reaction times (fig. 13). The comparison of the audiovisual benefit, specifically for unimodal visual median reaction times, among head-orienting and arm-orienting responses resulted in a low coefficient of determination (R^2 = 0.28). Similarly, when evaluating the audiovisual benefit of unimodal visual responses in terms of variance, a low coefficient of determination was observed (R^2 = 0.51). On the contrary, stronger correlations were apparent for the audiovisual benefits in the context of unimodal audio (R^2 = 0.74) median reaction time and unimodal visual variance (R^2 = 0.76) when comparing head and arm responses. Particularly with regard to unimodal visual stimuli, the audiovisual benefits demonstrate a dependence on the effector employed, leading to the rejection of Hypothesis IIb.

As these results have not been shown in previous literature, we performed a more in-depth analysis of the effector dependency within audiovisual benefits. When presenting audiovisual stimuli, the brain needs to weigh each input in order to program an adequate goal-directed response. We analyzed this weighting for each target location. These weights can either be obtained by solving eq. (5), using endpoint variance, or by directly fitting eq. (6) to the endpoints of all modalities (fig. 16).

We performed both analyses for the endpoint variances per target locations pooled over all participants. The correlation between these two independent approaches was good for both head and arm responses (R^2_{head} = 0.98, R^2_{arm} = 0.90). Interestingly, we observed that for the majority of locations visual weights were larger ($\geq$ 0.6) than auditory weights, clearly indicating that vision dominated the audiovisual location percept. This was especially pronounced for central locations. Interestingly at the most peripheral locations, for head responses the visual weights were smaller than 0.3, i.e., the auditory weights were larger than 0.7. This finding is interesting as it suggests that at these peripheral locations, auditory input dominates visual input. We interpret these results to reflect the fact that for central locations vision, guided by the fovea, dominates our perception, whereas in the periphery, for which visual acuity is low, audition dominates and aids us in aligning our fovea with objects of interest [60].

Note that this dominance of audition over vision for peripheral locations is not observed for arm responses. One possible explanation for this finding may be the tasks themselves and the internal reference frames used to translate sensory input to motor output. While the auditory system uses a head-centered reference frame (Cartesian coordinates), as the spatial cues are linked to the two ears, vision uses an oculocentric reference frame (polar coordinates). Arm responses are programmed in the body-centered frame and hence not linked to reference frames involved in sensory processing. The manual task required the participants to align a visual marker with the perceived target location. This visual element results in the necessity to use the oculocentric, visual reference frame rather than the craniocentric frame. As a consequence, the visual input may have been weighed more strongly in the manual task compared to the head-localization task.

The unimodal weightings pertaining to audiovisual reaction times (fig. 17) appear to support our interpretation. While for head responses weights were on average 0.5, we observed slightly higher weights at more central locations compared to peripheral locations. Clearly, the auditory system played an equal role in generating reaction times to audiovisual targets. With arm responses, we observed three distinct regions. For locations between $\pm 10°$ vision clearly dominated (visual weights $\approx$ 0.9), at intermediate locations (28°-13°) vision and the auditory system were about equal, and for the most peripheral locations tested the auditory system clearly dominated (visual weights $\approx$ 0.2). The latter is clearly in congruence with the idea that audition guides visual attention for peripheral locations [60]. Interestingly, we did not find this dominance for head responses, which is surprising since this was the case for the endpoint variance. It is not clear what caused this discrepancy, but timing issues with the head tracker in our setup may be the cause. Generally, in the used setup, we deem the extraction of response endpoints for head responses more reliable than the extraction of reaction times. The timing issues are elaborately discussed in Appendix A. Note that the arm responses do not suffer from this issue. Future experiments will need to validate the accuracy of the weights associated with head responses as presented in this study.

Be that as it may, these intriguing findings deserve further study. It is remarkable that given the same sensory input, the intended effector influences the weighing of sensory information. This seems to suggest that the system takes effector properties, such as inherent noise and required coordinate transformations, into account, when deciding how to optimally fuse information from individual modalities. In that regard, our results may be useful for the development of a sensor fusion algorithm. It may be fruitful to not only use the information on the signals that need to be fused, but also the properties of the output system when deciding how to weigh individual information channels.

In summary, our research has showcased the potential of audiovisual cues to enhance localization performance within an echoic environment. We showcased the similarity of localization behavior in an echoic room compared to an anechoic environment, highlighting the potential role of directional sound cues in real-life applications. Importantly, our observations on audiovisual benefits encompassed not only head movements but also extended to arm-localization tasks. This expansion of applicability underscores the range of activities that can derive benefits from audiovisual cues, including activities like driving and steering aircraft. Through the utilization of simulated sound sources, we have expanded the applicability of audiovisual cues to motor tasks that lack the presence of headphones and tasks that involve head and eye movements, as often encountered in real-world scenarios. Notably, audiovisual cues could significantly impact time-sensitive decision-making, offering practical implications like the incorporation of directional auditory alarms to ensure immediate visual orientation to critical matters [13], [15], [58].

Additionally, the resemblance in sound localization behavior between our results and experiments conducted in anechoic rooms highlights the possibility of conducting experiments for continuous sound localization [28]. Hence, this study serves as an initial stage in exploring audiovisual benefits in continuous tracking tasks which provides a foundation for further exploring how audiovisual cues can enhance performance and promote improved learning effects.

V. Summary and conclusions

This paper presents research on the potential of audiovisual benefits in head and arm localization tasks performed in an echoic room. Despite the mean absolute error (MAE) being (MAE_{head} = 4.9°, MAE_{arm} = 5.6°) larger than the offset between adjacent target positions (4.7°), the auditory and visual stimuli were sufficiently aligned to elicit audiovisual interactions. We observed reduced reaction times in audiovisual stimuli when compared to unimodal stimuli. In head responses, reaction times were reduced by 58 ms for visual stimuli and by 35 ms for auditory stimuli. Similarly, in arm responses, reaction times were shortened by 30 ms for both unimodal conditions. Furthermore, we observed reduced variances in audiovisual responses in comparison to unimodal auditory responses, with head responses showing a reduction of ($50^{\circ 2}$) and arm responses displaying a reduction of ($44^{\circ 2}$). The benefits compared to unimodal visual responses were participant-dependent. Most interestingly, we observed a modality-effector dependency in the weighting of the auditory and visual percept in the audiovisual responses. This suggests that the brain takes into account effector properties when deciding how to optimally fuse the unimodal percepts.

Overall, the ability to localize sound in an echoic environment combined with the audiovisual benefits observed in arm movements shows the potential of audiovisual benefits for motor tasks in real-life applications.

References

[1] F. Alex, R. A. Peters, and R. Stapleford, "Experiments and a model for pilot dynamics with visual and motion inputs," NASA, Tech. Rep., 1969.

[2] R. Hosman and H. Stassen, "Pilot's perception and control of aircraft motions," *IFAC Proceedings Volumes*, vol. 31, no. 26, pp. 311–316, 1998.

[3] S. C. F. Vrouwenvelder, "A cybernetic approach to point-to-point movements," Delft University of Technology, Tech. Rep., 2020.

[4] H. C. Hughes, P. A. Reuter-Lorenz, G. Nozawa, and R. Fendrich, "Visual-auditory interactions in sensorimotor processing: Saccades versus manual responses.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 20, no. 1, p. 131, 1994.

[5] M. M. Van Wanrooij, P. J. P. Bremen, and A. J. Van Opstal, "Acquired prior knowledge modulates audiovisual integration," *European Journal of Neuroscience*, vol. 31, no. 10, pp. 1763–1771, 2010.

[6] B. E. Stein and M. A. Meredith, "Multisensory integration," *Annals of the New York Academy of Sciences*, vol. 608, no. 1, pp. 51–70, 1990.

[7] R. Magill and D. I. Anderson, *Motor learning and control*. McGraw-Hill Publishing New York, 2010.

[8] G. Rosati, F. Oscari, S. Spagnol, F. Avanzini, and S. Masiero, "Effect of task-related continuous auditory feedback during learning of tracking motion exercises," *Journal of neuroengineering and rehabilitation*, vol. 9, no. 1, pp. 1–13, 2012.

[9] E. W. Vinje and E. T. Pitkin, "Human operator dynamics for aural compensatory tracking," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-2, no. 4, pp. 504–512, 1972. DOI: 10.1109/TSMC.1972.4309160.

[10] B. B. Johansson, "Multisensory stimulation in stroke rehabilitation," *Frontiers in human neuroscience*, vol. 6, p. 60, 2012.

[11] M. H. Thaut, G. C. McIntosh, R. R. Rice, R. A. Miller, J. Rathbun, and J. Brault, "Rhythmic auditory stimulation in gait training for parkinson's disease patients," *Movement disorders: official journal of the Movement Disorder Society*, vol. 11, no. 2, pp. 193–200, 1996.

[12] P. Bazilinskyy, L. van der Geest, S. van Leeuwen, B. Numan, J. Pijnacker, and J. de Winter, "Blind driving by means of auditory feedback," *IFAC-PapersOnLine*, vol. 49, no. 19, pp. 525–530, 2016.

[13] R. Sigrist, G. Rauter, R. Riener, and P. Wolf, "Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review," *Psychonomic Bulletin and Review*, vol. 20, pp. 21–53, 1 2013, ISSN: 10699384. DOI: 10.3758/s13423-012-0333-8.

[14] R. D. Sorkin, F. L. Wightman, D. S. Kistler, and G. C. Elvers, "An exploratory study of the use of movement-correlated cues in an auditory head-up display," *Human Factors*, vol. 31, no. 2, pp. 161–166, 1989.

[15] S. Carlile, B. Shinn-Cunningham, and A. Kulkarni, "Recent developments in virtual auditory space," *Virtual Auditory Space: Generation and Applications*, pp. 185–243, 1996.

[16] D. R. Begault, "Head-up auditory displays for traffic collision avoidance system advisories: A preliminary investigation," *Human factors*, vol. 35, no. 4, pp. 707–717, 1993.

[17] A. W. Bronkhorst, J. Veltman, and L. Van Breda, "Application of a three-dimensional auditory display in a flight task," *Human factors*, vol. 38, no. 1, pp. 23–33, 1996.

[18] A. J. Van Opstal, *The auditory system and human sound-localization behavior*. Academic Press, 2016.

[19] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[20] J. C. Makous and J. C. Middlebrooks, "Two-dimensional sound localization by human listeners," *The journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2188–2200, 1990.

[21] C. J. Plack, *The sense of hearing*. Routledge, 2018.

[22] M. M. Van Wanrooij and A. J. Van Opstal, "Contribution of head shadow and pinna cues to chronic monaural sound localization," *Journal of Neuroscience*, vol. 24, no. 17, pp. 4163–4171, 2004.

[23] J. C. Middlebrooks, "Sound localization," *Handbook of clinical neurology*, vol. 129, pp. 99–116, 2015.

[24] E. A. Macpherson and J. C. Middlebrooks, "Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited," *The Journal of the Acoustical Society of America*, vol. 111, no. 5, pp. 2219–2236, 2002.

[25] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1648–1661, 1992.

[26] F. L. Wightman and D. J. Kistler, "Headphone simulation of free-field listening. i: Stimulus synthesis," *The Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 858–867, 1989.

[27] P. J. P. Bremen, *Woods hole summer course: 'biology of the inner ear - systems part 3*, 2022.

[28] N. Niehof, M. M. Van Wanrooij, and A. J. Van Opstal, "Dynamic auditory localisation: Head tracking of virtual moving sounds," *Proceedings of the Master's Programme Cognitive Neuroscience, Nijmegen*, vol. 10, pp. 33–48, 2014.

[29] M. A. Meredith and B. E. Stein, "Spatial factors determine the activity of multisensory neurons in cat superior colliculus," *Brain research*, vol. 365, no. 2, pp. 350–354, 1986.

[30] M. A. Meredith, J. W. Nemitz, and B. E. Stein, "Determinants of multisensory integration in superior colliculus neurons. i. temporal factors," *Journal of Neuroscience*, vol. 7, no. 10, pp. 3215–3229, 1987.

[31] D. Alais and D. Burr, "The ventriloquist effect results from near-optimal bimodal integration," *Current biology*, vol. 14, no. 3, pp. 257–262, 2004.

[32] M. O. Ernst and M. S. Banks, "Humans integrate visual and haptic information in a statistically optimal fashion," *Nature*, vol. 415, no. 6870, pp. 429–433, 2002.

[33] B. D. Corneil, M. M. Van Wanrooij, D. P. Munoz, and A. J. Van Opstal, "Auditory-visual interactions subserving goal-directed saccades in a complex scene," *Journal of Neurophysiology*, vol. 88, no. 1, pp. 438–454, 2002.

[34] M. M. Van Wanrooij, A. H. Bell, D. P. Munoz, and A. J. Van Opstal, "The effect of spatial–temporal audiovisual disparities on saccades in a complex scene," *Experimental brain research*, vol. 198, pp. 425–437, 2009.

[35] "Visual-auditory interactions modulate saccade-related activity in monkey superior colliculus," *Brain research bulletin*, vol. 46, no. 3, pp. 211–224, 1998.

[36] L. K. Harrington and C. K. Peck, "Spatial disparity affects visual-auditory interactions in human sensorimotor processing," *Experimental Brain Research*, vol. 122, pp. 247–252, 1998.

[37] R. A. Stevenson, D. Ghose, J. K. Fister, *et al.*, "Identifying and quantifying multisensory integration: A tutorial review," *Brain topography*, vol. 27, pp. 707–730, 2014.

[38] H. H. Goossens and A. J. Van Opstal, "Human eye-head coordination in two dimensions under different sensorimotor conditions," *Experimental Brain Research*, vol. 114, no. 3, pp. 542–560, 1997.

[39] M. E. Berryhill, T. Chiu, and H. C. Hughes, "Smooth pursuit of nonvisual motion," *Journal of neurophysiology*, vol. 96, no. 1, pp. 461–465, 2006.

[40] M. P. Zwiers, A. J. Van Opstal, and J. R. M. Cruysberg, "A spatial hearing deficit in early-blind humans," 2001.

[41] M. P. Zwiers, A. J. Van Opstal, and G. D. Paige, "Plasticity in human sound localization induced by compressed spatial vision," *Nature neuroscience*, vol. 6, no. 2, pp. 175–181, 2003.

[42] D. R. Begault and M. T. Pittman, "Three-dimensional audio versus head-down traffic alert and collision avoidance system displays," *The International Journal of Aviation Psychology*, vol. 6, no. 1, pp. 79–93, 1996.

[43] B. Biguer, C. Prablanc, and M. Jeannerod, "The contribution of coordinated eye and head movements in hand pointing accuracy," *Experimental brain research*, vol. 55, no. 3, pp. 462–469, 1984.

[44] R. A. Schmidt, T. D. Lee, C. Winstein, G. Wulf, and H. N. Zelaznik, *Motor control and learning: A behavioral emphasis*. Human kinetics, 2018.

[45] S. C. Gielen, R. A. Schmidt, and P. J. Van Den Heuvel, "On the nature of intersensory facilitation of reaction time," *Perception & psychophysics*, vol. 34, pp. 161–168, 1983.

[46] M. Mulder, D. M. Pool, D. A. Abbink, *et al.*, "Manual control cybernetics: State-of-the-art and current trends," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 5, pp. 468–485, 2017.

[47] M. Van Paassen, O. Stroosma, and J. Delatour, "Dueca-data-driven activation in distributed real-time computation," in *Modeling and Simulation Technologies Conference*, 2000, p. 4503.

[48] L. D. Reid and M. A. Nahon, "Flight simulation motion-base drive algorithms: Part 1. developing and testing equations," *UTIAS Report, No. 296*, 1985.

[49] D. J. MacKay, "The evidence framework applied to classification networks," *Neural computation*, vol. 4, no. 5, pp. 720–736, 1992.

[50] J. Miller, "Divided attention: Evidence for coactivation with redundant signals," *Cognitive psychology*, vol. 14, no. 2, pp. 247–279, 1982.

[51] E. Fehrer and D. Raab, "Reaction time to stimuli masked by metacontrast.," *Journal of experimental psychology*, vol. 63, no. 2, p. 143, 1962.

[52] M. Gondan and K. Minakata, “A tutorial on testing the race model inequality,” *Attention, Perception, & Psychophysics*, vol. 78, pp. 723–735, 2016.

[53] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[54] G. S. Kendall, “The decorrelation of audio signals and its impact on spatial imagery,” *Computer Music Journal*, vol. 19, no. 4, pp. 71–87, 1995.

[55] C. M. Harris and D. M. Wolpert, “The main sequence of saccades optimizes speed-accuracy trade-off,” *Biological cybernetics*, vol. 95, no. 1, pp. 21–29, 2006.

[56] M. M. Van Wanrooij and A. J. Van Opstal, “Relearning sound localization with a new ear,” *Journal of Neuroscience*, vol. 25, no. 22, pp. 5413–5424, 2005.

[57] G. R. Barnes, “Cognitive processes involved in smooth pursuit eye movements,” *Brain and cognition*, vol. 68, no. 3, pp. 309–326, 2008.

[58] D. Begault, E. M. Wenzel, M. Godfroy, J. D. Miller, and M. R. Anderson, “Applying spatial audio to human interfaces: 25 years of nasa experience,” in *Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space*, Audio Engineering Society, 2010.

[59] D. H. Raab, “Statistical facilitation of simple reaction times.,” *Transactions of the New York Academy of Sciences*, 1962.

[60] H. E. Heffner and R. S. Heffner, “The evolution of mammalian hearing,” in *AIP Conference Proceedings*, AIP Publishing, vol. 1965, 2018.

II

Preliminary Report

Previously graded under AE4020

Introduction

While multisensory benefits have been extensively studied in tasks involving quick eye and head movements, their application in more complex motor tasks remains understudied. Visual and haptic cues have long been recognized as playing a key role in the execution of motor tasks. However, the potential benefits of audio cues in complex motor tasks have been largely ignored despite their potential application in various human-machine interaction scenarios.

Recent studies have suggested that audio cues can be used to support blind driving, increase learning during rehabilitation therapy, or increase performance in complex motor tasks [2] [22] [44]. Most of these studies have been focused on performance increases in real-life situations but are limited in their fundamental explanation of how this increased performance is achieved. Additional auditory cues have been shown to increase situational awareness and reduce workload for pilots [46]. For example, directional audio cues can provide pilots with additional information on spatial localization, which is especially useful in fighter jets where visual cues might be misleading and lead to spatial disorientation [45].

Bending and reflection of sound waves are used by humans to localize sounds. Interaural Level Differences (ILDs) in audio cues are a commonly studied and simple technique to create directional audio cues. Much research has been performed on the localization of audio cues in anechoic rooms. However, in echoic rooms, localization might be less accurate due to reflections. Most of these studies involve rather complex setups that are difficult to use in real-life situations.

This report aims to explore the audiovisual benefits of a discrete manual control task focusing on directional audio cues created using Interaural Level Differences (ILDs) by answering the following research question:

> What audiovisual benefits can be observed in a discrete manual localization task providing multisensory cues in an echoic environment?

In this report first, a literature overview will be provided which starts with a chapter on audiovisual benefits, this includes a review on audiovisual benefits in motor tasks as well as a review of fundamental research performed on this topic. As this report focuses on a localization task, an explanation of sound localization principles will be given in chapter 2. In chapter 3, an elaboration of the sensory processing mechanisms and creation of motor commands is presented. After that, in chapter 4, the results and conclusions of the preliminary results are provided. Lastly, the experiment proposal and an explanation of the proposed data analysis are given in chapter 5.

1
Audiovisual benefits

When observing the world around us multiple senses are triggered. Hence, the human brain is skilled in combining all information received, a process referred to as multisensory processing. In this chapter, first, a general introduction is given to multisensory benefits in motor tasks. After that, different methods to quantify multisensory benefits are explained.

1.1. Audiovisual benefits in motor tasks

During a motor task, multiple types of feedback can be provided to improve performance such as visual, auditory, haptic, and multimodal feedback. Visual feedback is important in, for example, eye-hand coordination tasks, such as catching a ball [25]. However, introducing other modes of feedback, in which feedback refers to externally created stimuli that provide information on the movement, can also have benefits. When considering auditory feedback three categories can be defined: auditory alarms, movement sonification, and error sonification [44]. Sonification refers to the use of non-speech sounds to convey information to a listener by changing sound characteristics such as frequency and amplitude. For aerospace applications only auditory alarms are used, as during flight sound is already used as the main method of communication between e.g. pilots and air traffic control. These auditory alarms do not contain any spatial information but are meant to make the pilot aware of certain warnings. An additional continuous acoustic cue would result in an overload of information on the auditory 'channel'. However, directional audio cues have been used in static localization tasks to reduce reaction times and increase accuracy in localization for fighter pilots [46]. Multimodal feedback is most relevant in complex tasks as the distribution of information leads to a decrease in workload [44].

Various research has been performed on the addition of audio cues to increase motor performance or motor learning. The scope of introducing feedback goes beyond increased performance and learning effects, as it can also enhance compensatory mechanisms to overcome the loss of motor functions [44] [23]. A well-studied area is the use of auditory feedback in rehabilitation therapy [22]. It has been shown that increased learning behavior is observed even when additional feedback is removed. In some specific conditions adding extra sensory information can lead to a reduction in workload, for example in Parkinson's patients when external feedback is provided by rhythmic auditory stimuli [50]. Also, audio feedback has been proven to be successful in a blind driving experiment in which subjects had to follow a path based solely on auditory feedback [2].

1.2. Measures for audiovisual benefits

The aforementioned research focused mostly on improving performance for complex motor tasks but, did not extend to underlying principles on why this feedback positively affected performance. However, underlying neural pathways and principles that enhance (or suppress) audiovisual benefits have been widely studied for simple motor responses [21] [52] [47]. Also, different types of measures have been developed to quantify the audiovisual benefits [49]. In this section, firstly, an introduction will be provided on measures based on the reaction of single cells. After that, different behavioral measures for audiovisual benefits will be explained.

1.2.1. Neural processing

When examining the processing of sensory inputs, it is important to differentiate between two types of stimuli: unimodal and multimodal stimuli. Unimodal stimuli involve the presentation of a single sensory input, while multimodal stimuli involve the presentation of stimuli that are processed by different sensory channels. The phenomenon of multisensory integration, which involves the combination of information from multiple sensory modalities, was initially investigated by Woodworth [61]. The response to multimodal stimuli can be influenced by temporal and spatial factors, either enhancing or suppressing the response [28] [27]. In cases of enhancement, firing rates for multimodal stimuli are greater than those for unimodal stimuli, as depicted in fig. 1.1 (left side) [49]. Meredith and Stein introduced the principle of superadditivity, also known as multisensory integration, which suggests that the response to multisensory signals is greater than the sum of the individual unisensory responses [47], which is shown in fig. 1.1 (lower left side) [49]. When no integration occurs, the firing rate for multimodal stimuli is equal to the maximum firing rate observed for either the unimodal audio or visual response, as shown in fig. 1.1 (top right). On the other hand, a response can be suppressed, resulting in a firing rate for multimodal stimuli that is even lower than the maximum firing rate observed for unimodal audio and visual stimuli. This is illustrated in fig. 1.1 (lower right).

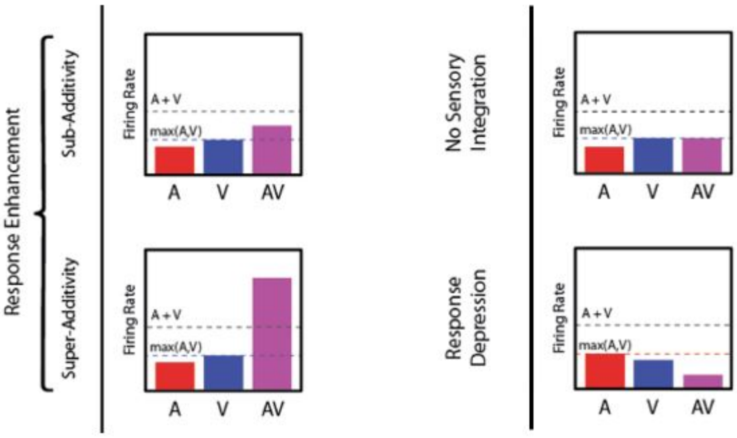


Figure 1.1: Firing rates of single neurons in response to audio (a), visual (v), and audiovisual (av) stimuli corresponding to different levels of audiovisual interaction. Dotted lines present the maximum firing rate of the responses to the unimodal audio and visual stimuli (max(A,V)) and the sum of these unimodal responses(A+V) [49].

Multisensory integration can be observed when the properties of stimuli promote the fusion of information from two modalities, resulting in the perception of a single object. Conversely, if the stimuli properties favor the dissociation of sensory information into two distinct objects, there is little to no response enhancement observed. The amount of response enhancement depends on three main principles [28]:

1. **Temporal proximity**: Events should be perceived in close temporal proximity to each other to achieve maximum benefit.
2. **Spatial proximity**: Events should be perceived in close spatial proximity to each other in order to achieve maximum benefit.
3. **Principle of inverse effectiveness**: When multimodal stimuli are presented near threshold values, enhancement is larger compared to when stronger stimuli are presented.

Perception is key for multisensory integration; stimuli do not have to be spatially close but have to be perceived as such for the brain to perceive a single object [11] [55]. To prove multisensory integration multiple variations must be tested that also show reduced enhancement when signals are spatially and temporally not aligned [14] [17]. Also, response enhancement should be small when at least one of the stimuli presented is strong, to prove the principle of inverse effectiveness.

1.2.2. Behavioral Measures

Multisensory interaction and multisensory integration are two related but distinct concepts in the field of sensory perception. Multisensory integration, discussed in the previous section, refers to the neural processes that occur in the brain to combine and integrate information from multiple sensory modalities [48] [58] [14].

Multisensory interaction, on the other hand, refers to the ways in which the presence of multimodal stimuli influences behavioral measures such as reaction time and accuracy. When this interaction leads to behaviorally measurable improvements, such as reduced reaction times and increased accuracy in localizing stimuli, it is termed an audiovisual benefit. Hence, multimodal stimuli can be presented resulting in measurable improvements without sensory information being integrated [49].

Most research concerning audiovisual benefits has been conducted using light flashes and short sounds (50-100 ms) [21] [52]. Figure 1.2 depicts a representative outcome from such an experiment, illustrating the relationship between reaction time (in milliseconds) and localization error (in degrees) for different types of stimuli [52]. In this study, head-orienting responses to unimodal visual targets (represented by red dots) and unimodal auditory targets (represented by blue squares), as well as multimodal audiovisual targets (represented by green diamonds), were measured. The ellipses displayed in fig. 1.2 represent the mean plus standard deviations of the localization error and reaction time for each stimulus category. Observations from this experiment indicate that, generally, reaction times for auditory stimuli are faster but less accurate compared to visual stimuli. This is apparent in fig. 1.2, where the blue circle (representing audio responses) appears larger along the y-axis compared to the red circle (representing visual responses) but is more shifted towards the right. Interestingly, the distribution of localization error for the multimodal stimuli (green circle) resembles that of the visual stimuli (red circle), while demonstrating much faster reaction times compared to the unimodal visual stimuli. Additionally, reaction times for the multimodal stimuli are faster than those for the unimodal auditory stimuli. These findings suggest the presence of audiovisual benefits when audiovisual stimuli are presented. Different methods have been developed to quantify the audiovisual benefits using behavioral measures of which detection rate, accuracy, and reaction time are discussed below.

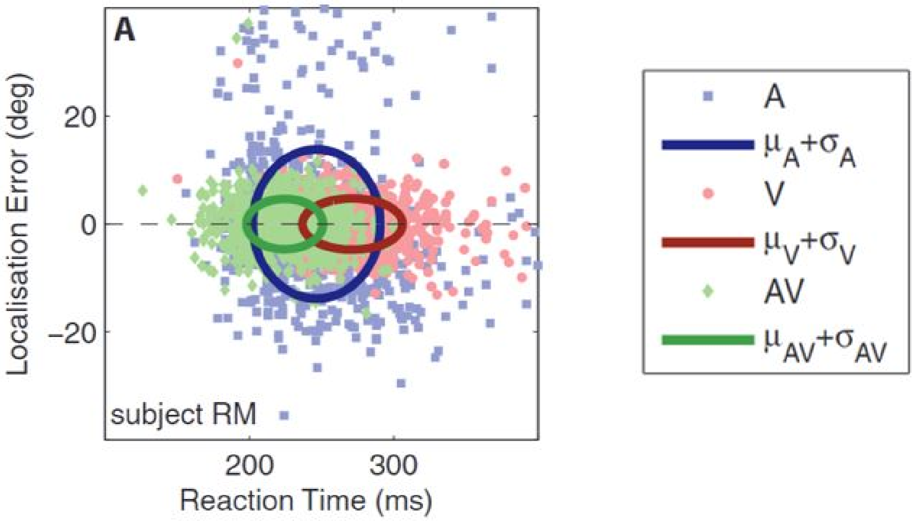


Figure 1.2: In this study, head-orienting responses were measured to unimodal visual targets (represented by red dots) and unimodal auditory targets (represented by blue squares), as well as multimodal audiovisual targets (represented by green diamonds). The corresponding ellipses show the mean plus standard deviation for the responses to specific stimuli. These results suggest that multimodal stimuli improve reaction times without compromising accuracy, highlighting the advantages of multimodal stimuli [52].

Detection rate

An intuitive way to assess the potential benefit of multimodal stimuli for the detection rate is by considering both unisensory responses and comparing those to the multimodal responses. This maximum criterion is defined in eq. (1.1) [49]. In which $\hat{p}(AV)$ is the detection rate for audiovisual stimuli and $p(A)$ and $p(V)$ are the detection rates for the unimodal audio and visual stimuli, respectively.

$$\hat{p}(AV) > \max[p(A), p(V)] \tag{1.1}$$

If eq. (1.1) holds, this only tells us that information from both senses is used but not if the information is also integrated. When multisensory integration occurs and independence between the detection rates is assumed, equation eq. (1.2) holds [49]. The right side of equation eq. (1.2) is the summed detection rate of the unimodal stimuli subtracted by the possibility of both stimuli being detected in the same trial.

$$\hat{p}(AV) > p(A) + p(V) - p(A) \cdot p(V) \tag{1.2}$$

In conditions where stimuli are presented at intensities close to the threshold of detection, the most multisensory benefit is observed following the principle of inverse effectiveness explained earlier.

Accuracy

Here, accuracy refers to how closely the participant's endpoints match the target location. A simple linear regression model can be used to quantify the target-response relation shown in eq. (1.3) [52]. In which the response angle is defined as a_r and the target angle is provided by a_t. The linear regression parameters a and b are solved by minimizing the mean squared error [35]. The localization error is defined as the value of the residuals between the data and the linear fit.

$$a_r = a \cdot a_t + b \tag{1.3}$$

Reaction times

A commonly known model to quantify audiovisual benefits for reaction times is the statistical facilitation method also referred to as the race model [31] [15] [13]. This method relies on the principle that the benefit of providing two stimuli results from the so-called race between different unimodal stimuli and that the first neuronal representation arriving at a decision stage elicits a reaction. By performing a simple statistical analysis it can already be concluded that when multimodal stimuli are presented, the reaction times will always be faster compared to unimodal stimuli even if no interaction takes place. For the race model, the cumulative distribution function (CDF) of the measured reaction times is compared to the upper bound of the race model, which is calculated using eq. (1.4). In this equation, $F_A(t)$ and $F_V(t)$ are the probabilities that the response started at time t for unimodal audio and visual stimuli, respectively. These probabilities multiplied is the probability that a response occurs at time t for both modalities.

$$F_{AV}(t) = F_A(t) + F_V(t) - F_A(t) \cdot F_V(t) < F_A(t) + F_V(t) \tag{1.4}$$

In fig. 1.3 the probability distribution of unimodal audio (red) and visual (blue) responses is provided combined with two examples of responses to multimodal stimuli (purple) [49]. The race model, created by selecting the fastest unimodal response for each trial, is shown in black. Violation of the race model as expressed in eq. (1.4), implies that the benefit cannot be attributed solely to one of the unimodal responses and multisensory interaction is taking place which is displayed in fig. 1.3b. The amount of enhancement is shown in fig. 1.3e. In fig. 1.3c the CDF for the second multisensory response is given, for these response no enhancements are observed and the CDF for the multisensory response stays below the race model for all responses. The difference between the multisensory response CDF and the race model is referred to as the Miller inequality and is shown in fig. 1.3e and fig. 1.3f [31]. Positive values indicate multisensory integration, however, negative values do not imply any type of interaction [31].

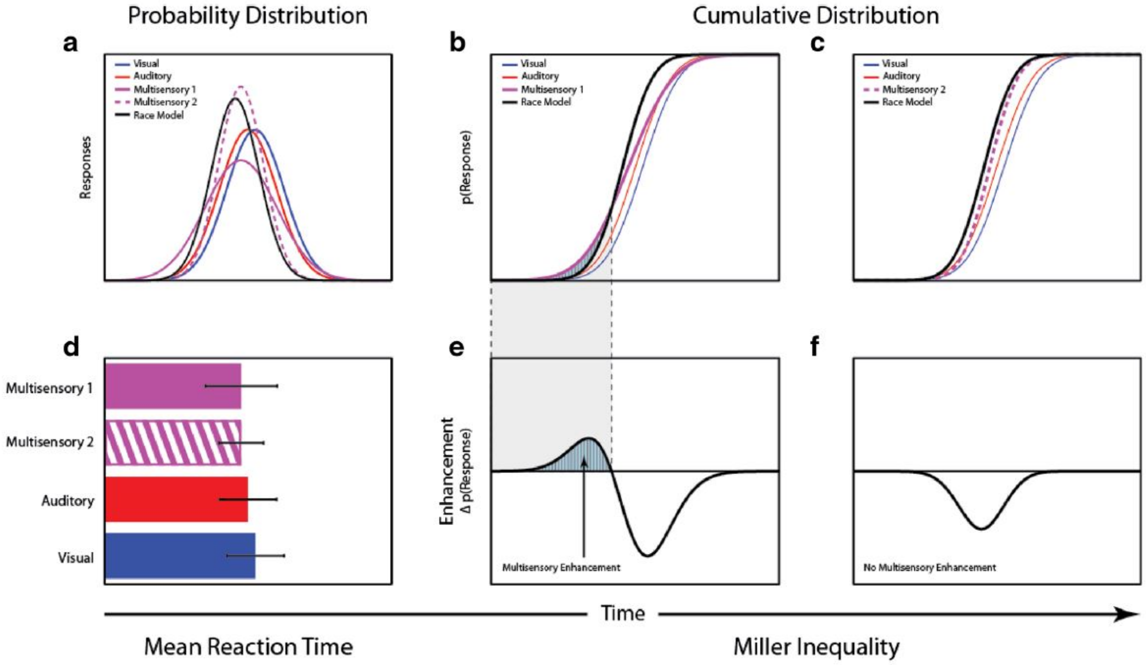


Figure 1.3: Mean reaction times of responses to unimodal stimuli as well as multimodal stimuli (d) and compared to the race model using a distribution of reaction times (a). The CDF of the multimodal responses is compared to the race model for two cases (b,c). Comparisons can be made using the Miller inequality which is the difference between the CDF of the responses and the (e,f). Positive enhancement implies audiovisual integration [49]

1.3. Conclusion

Numerous studies have provided evidence supporting the positive impact of incorporating audio feedback into complex motor tasks. Also, several studies focused on investigating the audiovisual benefits during simple motor tasks, such as eye and head movements. Specifically, when weak stimuli are aligned spatially and temporally, significant benefits can be observed, including reduced reaction time and increased task accuracy. To exploit the potential of additional audio feedback, further research is needed to examine its influence on more complex motor tasks.

2

Principles of spatial hearing

While the visual system uses direct one-to-one mapping to localize objects, the auditory system requires the involvement of both ears to construct a spatial representation of sound waves. This chapter starts with a brief introduction to the basic principles of sound, followed by an explanation of the anatomical structure of the ear. The working mechanisms involved in sound localization are then presented. Finally, the influence of the acoustic environment on sound localization as well as different techniques used to present audio stimuli in experimental setups are discussed.

2.1. The theory of sound

Sound is created by pressure differences which can be evaluated over time creating waves. Each pressure wave has several characteristics such as a period (T), frequency f, wavelength (λ), phase (ϕ), and amplitude (A) which are related by eq. (2.1).

$$y = A\,sin(2\pi f t + \phi) \tag{2.1}$$

The frequency of a wave is related to the period by eq. (2.2).

$$T = \frac{1}{f} \tag{2.2}$$

Sound is propagated in all directions and changes when it travels through different media and interacts with objects. The power of the sound spreads in a sphere that is centered around the source. When sound waves encounter a boundary they will be reflected or absorbed.

The reflection depends on the type of boundary the sound encounters and on the incidence angle. Impedance defines the resistance of a medium being moved [51]. Dense and stiff objects reflect the sound wave with higher energy compared to other objects. The energy that is not reflected by the object is transmitted through the object or absorbed. When the energy is transmitted the sound waves produce vibrations in the object. The energy that is absorbed by an object is converted to thermal energy. Lastly, sound waves can bend or diffract around objects. The amount of diffraction depends on the frequency of the sound; low-frequency sound waves diffract more compared to high-frequency sound waves [51]. If the object is smaller than the wavelength, the sound will be unaffected by the object. The wavelength λ is given by eq. (2.3). In which c is the speed of sound.

$$\lambda = \frac{c}{f} \tag{2.3}$$

The head can be described as a low-pass filter as it is acoustically transparent for $f < 2\ kHz$, assuming $\lambda \approx 17\ cm$, which is the diameter of a 'standard head' [51]. This is important in the context of binaural cues which is discussed section 2.3.

2.2. Anatomy of the ear

Sound plays a large role in communication. In contrast to vision, for which the field of view is limited, hearing allows us to monitor sounds all around us. The ear can be divided into three parts; the outer ear, middle ear, and inner ear, as shown in fig. 2.1 [19]. Sound waves impinge on the outer ear and reach the tympanic membrane via the external auditory meatus [7]. The middle ear performs impedance matching (air to salt water) and the inner ear can be conceived of as a frequency analyzer. Information on the decomposed sound is transmitted via individual frequency channels in the auditory nerve to the Central Nervous System (CNS) [51]. It is thought that the CNS reconstructs the sound from the information available in individual frequency channels. The result of this reconstruction ultimately leads to the perception of the sound by the listener.

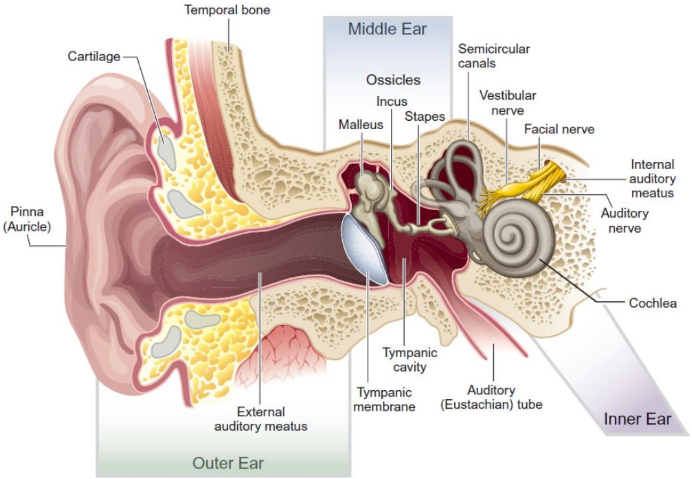


Figure 2.1: Overview of the anatomy of the ear consisting of three main parts: the outer ear, middle ear, and cochlea[19]

2.3. Sound localization principles

When discussing sound localization two types of directions are considered: the azimuth angle and elevation angle [51]. The azimuth angle is the horizontal plane through the head (left/right) and the elevation angle is defined as the vertical plane through the head (up/down), schematically shown in fig. 2.2 [51]. Different mechanisms are used for spatial localization in elevation and azimuth, which are discussed below.

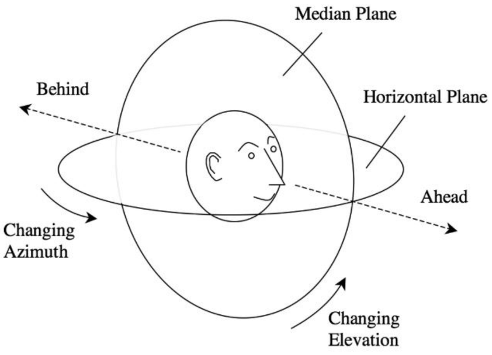


Figure 2.2: Schematic overview of azimuth and elevation direction used in sound localization [51]

2.3.1. Localization principles for azimuth

When examining sound localization in terms of azimuth (horizontal direction), two crucial principles must be addressed: interaural time differences (ITDs) and interaural level differences (ILDs). Early research on ITDs and ILDs was conducted by Rayleigh [38]. These concepts fall under the category of binaural listening, as both ears play a role in sound localization [51].

ITDs arise from variations in sound arrival time between the right and left ears. As depicted in fig. 2.3, when a sound source is located on the left (SL), the left ear receives the sound earlier than the right ear. The brain utilizes the phase difference and onset difference to determine the sound source location [51]. The brain uses phase locking, neuronal responses locked to a certain phase of a sine wave, to determine the time difference [51]. ITD processing is limited to sounds of approximately 1.5 kHz [38].

On the other hand, ILDs depend on the level differences between sounds reaching each ear. As illustrated in fig. 2.3, when a sound source is positioned on the left (SL), the sound waves arriving at the left ear (LL) are louder (indicated by a thicker wave). The head acts as a barrier, causing attenuation of sound waves traveling from the left to the right ear for higher frequencies [51].

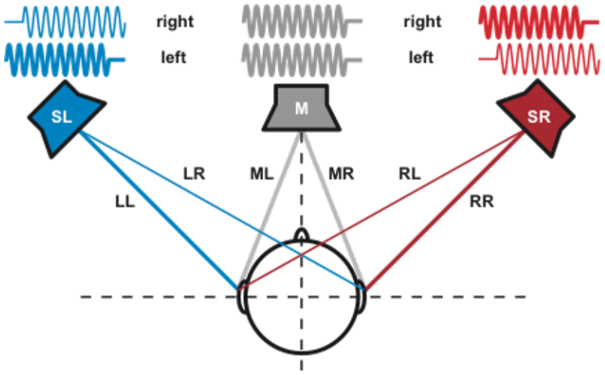


Figure 2.3: Schematic overview of interaural time and level differences for sound sources presented on the left (SL), middle (M), and right (SR). The thickness of the sound waves presents the sound level difference between the ears. Note that there is a phase and time of arrival differences for the sound sources presented on the right and the left. No time or level differences are encountered for the sound source in the middle (source: Peter Bremen)

This damping is referred to as the head shadow effect (HSE), and for broadband noise can be estimated by eq. (2.4). Equation 2.4 is plotted in fig. 2.4 and a linear relationship between the sound intensities perceived at each ear is observed between approximately -30 and 30 degrees in azimuth direction [53]. The level difference between the ears is used by the brain to identify the location of the sound [51]. For frequencies above 2-3 kHz, the range in which wavelengths of sounds are shorter than the size of the head, ILDs are more dominant in sound localization [30]. Later, this theory was confirmed for broadband noise [60]. Localization of sounds in the range 4-16 kHz is dominated by ILDs cues even if contradicting ITDs cues are presented [24].

$$\mathrm{HSE}(\alpha) = 10\sin(0.13\alpha + 0.3)\mathrm{dB} \tag{2.4}$$

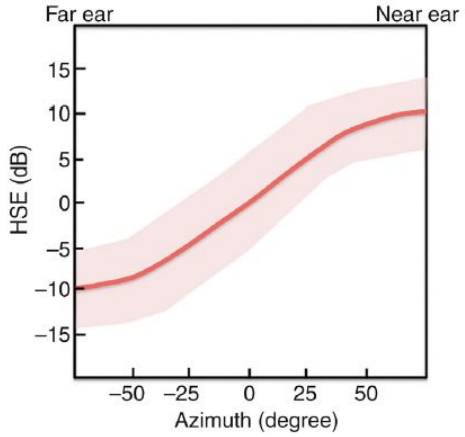


Figure 2.4: Measured level differences at two ears for a specific subject with a fitted curve according to eq. (2.4) [51]

2.3.2. Cone of confusion

In humans, the aforementioned mechanisms do not provide enough information to solve the localization problem, as multiple locations can cause the same ILDs or ITDs. This area is referred to as the cone of confusion [30]. An example of such an area is shown in fig. 2.5. In which d represents the distance from the sound source S to each ear, r the radius of the head, and $d \gg r,$ such that $\alpha \simeq \alpha_R \simeq \alpha_L$.

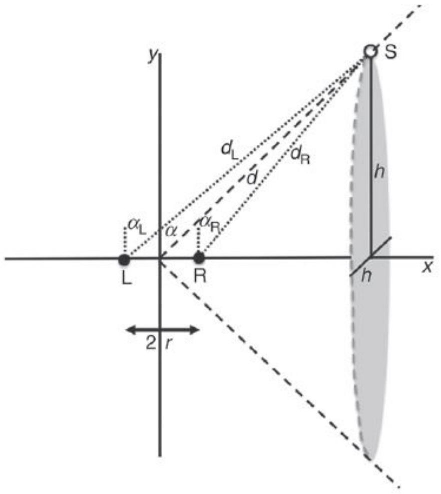


Figure 2.5: The cone of fusion consists of points with the same interaural differences in which d represents the distance from each ear to the sound source S, $2r$ presents the diameter of the head, and $d \gg r,$ such that $\alpha \simeq \alpha_R \simeq \alpha_L$[51]

Considering a symmetrical head, it can be deduced that at least one of these locations will be positioned at the midsagittal plane. Consequently, humans require additional spatial information to effectively resolve this localization problem.

2.3.3. Localization principles for elevation

Extra spatial information of the sound source is collected using the reflection of the sound waves on the body and specifically on the pinnae (part of the outer ear). The reflection of the sound waves depends on the angle of incidence and the shape of the pinnae, which is unique for everyone [51]. A schematic overview of these reflections is provided in fig. 2.6 [51]. As the shapes of the pinnae (head and torso) are very similar for the right and left ear, localization for elevation is essentially monaural [51].

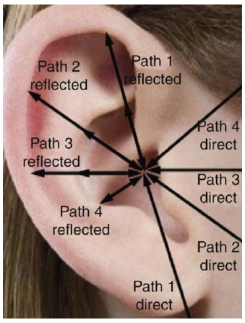


Figure 2.6: Schematic overview of the reflected paths in the pinna which works as a direction-dependent spectral filter [51]

The pinnae act as a filter that attenuates and amplifies frequencies for specific elevation. This filter is also referred to as the head-related transfer function or HRTF [59]. Such a filter is shown in fig. 2.7 for azimuth 0 degrees as a function of frequency and elevation [18]. Variation in attenuation and amplification is displayed using colors, in which dark colors present attenuation and light colors present amplification. The most noticeable characteristic of the filter is a linear increase in narrow regions of low acoustic energy, referred to as spectral notch, as the elevation increases for frequencies above 3-4 kHz. This is shown in fig. 2.7 by darker areas at higher elevations. Below 3-4 kHz limited differences are seen in filter characteristics for different elevations. Meaning that elevation perception is worse for low-pass signals compared to signals containing high-pass noise at azimuth 0.

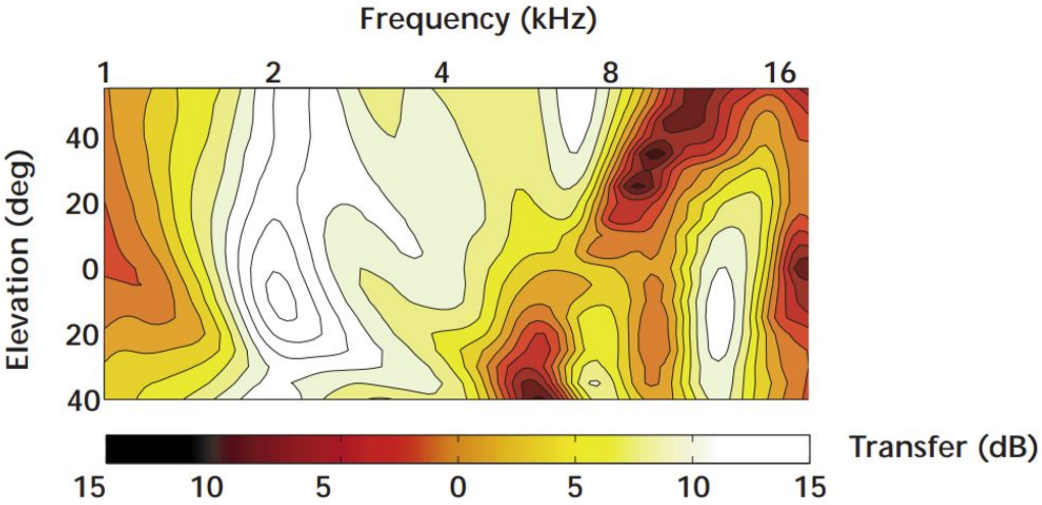


Figure 2.7: Pinna transfer function of the right ear for one subject. Transfer functions are shown as a function of frequency and elevation of the sound source. The amplitude (in dB) of the transfer function is represented by color. At zero dB, the sound pressure amplitude of a tone at a specific frequency and elevation is unaffected by the presence of the head and pinna. Light colors indicate sound amplification, while dark areas indicate sound attenuation. [18]

Even with this new source of information, the problem remains ill-posed, as the sensory spectrum at the eardrum defined as $S(f;\epsilon_T)$ consists of a multiplication of the direction-specific HRTFs $H(f;\epsilon_T)$ and the actual frequency spectrum of the sound source $T(F)$, which is unknown. This is summarized in eq. (2.5).

$$S\left(f;\varepsilon_T\right) = H\left(f;\varepsilon_T\right)\cdot T(f) \tag{2.5}$$

A model to solve this ill-posed problem was proposed by Ege [12] and is shown in fig. 2.8. The target location $T*$ consist of a specific azimuth $(\alpha*)$ and elevation $(\epsilon*)$ angle. As explained in the previous section, the azimuth location is determined using binaural listening principles (ILDs and ITDs), after which a maximum likelihood estimation is used (MLE). The first step in solving the ill-posed problem for response elevation localization is applying a cross-correlation. In the absence of noise, only peaks will be observed for the target elevation angle. Hence, the problem can also be presented as a maximum likelihood estimation (MLE) $(L(\epsilon|\epsilon_T)$.

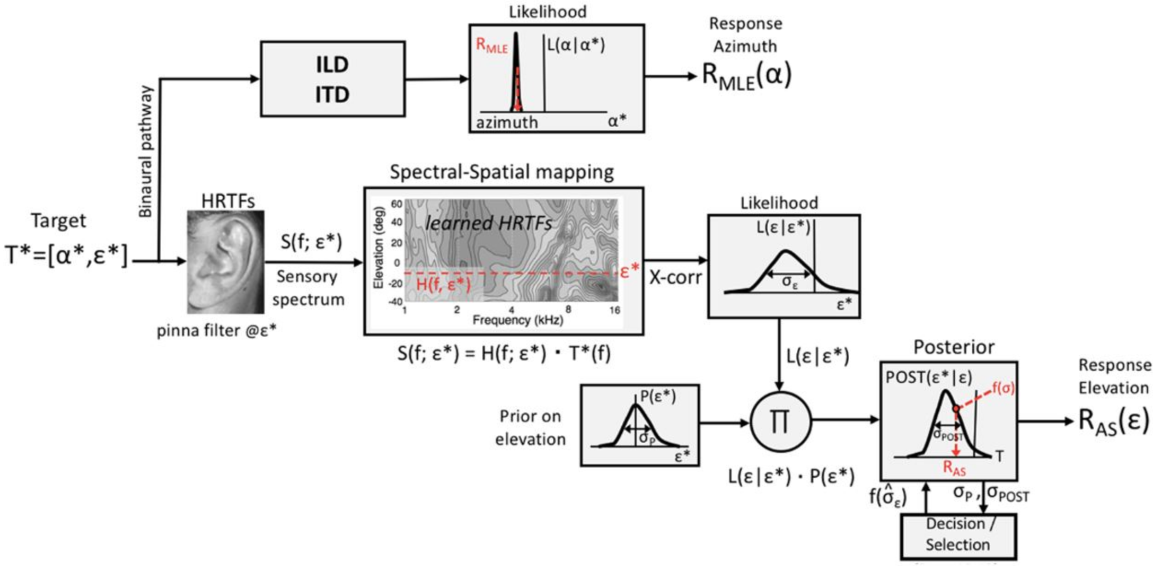


Figure 2.8: A neuro-computational model has been developed to explain how humans perceive sound localization in both azimuth (horizontal direction) and elevation (vertical direction). The neural pathways responsible for extracting these coordinates operate independently, as the mechanisms involved in processing azimuth and elevation differ significantly. Generally, the estimation of azimuth is more accurate than that of elevation, resulting in narrower likelihood functions and a more reliable maximum likelihood estimation (MLE). [12]

In reality, however, noise is included in the signal which can lead to large localization errors. Hence it is thought that prior knowledge also contributes to the final response elevation. This spatial prior $P(_T, \epsilon_T)$ ensures a more precise posterior. The decision-making is finally performed by a so-called maximum-a-posteriori (MAP) estimate.

Even though localization in elevation is mostly monaural, still the brain receives information from the right as well as the left ear. This information is combined by creating a spatial perception and by performing binaural weighting which is schematically shown in fig. 2.9. The order of these two steps is still unknown [54].

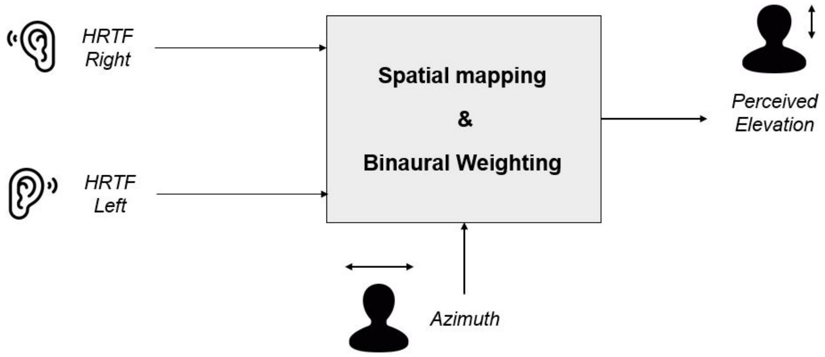


Figure 2.9: Schematic overview of the creation of the final perception in elevation combining spatial perceptions created by the right and left ear using HRTFs. Note that the percept in azimuth also influences the elevation percept.

2.3.4. Localization of moving sound

The principles previously explained are thoroughly tested for the localization of static sounds. Much less is known about the representation of moving sounds. Visual targets that move through the environment, can be tracked with smooth pursuit eye movements. Additionally, the visual system can make use of retinal slip, which provides motion cues that can be exploited by the system [36] [39]. In the auditory system, such cues do not exist as there is no ‘cochlear slip’. As a result, it is not clear whether the sound location is being calculated in a discrete (glimpses) or continuous way [40]. Smooth eye movements to auditory targets seem to be practically non-existent, rather than making a smooth eye movement, subjects make a series of saccades [4]. Also, people tend to move their heads toward the audio stimuli presented and it has been demonstrated that humans are capable of accurately tracking a moving sound source with their head both with physically moving sources and simulated phantom movement [32] [10][26]. To our knowledge, sensorimotor processing with combined visual and auditory targets in dynamic tasks has not been studied to date.

2.4. Experimental settings

Depending on the goal of experiments different experimental conditions can be chosen to assess sound localization performance. In this section, firstly, different sound fields for experiments are discussed. After that, the different methods to present audio cues will be explained.

The theoretical framework presented in this discussion is based on the work outlined in the booklet by Philip Joris and Peter Bremen [9]. The concept of the near-field region refers to an area that extends approximately one wavelength of the sound or three times the largest dimension of the sound source. Managing sound levels within this region can be challenging due to its poorly defined boundaries. Transitioning from the near-field region, the free field region begins, characterized by a decrease in sound level by six decibels with each doubling of the distance from the source. Additionally, sound propagation in this region is uniform, making it suitable for acoustically controlled experiments. Sound-attenuating foam can be used to minimize reflections and to create free-field conditions in a so-called anechoic room. Typically, anechoic rooms are used for auditory localization experiments. In real-life scenarios such as cockpits and cars, the sound waves in the environment interact with surfaces such as walls, ceilings, and floors, leading to reflections and creating a reverberant field. In the context of this research, this condition will be referred to as an echoic room. A schematic overview of these regions is given in fig. 2.10.

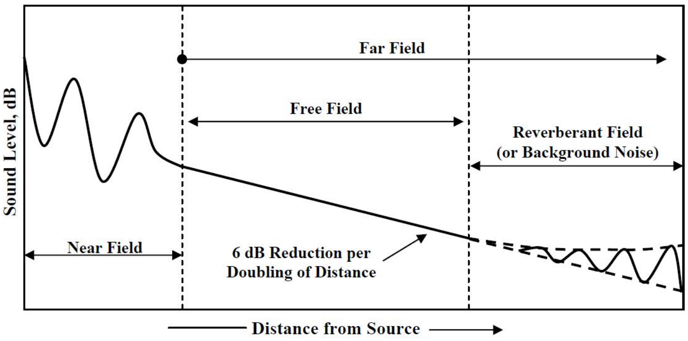


Figure 2.10: Definitions of different sound fields. The near field spans roughly one wavelength of the sound or three times the largest dimension of the sound source, after that the free field starts. The reverberant field starts where free field waves are reflected back. The free-field condition is mostly used for localization experiments due to the uniform sound propagation. [37]

If experiments are performed in an echoic room, reflections highly affect the localization. Some reflections of a sound source S perceived by listener L are shown in fig. 2.11, in which $[w(\tau)]$ is the effect of sounds being filtered by the wall. The energies of the different signals arriving at the listener L are shown by the differences in thicknesses, in which the thickest line has the most energy. The auditory system uses the relative energies of the different sound paths to estimate the distance to the source [29]. The timing differences (ΔT) are caused by the path length differences between the direct signal and the reflected ones.

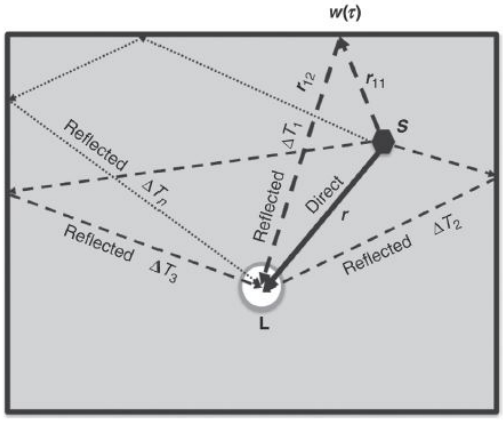


Figure 2.11: Overview of reflections created by a sound source (S) and perceived by a listener (L). The timing differences (ΔT) are caused by the path length differences between the direct signal and the reflected ones. The energies of the sound waves are shown by the variances in thicknesses of the paths shown, the thickets lines having the most energy [51]

Experiments can be conducted in various sound fields, and multiple techniques can be used to generate the audio stimuli. The following section provides a description of the most relevant techniques.

1. *Single speaker per location*: In this setup, there is one speaker positioned at each target location and the sounds are presented from their actual positions. Although it is expensive to use multiple speakers when testing multiple target locations, it is beneficial that the coupling between different acoustic cues is maintained.

2. *Simulation of target locations using two speakers*: The interaural level differences (ILDs) can be simulated using two speakers by adjusting their volume linearly based on the target angle of the sound source [32]. This setup is simpler than the previous one mentioned, however, the accuracy of acoustic cues is compromised as they have to be simulated.

3. *Virtual acoustic space*: Headphones can be used to simulate acoustic cues. Virtual acoustics is based on the principle that the auditory system relies solely on the acoustic pressure at the eardrums to perceive spatial information, and linear filters can be utilized to replicate the sound wave-to-eardrum transformation [59]. Virtual acoustics is particularly effective when non-acoustic signals, such as eye- and head movements, can be somehow excluded from the perceptual task [51].

2.5. Conclusion

Interaural level differences serve as an important cue for sound localization in azimuth, whereas head-related transfer functions are used in elevation. The choice of experimental settings depends on the specific research objective. Real-world situations impose certain limitations, such as a restricted number of speakers and the presence of reflections. This makes a setup involving two speakers with linear volume adjustments an appropriate choice for investigating the potential benefits of (additional) audio cues.

3

Measuring localization behavior

In the previous chapters, the fundamentals of audiovisual integration and sound localization have been explained. This chapter will focus on sensorimotor processing and issues related to reference frames, coordinate transformation, and the analysis of localization behavior.

3.1. Human sensorimotor processing

Localization tasks in humans can be described as a simple input-output model including three main steps; the processing of the sensory cues, the spatial perception created from this, and the motor response executed to a specific location [51]. This is schematically shown in fig. 3.1. The inputs to the system are the stimuli, e.g. a sound or a light flash. The outputs of the system are motor responses, e.g. head or eye movement or a simple button press. The type of motor output (eye/head/arm) is referred to as the effector.

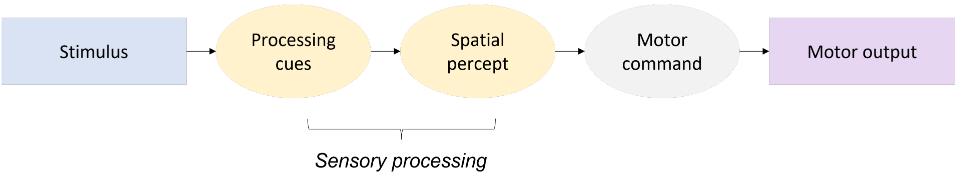


Figure 3.1: The steps in a localization task presented as a simple input and output model consisting of three main steps; sensory processing, the creation of a spatial percept, and the generation of a motor command

Measuring sensorimotor processing in humans can be challenging due to the complex nature of the mechanisms involved. However, several methods have been developed to assess the interplay between sensory processing and motor output generation. One approach is to measure natural gaze-orienting behavior [16]. Here, gaze, also called eye-in-space position, indicates the sum of the eye-in-head and head-in-space positions. The subject is asked to make coordinated eye-head movements toward target stimuli while their gaze position is being recorded. Parameters extracted from the measured movement trajectory, i.e. saccade, provide information about response latency and accuracy of human localization behavior [16]. Using gaze to characterize sensorimotor processing is motivated by the fact that, initially, visual information is encoded in an oculocentric, eye-centered, reference frame F (fig. 3.2) [51]. A specific target in space T is encoded relative to the retina in the eye T_E. The eyes are free to rotate within the head, the eye-in-head orientation is given by E_H in fig. 3.2.

As discussed in chapter 2, the location of a sound source needs to be reconstructed from binaural and monaural cues arising from the distance between the two ears and the filter characteristics of torso, head, and pinnae. As a result, the auditory space is encoded in craniocentric, head-centered, coordinates H (fig. 3.2).

When discussing reference frames in this context two types of reference frames are considered; egocentric and allocentric reference frames. Here, egocentric reference frames indicate that the location of objects is encoded relative to the observer's body, while allocentric reference frames measure the relative distances and orientations of objects in the world to each other, regardless of the observer's position [51]. The aforementioned oculocentric and craniocentric reference frames are examples of egocentric reference frames. To program goal-oriented eye-head movements towards visual, auditory, or audio-visual stimuli (fig. 3.2 T, O_2, O_3), these two reference frames need to be aligned. Additionally, the range of eye movements is limited (fig. 3.2 oculomotor range (OMR)), and localization of targets at angles greater than ≈ 10 degrees have to be performed by combined head-eye movements rather than eye-movements movements only [16]. Also, for programming an appropriate gaze output towards real-life targets, eye and head movements need to be coordinated and target location needs to be transformed from egocentric reference frames into the world-centered (fig. 3.2 W) allocentric reference frames.

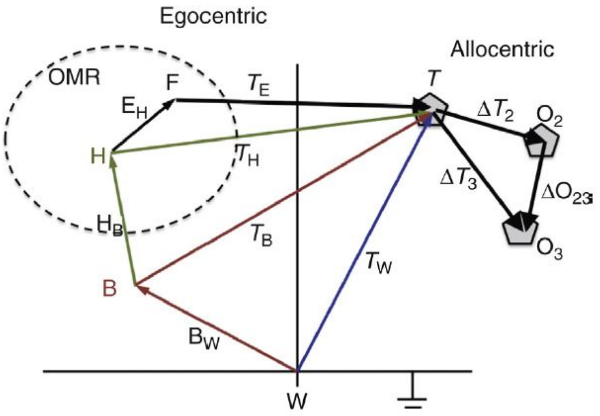


Figure 3.2: Different coordinate systems used in localization tasks; world-centered (W), body-centered (B), head-centered (H), eye-centered (F), target in the world (T) and two other landmarks (O_2 and O_3). Also, the Oculomotor range (OMR) is shown. [51]

How this transformation is achieved by the brain is poorly understood [51]. However, several hypotheses have been proposed regarding these coordinate transformations. For sound localization, it has been suggested that sensory inputs are transformed by making use of binaural and HRTFs cues to specify the target in craniocentric coordinates [51]. By including proprioceptive information on the initial head position target location is transformed into allocentric world coordinates and fed into the gaze motor map. The gaze motor map also receives information on the current eye and head positions, which are needed to program appropriate eye-centered and head-centered motor commands [51].

The coordinate transformations that take place during audiovisual localization tasks are currently not fully understood. However. audiovisual benefits can be observed when measuring saccadic eye movements [11]. Multisensory integration can also be observed by tracking head movements, however, head movements are less accurate compared to eye movements [16] [52]. Further research is needed to better understand the complex transformations involved in audiovisual localization.

In research on smooth pursuit tasks, it was established that humans are not able to track moving sounds using smooth pursuit eye movements whilst this is possible for moving visual targets [4]. However, moving sounds can be tracked continuously using head movements [26]. Next to that, smaller eye-head latency is observed for auditory stimuli compared to visual stimuli [16]. Based on these findings, and acknowledging the fact that these coordinate transformations cause inaccuracies and latency, it can be hypothesized that using a modality-matched effector (eye → visual targets or head → auditory targets) for goal-directed orienting is more favored. Secondly, it can be hypothesized that the motor command and properties of the response, e.g., duration or peak velocity, are independent of the modality of the target and are solely linked to the effector. This is already shown for head and eye movements where the head has larger inertia than the eye, therefore, head peak velocities are lower than those of the eye for targets at the same location irrespective of stimulus modality [16]. Next to that, similar peak velocities are observed for head movements for different unimodal stimuli [16], which suggests that at least for head movements, the type of stimuli has limited influences on the motor behavior observed.

3.2. Sensorimotor processing in a manual control task

Responses to audio and visual stimuli can also be measured using arm movements. In such experiments, participants are typically required to point toward a visual or auditory target [63] [62]. This research focuses on steering tasks as performed by car drivers or airplane pilots who typically interface with the vehicle via a steering wheel or joystick. Typically, these interfaces reduce the possible degrees of freedom (DOF) of the arm movement and limit the natural arm movement. Irrespective of the DOF, arm position is encoded in body-centric coordinates such that targets initially encoded in oculocentric or craniocentric coordinates need to be transformed into the appropriate reference frame to allow the programming of correct orienting responses. This adds a layer of complexity to sensorimotor processing and is schematically shown in fig. 3.3. In which f_t is the target presented, e is the error between the target and the output, u is the arm movement initiated by the subject, and x is the output of the system. Also, the motor command required for arm movements is more complex compared to eye or head movements, which can lead to longer reaction times and more variability in the data [3]. Note that the reasoning explained above also holds for a situation when short stimuli (100 ms) are presented, in that case, f_t is memory-based.

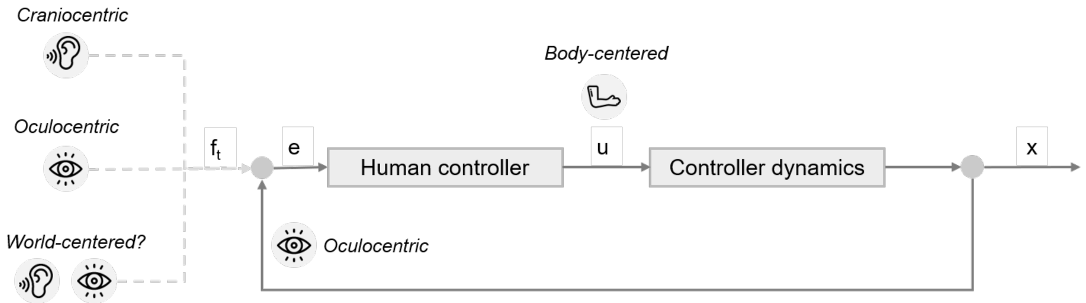


Figure 3.3: Stimuli are presented in different reference frames compared to the reference frame used for the motor command (body-centered). Hence, a coordinate transformation needs to take place. f_t is the target presented, e is the error between the target and the output, u is the arm movement initiated by the subject, and x is the output of the system.

As it is expected that the characteristics of the motor output (effector) are independent of the modality of the target, different properties of the response must be compared. A phase-plane diagram is an intuitive way to present multiple movement characteristics such as velocity and acceleration. In phase-plane representations, the change in the controlled state is plotted against its rate. Phase-plane representations are very suitable to analyze performance in a discrete step tracking performance task and have been used previously to assess the influence of motion cueing on performance [43] [34] [33].

Even though it is still uncertain how motor plans for eye, head, and arm movements are created, it has been established that arm-pointing accuracy is low in head-restrained conditions [6] [42]. This is likely caused by the suppression of the natural coordination of the eye-head-arm movement. Furthermore, in eye-hand coordination tasks it has been observed that the eye leads the hand [15]. These findings are inline with the presented hypotheses that modality-matched effectors are favored.

3.3. Dynamic localization tasks

In a continuous closed-loop tracking task, the target remains perceivable during the response and needs to be followed. Continuous feedback can be used by subjects to adapt and update the motor command. Much research has been performed on the influence of visual cues and motion cueing in discrete and continuous tasks [1] [20] [57]. These models provide a high understanding of different behavioral aspects of manual tracking performance. Also, continuous auditory feedback during tracking tasks can decrease response latency and improve tracking accuracy [41] [56]. How these effects combine under audio-visual stimulation is yet unknown.

3.4. Conclusion

Human processing of stimuli resulting in specific motor output (effector) can be categorized into two stages: sensory processing and motor command generation. Here it is hypothesized that the generation of the motor command remains unaffected by the modality of the stimulus. This implies that factors like peak velocities are solely determined by the type of effector. Furthermore, it can be hypothesized that using an effector that matches the modality of the stimulus (such as using the eyes for visual targets and the head for auditory targets) during localization tasks is favored. By performing a manual localization task with short stimuli ($\approx$ 100 ms), the target disappears before the orienting response is initiated. These experiments allow for the characterization of the influence of the effector on the localization response.

4

Preliminary results

In the previous chapters, a review of the relevant literature has been presented. The current chapter focuses on the preliminary experiments. The objectives of this chapter are to define the research objectives, describe the experimental setup, present the obtained results, and to provide the conclusions.

4.1. Research question and hypotheses preliminary experiments

The audiovisual benefits observed in eye and head movements have been discussed in previous chapters as well as potential benefits in more complex motor tasks. These experiments have been performed mostly in anechoic rooms with elaborate setups. In daily life situations, such as cars and cockpits, significant limitations are present, including a limited number of speakers and an echoic environment. To better understand the benefits of audio cues in this environment, it is essential to investigate both static and dynamic tasks. As a first step, it is explored whether it is possible to elicit or simulate discrete sound location perception with only two free-field speakers in a reflective environment. This initial investigation is mandatory and serves as a prerequisite for further exploration of the benefits of audio cues in both static and dynamic tasks. The main research question for this preliminary experiment is defined as follows:

> **To what extent does manipulating the relative sound level of two speakers affect the elicitation of discrete sound location perception in an echoic environment?**

In accordance with the research question, several hypotheses can be formulated, as outlined below.

1. Three different types of audible noise are introduced; low-pass noise (0.5 - 1.5 kHz), broadband noise (0.2 - 20 kHz), and high-pass noise (4 - 20 kHz). Only interaural-level differences (ILDs) and not interaural-time differences (ITDs) are manipulated. As described in section 2.1 ILDs are used for sound localization for high-frequency sounds only. It is expected that for high-pass and broadband noise higher localization accuracy will be observed compared to low-pass stimuli. It is expected that for high-pass and broad-band noise, localization accuracy will be more accurate, and responses will have lower variance, compared to low-pass noise. Compared to the high-pass stimuli, lower accuracy is expected for broadband stimuli. The broadband stimuli contain correct ILDs but unchanged incorrect ITDs. This creates uncertainty in the localization which is expected to result in a lower accuracy and higher variance.

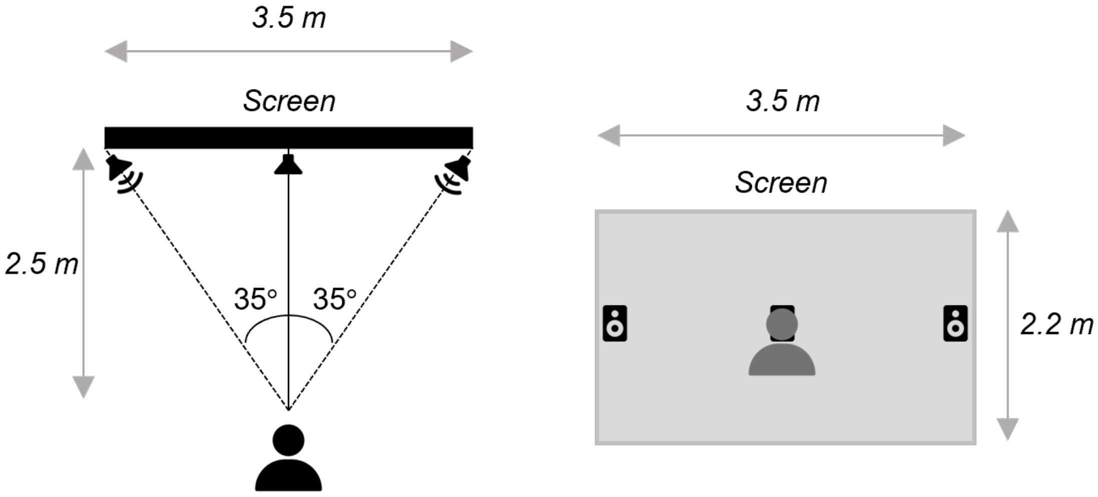
2. Misalignment in the two speakers may lead to decorrelation of the two sound signals, resulting in a defuse sound percept which is observed by a higher variance in elevation and azimuth percept. It is expected that these differences will be most pronounced around azimuth 0 degrees, as both signals are at full strength at this position.

3. As audio stimuli are processed in the craniocentric reference frame, head position will greatly affect the elevation and azimuth percept of presented stimuli. The position of the subject's interaural axis plays a key role in this as well as the subject's alignment with the center line.

4. It is hypothesized that the motor command for arm movements is more complex compared to head movements. Also for audio cues, head movements are considered the most natural effector. This is expected to introduce more execution noise resulting in reduced localization accuracy and increased reaction times compared to head movements. Nevertheless, the manual orienting task should be executable by the subject

4.2. Setup

This research aims to investigate the effect of audiovisual cues in an echoic environment, such as cockpits and cars, on performance in a manual control task. The facilities of the human-machine interaction lab at the Aerospace Engineering faculty of the Delft University of Technology are used to simulate such an environment. This facility is regularly used to research manual control tracking tasks for automotive as well as aerospace applications. The facilities used in this room consist of two speakers which are part of the Logitech Z906 5.1 surround system. The visual cues and joystick output are projected using a BenQ TH690ST projector on a white screen that has a width of 3.5 m and a height of 2.2 m. The subject is positioned in the middle of the room at an equal distance from both speakers as shown in fig. 4.1a and fig. 4.1b. The maximum attainable angle for stimulus presentation is 33 degrees (fig. 4.1a).

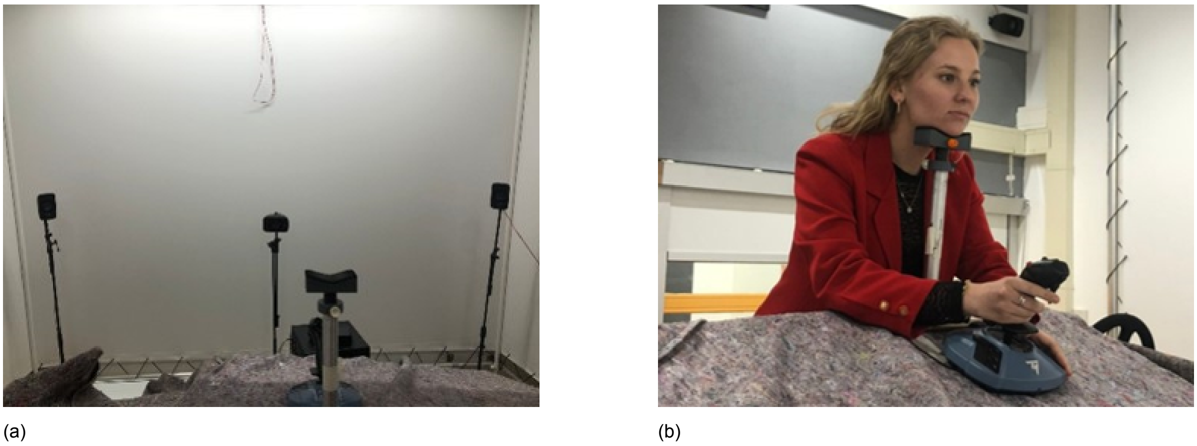


(a)

(b)

Figure 4.1: a) Top view of the setup used for the experiments including the three speakers of which the middle one is not functional. Maximum angles are 35 degrees. (b) Back view from setup including the three speakers and sizes of the screen.

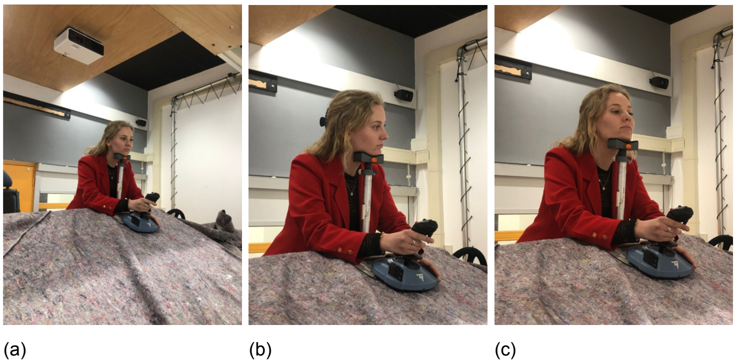
Previous research has shown that two speakers are sufficient to simulate different target locations [32]. For this experiment, two functional speakers are used at the corners of the room. In localization tasks, responses to audio stimuli can be pulled toward a specific location if a strong spatial cue is present, hence, one dummy speaker is placed in the middle to prevent the localization from being pulled toward the sides [5]. In fig. 4.2a the three different speakers are shown as well as the chin rest used. During experiments, the subject was standing and leaning forward, resting the elbows on the bonnet and positioning their head on the chin rest. The gain-controlled joystick (Thrustmaster TCA sidestick Airbus edition joystick) is positioned in front of the subject as shown in fig. 4.2b.


(a)

(b)

Figure 4.2: a) Position of speakers, including two functional speakers on the left and right side and one dummy speaker in the middle. The dummy speaker is used to prevent localization from being pulled toward the sides. b) Position of a subject during experiments. Note the chin rest and the joystick. The car bonnet has been covered with carpet to minimize acoustic reflections.

As elaborated upon in chapter 2, the position of the head greatly affects the perceived location of the stimuli presented. To reduce variability head movements were reduced using a simple chin rest. Even though this stand limited the movements of the head, still, a large variation in head positions was observed between different runs and subjects. An example of a neutral head position is shown in fig. 4.3a, and a shifted head position to the side and upwards are shown in fig. 4.3b and fig. 4.3c, respectively. Note that fixating the head impacts the natural movement of the arm and influences the localization ability [42], hence, in future experiments, the head position will be measured and corrected.


(a)

(b)

(c)

Figure 4.3: Overview of different head positions when using a simple chin rest. Note that different head positions result in different spatial hearing characteristics. a) Head position that is considered neutral b) Head tilted towards to the side compared to neutral position c) Head tilted upwards compared to neutral position

The manual output is measured using a Thrustmaster TCA sidestick Airbus edition joystick. A picture of the subject during the trials is provided in fig. 4.2b. In fig. 4.4 the output of the gain-controlled joystick projected on the screen, the red dot, is shown and compared to the location of the virtual sound. A deflection of the joystick directly relates to a proportionate position change. The error is defined as the difference between the final output and the real location of the virtual sound.

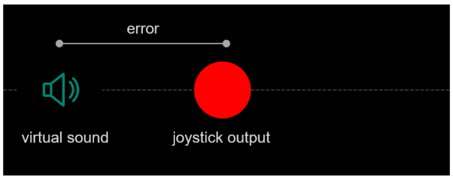


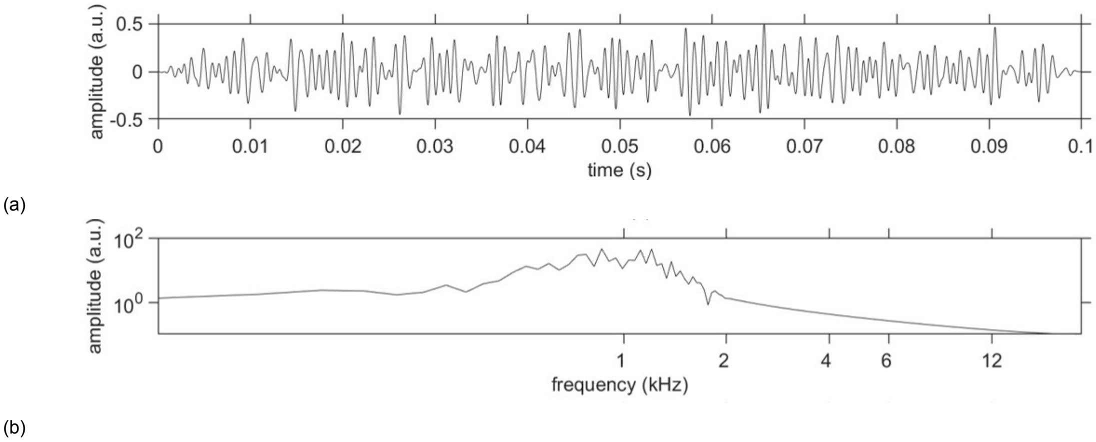
Figure 4.4: Joystick output and virtual sound location presented on the black screen. The error is defined as the difference between the simulated sound location and the joystick output.

The experiments will run on 3 different Dell desktops with a Linux operation system (Ubuntu 18.04.2 LTSS) on the Delft University Environment for Communication and Activation (DUECA). DUECA is a module-based real-time software. A big advantage of this software is that the software itself accounts for software timing latency between the different modules (joystick, speakers, head tracker) and the simulations can easily be distributed over multiple computers. The different sounds are generated using the signal processing toolbox 9.1 of MATLAB. This toolbox is user-friendly and allows for easy visualization of the created signals. For the data analyses of the results, this toolbox will also be used.

4.2.1. Input signal

As explained in chapter 2, sound localization for the azimuth direction mainly relies on two important phenomena: interaural time differences(ITDs) and interaural level differences (ILDs). Previous research has shown that ILDs can be simulated effectively by changing the relative volume of the two audio speakers [32]. To verify that proven localization principles for head and eye movements in anechoic environments, hold for manual localization tasks in echoic environments three different auditory stimuli were tested including low-pass (0.2 - 20 kHz), broadband (0.2 - 20 kHz), and high-pass noise (4 - 20 kHz). All auditory stimuli were ramped in the first 5 ms using a sinusoidal ramp to avoid click artifacts.

The low-pass signal contained frequencies ranging from 0.5 - 1.5 kHz and the corresponding time waveform (fig. 4.5a) and magnitude of the frequency spectrum (fig. 4.5b) are shown fig. 4.5. Note that in this frequency range, only ITDs are used for localization. In this experiment, ITDs are not modified.

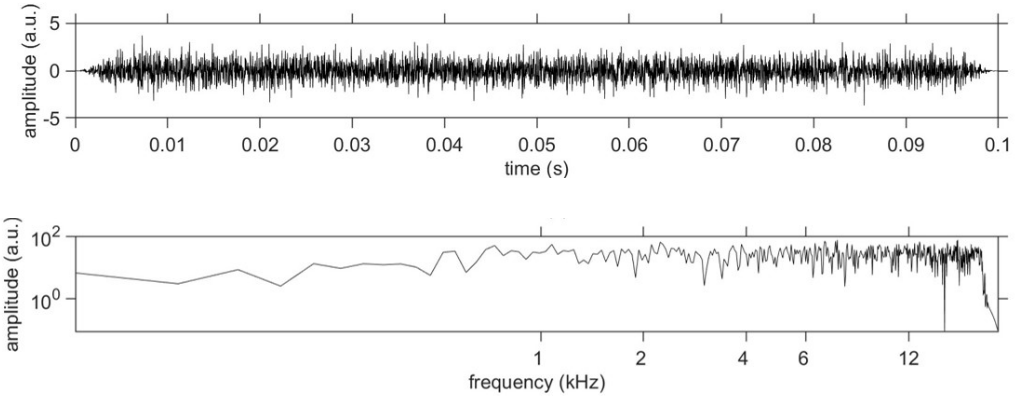


(a)

(b)

Figure 4.5: The low-pass signal contained frequencies ranging from 0.2 - 1.5 kHz and a) the corresponding time waveform b) the magnitude of the frequency spectrum

The broadband signal contained frequencies ranging from 0.2 - 20 kHz and the corresponding time waveform (fig. 4.6a) and magnitude of the frequency spectrum (fig. 4.6b) are shown in fig. 4.6. In this frequency range, both ILDs and ITDs are used for localization.

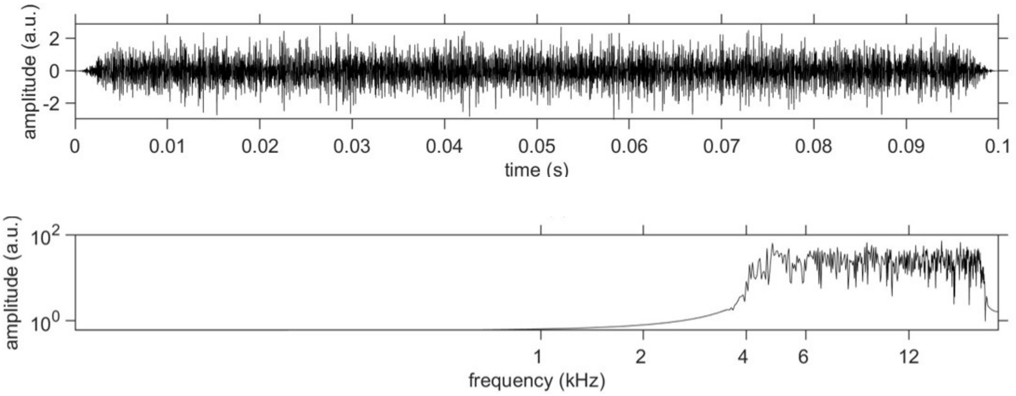


(a)

(b)

Figure 4.6: The broadband signal contained frequencies ranging from 0.2 - 20 kHz and a) the corresponding time waveform b) the magnitude of the frequency spectrum

The high-pass signal contained frequencies ranging from 4 - 20 kHz and the corresponding time waveform and magnitude of the frequency spectrum are shown in fig. 4.7a and fig. 4.7b, respectively. Note that in this frequency range, solely ILDs are used for localization.



(a)

(b)

Figure 4.7: The high-pass signal contained frequencies ranging from 4 - 20 kHz and a) the corresponding time waveform b) the magnitude of the frequency spectrum

A simulated head shadow effect, as explained in chapter 2, is created by manipulating the relative sound levels for the left and right speaker, as presented by Niehof [32]. First, a relative position vector was defined to specify the azimuth angle of the target, α_{target} (in degrees) with respect to the maximum angle α_{max} (in degrees) defined in eq. (4.1).

$$I_{\text{rel}_\alpha} = \frac{\alpha_{target}}{\alpha_{\max}} \tag{4.1}$$

Where α_{target} is always smaller than or equal to the maximum deviation angle from the center line, defined as 35 degrees in all experiments.

$$\alpha_{target} \leq \alpha_{\max} \tag{4.2}$$

Resulting in a scaling vector $Irel$ which varies between [-1, +1]. Sound intensities for the left (I_L) and right (I_R) speakers were then calculated using eq. (4.3)

$$\begin{aligned} I_L &= 0.5 \cdot (1 - I_{\text{rel}}\,(t)) \\ I_R &= 0.5 \cdot (1 + I_{\text{rel}}\,(t)) \end{aligned} \tag{4.3}$$

The original signals created were then multiplied with I_L and I_R, respectively, to determine the signals for the left and right speakers. An example of this relative scaling for a target at location -35 degrees azimuth (left) is given in fig. 4.8.

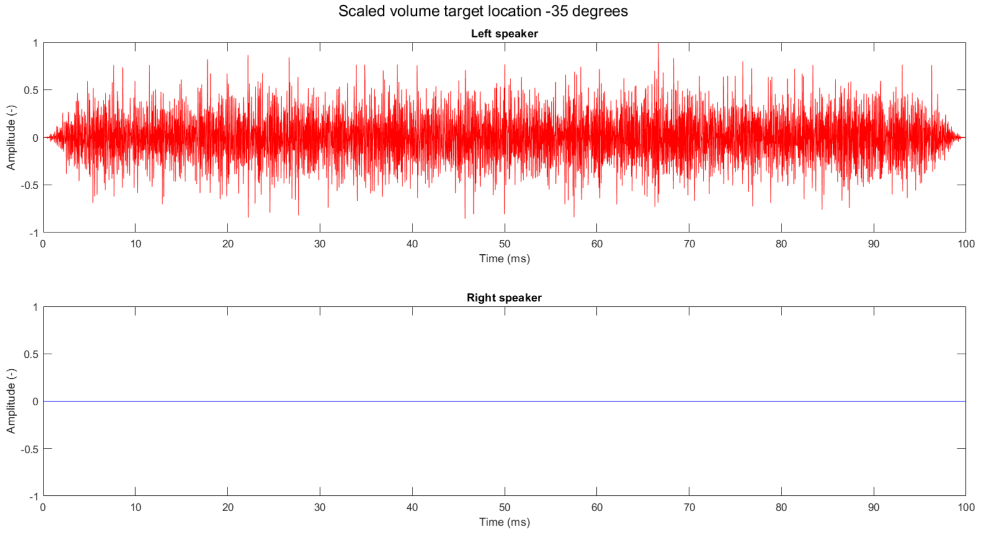


Figure 4.8: Example of relative volume scaling between the right and left speaker for a target presented at -35 degrees (left)

The target locations (α) ranged between -35 degrees (left of the subject) and 35 degrees (right of the subject) in steps of 5 degrees. The stimuli were presented randomly but no conditions were created to assure the same amount of repetitions for each target. This will be included in the final experiments.

4.2.2. Task description and experimental procedure

The subjects were instructed to move the joystick output, the red dot, as accurately and as fast as possible to the perceived location. This contradicting instruction was provided on purpose, as subjects should not focus on one specific movement characteristic. The subject was able to move the joystick in elevation as well as azimuth. fig. 4.9 provides an overview of the trial structure. The trial starts with a pseudo-random wait interval that ranges from 1-3 seconds which is introduced to remove predictability in the stimulus presentation. After this, an audio stimulus was presented with a duration of 100 ms. Depending on the experiment either low-pass, broadband, or high-pass stimuli were presented. Next, subjects needed to move the projection of the joystick, a red dot, to the perceived location. When subjects wanted to confirm their end location they needed to press the button. Subjects confirmed the end location of their movement by pressing a button on the joystick. This was done mainly to simplify data analysis. The trial was completed after 7.0 seconds to provide sufficient time for subjects to execute the motor response and to have some seconds of rest between the trials.

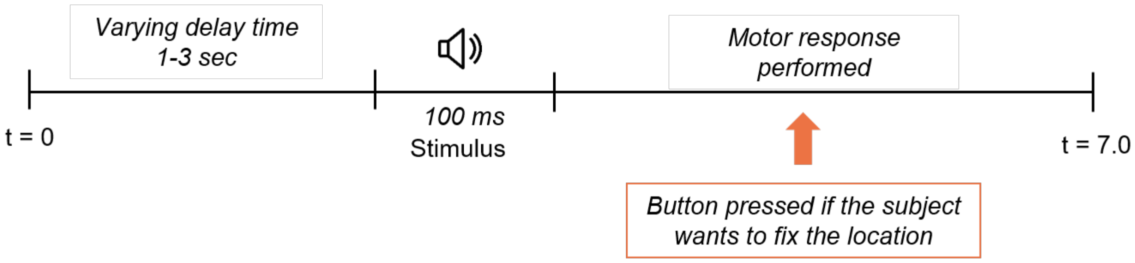


Figure 4.9: Overview of one trial starting with a varying time delay to remove predictability. The endpoints are determined based on the position of the joystick at the time the button was pressed.

In total 5 subjects participated in this experiment. All participants were male, right-handed, and 22-43 years old. The experiment procedure for each subject is shown in fig. 4.10. Each experiment started with a short briefing after which the measurements started. Generally, subjects performed 3 blocks of experiments all taking approximately 12 min. After two blocks, subjects had a short break of 5 minutes after which they started the final block. Each block consisted of 100 trials with one specific noise stimulus (low-pass, broadband, or high-pass). The stimulus order was pseudo-randomized across blocks and participants. The target locations in azimuth, ranging between -35 degrees and 35 degrees in steps of 5 degrees, were presented randomly but it was not assured that the number of repetitions per target location was the same. This will be corrected for the final experiment.

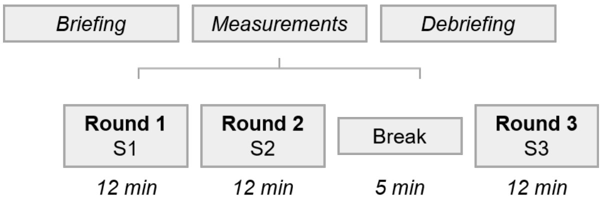


Figure 4.10: Experimental procedure of experiments including three rounds of experiments each consisting of 100 trials.

4.3. Results

In this section, firstly, the results of the responses in elevation and azimuth are presented. After this, the results for the azimuth responses are compared with the target locations using a linear regression model.

4.3.1. Responses in azimuth and elevation

Figure 4.11 depicts the responses to broadband stimuli of subject 4. The large and small circles indicate target locations and response locations for individual trials, respectively. Additionally, target and response locations are connected by thin lines and displayed in the same color. Average response endpoints are indicated with squares and standard deviation in azimuth and elevation are depicted as thick lines. From this figure three main findings become clear. Firstly, the subject's azimuth responses are correlated with the simulated azimuth locations. However, azimuth responses were biased towards more peripheral locations. Locations beyond ±20 degrees were perceived as originating from the speaker location at 35 degrees. Secondly, the stimuli elicited clear elevation responses for simulated locations close to the center line. For instance, the stimulus at 0 degrees azimuth elicited a response at 10 degrees elevation. Elevation responses rapidly decreased with the increasing peripheral location of the sound. Thirdly, responses were not symmetric across the center line. That is, the subject or perceived center line did not coincide with the physical center line as defined by the speakers.

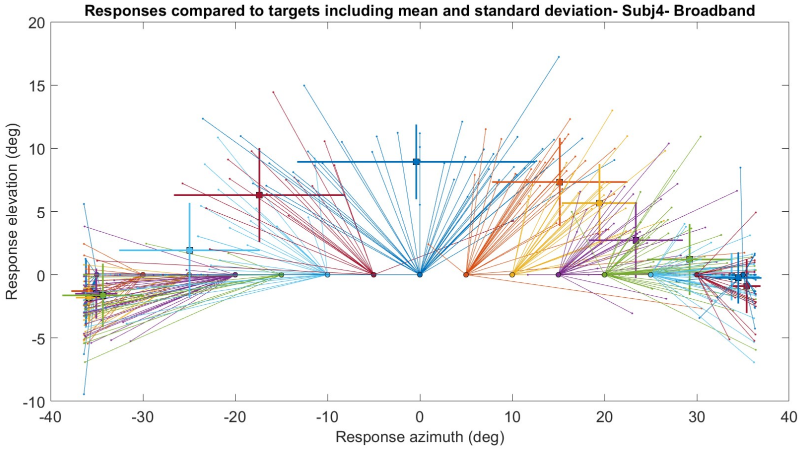


Figure 4.11: Responses to broadband stimuli in elevation and azimuth of subject 4. The large and small circles indicate target locations and response locations for individual trials, respectively. Additionally, target and response locations are connected by thin lines and displayed in the same color. Average response endpoints are indicated with squares and standard deviation in azimuth and elevation are depicted as thick lines.

In fig. 4.12 the average responses of all subjects per target location for high-pass stimuli are shown. Again, target locations are displayed using colored circles, and average responses per target location are displayed using colored small squares. The averages per subject are connected with a specific line type. Note that all targets were presented at an elevation of 0 degrees.

From this figure, several conclusions can be drawn. First of all, the elevation percept depends on the azimuth location; the average elevation perception decreases with increasing peripheral location for all subjects. However, the values of the elevation angles perceived, differ significantly between subjects. Subject 3 and subject 5 even perceived negative elevation angles. Next, it can be observed that average responses for target location azimuth 0 degrees (blue dot in the middle), are not aligned for all subjects indicating differences in the perception of the center line between subjects.

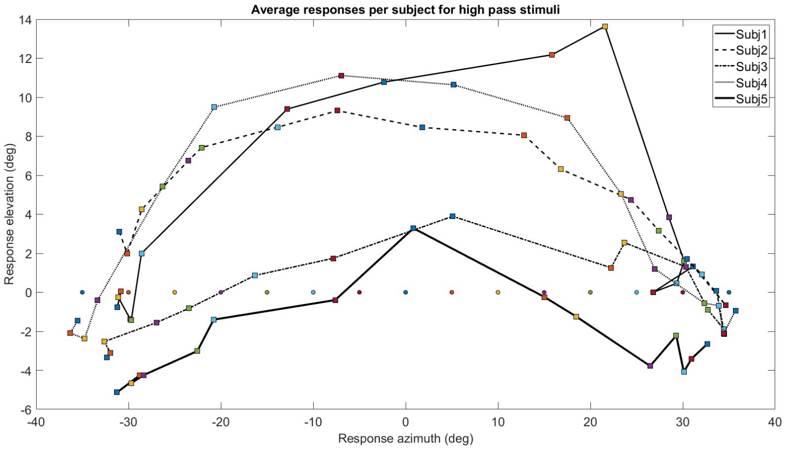


Figure 4.12: Average response for all subjects for high pass stimuli. Target locations are displayed using colored circles, and average responses per target location are displayed using colored small squares. The averages per subject are connected with a specific line type. Note that all targets were presented at an elevation of 0 degrees.

In fig. 4.13 the average responses of all subjects for high-pass and broadband stimuli are displayed. Again, target locations are presented using circles. Average response endpoints to broadband stimuli and high-pass stimuli are presented using squares and triangles, respectively. Average responses are connected to the target location using thin color lines, the different line types and colors correspond to specific subject and target locations. By visually comparing the responses to broadband and high-pass stimuli, no clear differences can be found in response accuracy.

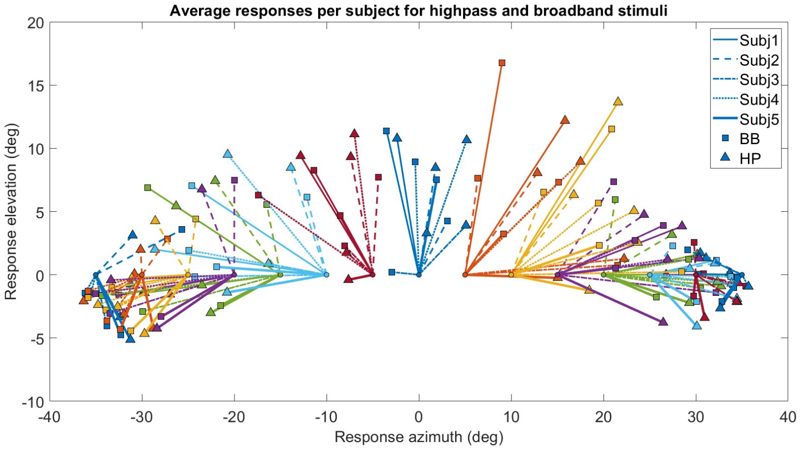


Figure 4.13: Average azimuth and elevation response for all subjects for broadband (BB) and high pass (HP) stimuli. Target locations are presented using circles. Average response endpoints to broadband stimuli and high-pass stimuli are presented using squares and triangles, respectively. Average responses are connected to the target location using thin color lines, the different line types and colors correspond to specific subjects and target locations.

An important aspect when evaluating endpoint accuracy is the standard deviation of the responses. In fig. 4.14, the standard deviation of the endpoint responses for broadband and high-pass stimuli are compared for azimuth and elevation. Averages responses for all target locations and subjects, each subject having a specific marker, are compared. A dotted gray line is plotted representing the one-on-one relation between the standard deviations. Points presented above this line indicate a higher standard deviation for high-pass responses compared to the standard deviation of broadband responses for a specific target location and subject. The points presented below the gray line indicate the opposite, meaning, a larger standard deviation for broadband responses compared to high-pass responses. Evaluating fig. 4.14 no clear clustering of points can be observed for elevation and azimuth.

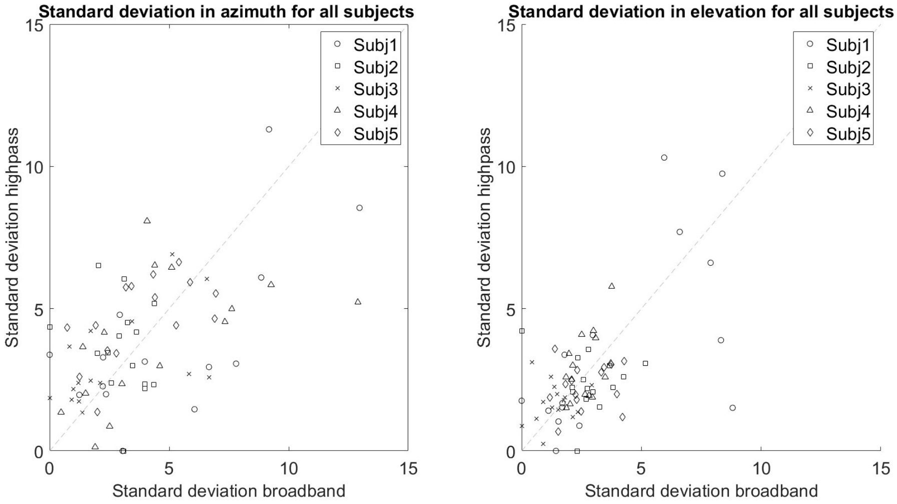


Figure 4.14: Comparison of the standard deviation of the endpoint responses for all subjects in azimuth and elevation. The gray line indicates a one-on-one relationship between the standard deviation of broadband and high-pass stimuli. If the standard deviation for broadband stimuli would be higher a clustering of responses is expected in the lower right corner.

4.3.2. Response accuracy in azimuth

This section aims to quantify the localization accuracy in azimuth by presenting a detailed comparison between the azimuth responses and the target locations.

The responses of subject 2 to all stimuli (high-pass, broadband, and low-pass) are depicted in fig. 4.15 using gray dots, which are sorted according to the target location. The dotted gray lines show a one-on-one relation between the target and response locations. The colored lines represent the best linear fit for the response, while the R^2 value indicates the deviation from this best-fit line. This method for assessing response accuracy was introduced in chapter 1.

Based on these observations, several conclusions can be drawn. Firstly, the responses to high-pass, broadband, and low-pass noise for peripheral target locations (25, 30, and 35 degrees) are similar. In all three cases, the R^2 coefficient is close to 1, indicating minimal deviation from the linear fit. Moreover, it is evident that the colored lines representing the best fit do not align perfectly with the gray dotted lines.

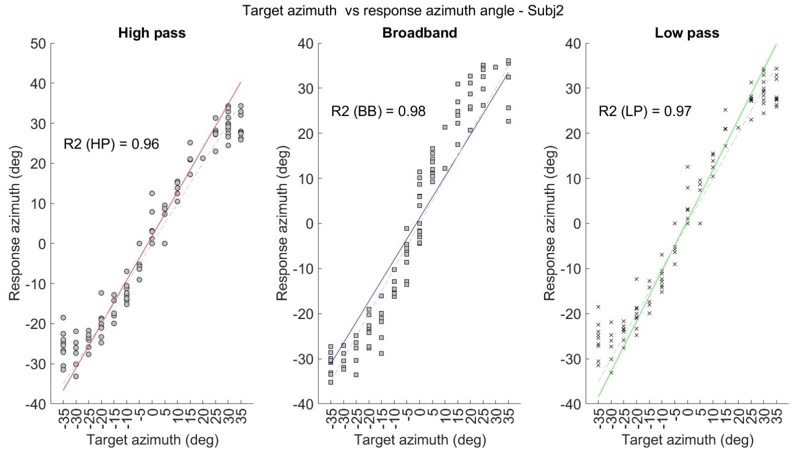


Figure 4.15: Responses plotted in gray dots and sorted for each target location and shown in per stimulus type. The gray dotted lines indicated a one-on-one relation between the target and the response. The colored lines show the best linear fit of the responses and the coefficient of determination (R^2) indicates the deviation of the responses to this linear fit. Values close to 1 indicate a good fit.

These differences can be attributed to the gain and bias of these lines, which are summarized for all subjects in Table 4.1. Note that for subject 4 and subject 5, no responses to the low-pass noise were measured due to the time constraints of the subjects. A linear fit with a gain of 1 and a bias of 0 combined with an R^2 coefficient equal to 1, indicates that responses were precisely at the target location. From Table 4.1, no significant differences in localization performance can be observed between high-pass, broadband, and low-pass stimuli. The implications of these findings will be discussed in the next section.

	Subj1			Subj2			Subj3		
	HP	*BB*	*LP*	*HP*	*BB*	*LP*	*HP*	*BB*	*LP*
Gain	1.17	1.13	1.17	1.10	0.92	1.12	1.23	1.22	1.18
Bias	-0.36	-2.55	-2.76	1.86	1.20	0.68	2.80	-0.79	-1.24
R^2	0.91	0.94	0.94	0.96	0.98	0.97	0.96	0.95	0.96

	Subj4		Subj5	
	HP	*BB*	*HP*	*BB*
Gain	*1.30*	1.33	1.13	1.15
Bias	*1.54*	-1.61	1.20	0.00
R^2	0.95	0.95	0.95	0.97

Table 4.1: Bias, gain, and the coefficient of determination (R^2) for all regression lines sorted per subject and three types of audio stimuli: high-pass (HP), broadband (BB), and low-pass (LP). Note that no significant differences between the parameters (Bias, gain, and R^2) are observed for the different stimuli types.

4.4. Discussion

In this section, the results presented will be compared with the hypotheses, and recommendations for the final experiments will be given.

The first hypothesis stated that high-pass stimuli would yield the best performance, however, considering fig. 4.13 no differences can be observed between the average endpoint values for broadband and high-pass responses. Similarly, the comparison of standard deviations in fig. 4.14 does not indicate significant differences in variance for high-pass and broadband responses. For localization of broadband noise, both interaural level differences (ILDs) and interaural time differences (ITDs) are used. It could be that the correct ILD cues are utilized for localization while the incorrect ITD cues are disregarded, as previously observed in the frequency range of 4-20 kHz [24]. Surprisingly, when considering the regression fits for the low-pass response, responses appear to be accurate, even though the ILD cues are not used for localization in this frequency range. However, further investigation revealed the presence of a high-frequency click in all stimuli due to software delays. In hindsight, subjects likely used the high-frequency components of this click to localize the stimuli.

For sound sources precisely positioned at the center line, interaural level differences are zero. However, due to the limited testing of the speakers used, small volume and timing differences may exist, causing the sounds perceived from the left and right to be slightly different. This disparity creates a diffuse percept, resulting in significant variance in responses for azimuth 0 degrees. For targets located to the left or right of the center line, clear level differences between speakers are present, generating a clear left or right percept. It is noteworthy that elevation perceptions around azimuth 0 are consistently higher compared to other positions.

The high elevation perception may be influenced by the head position and the position of the speakers relative to the aural axis. Although this does not explain the variance in elevation for different azimuth positions, it does account for the mostly upward percept in elevation observed among participants, as depicted in fig. 4.12. The head position also affects the perception of azimuth, most clearly shown by the responses observed for target location azimuth 0 degrees. The deviations at the center line create substantial variation in azimuth averages between subjects shown in fig. 4.13. Consequently, this leads to a perceptual shift for other target locations presented in fig. 4.12. Given these variations in head positions, comparing results between subjects becomes difficult. Furthermore, since the stimuli were designed assuming the subjects were positioned exactly in the middle of the speakers and looking straight ahead, any head shift would result in incorrect acoustic cues that could influence localization. Therefore, in the final experiment, head movements will be tracked, and extensive calibration experiments will be conducted.

Finally, it was hypothesized that joystick output serves as a reliable measure to assess localization performance. Although it remains difficult to determine whether the inaccuracy arises from additional noise created by the effector (motor output) or incorrect sensory processing, the coefficients of determination (R^2) show a good fit. To determine whether the noise originates from the processing of the sensory cues or the motor command, head movements will be measured in the final experiments.

4.5. Conclusion

In an echoic room, an auditory manual localization task can be conducted using two speakers to simulate Interaural Level Differences (ILDs). However, the obtained responses were found to be less accurate compared to previous results obtained using head movements as a measure. Additionally, considerable variability in the upward elevation percept was observed, which can be attributed to the position of the subjects' aural axis as well as the diffuse percept created by slight differences between the two speakers. For the final experiments, it is recommended to perform head movement measurements and conduct calibration experiments to assess the influence of head position on the obtained results. Also, the analysis of head movements can be used to investigate whether the accuracy and variability of responses results from sensory processing or from the type of effector.

5

Experiment Proposal and Data Analysis

The previous chapter described the results of preliminary experiments using a manual pointing task to evaluate human sound localization behavior. It was concluded that simulated target locations could be adequately localized. The research questions and hypotheses for the final experiments and the proposed experimental setup will be discussed in this chapter.

5.1. Research Questions and hypotheses final experiments

Both static and dynamic tasks need to be investigated to better understand the benefits of audio cues for a manual control task in echoic environments. As a first step, the possibility of eliciting discrete sound location percepts with only two free-field speakers in an echoic room was explored. Here, the focus will be on analyzing localization performance by quantifying reaction times and endpoint accuracy for unimodal and multimodal stimuli. Additionally, manual responses will be compared with head movement localization responses. The main research question for this experiment is defined as follows

> What audiovisual benefits can be observed in a discrete manual localization task providing multisensory cues in an echoic environment?

As discussed in chapter 3, extensive studies have been conducted on human audio, visual, and audiovisual localization behavior by measuring head, eye, or gaze movements. By using head movements as an effector, the results can be compared to the available literature. Hence, the effectiveness of simulated interaural level differences (ILDs) in an echoic environment can be evaluated. By comparing arm and head responses, the effect of sensory processing and motor commands can be dissociated. If the audiovisual benefit is a result of sensory processing, similar benefits are expected irrespective of the effector. However, parameters such as reaction time and localization accuracy may still differ between the effectors. The following hypotheses will be tested:

1. **Sound localization:**
    - Discrete sound-source locations can be adequately simulated using a two-speaker setup (azimuth +/- 33 degrees). By imposing volume changes that mimic ILDs at the listener's ears, a resolution of 4.7 degrees can be achieved in an echoic room.
2. **Multisensory benefits**
    (a) Compared to unimodal stimuli, faster reaction times and decreased localization variance are expected for multimodal stimuli. With accurate sound localization, the spatial percepts of the auditory and visual stimuli are expected to align sufficiently, resulting in audiovisual interactions tested with the Race Model for reaction times.

    (b) Audiovisual benefit is independent of the effector. The magnitude of reduction in reaction time and endpoint variance with audiovisual stimuli compared to unimodal stimuli is expected to be the same for head and arm movements.

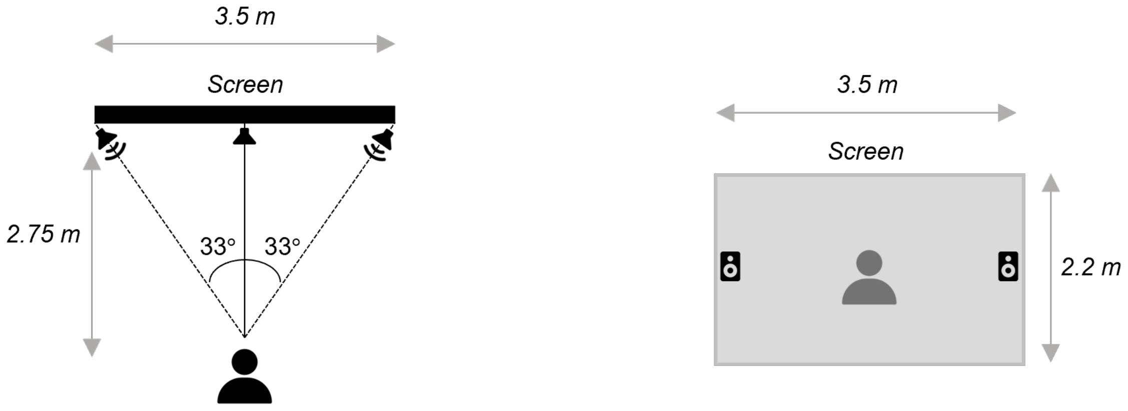
(c) As discussed in chapter 3, due to the different coordinate transformations involved in translating stimulus coordinates to effector coordinates as well as more execution noise, it is expected that reaction times are slower and localization accuracy is reduced for arm movements compared to head movements.

3. **Motor behavior**

   - Within one effector, movement parameters such as peak velocity are not expected to differ across audio, visual, and audiovisual stimuli. As it is hypothesized that the motor command is not influenced by the stimulus modality.

5.2. Setup

This research aims to investigate the effect of audiovisual cues in an echoic environment, such as cockpits and cars, on performance during a manual control task. Hence, the facilities of the human-machine interaction lab at the Aerospace Engineering faculty of the Delft University of Technology are used to simulate such an environment. This facility is regularly used to research manual control tracking tasks for automotive as well as aerospace applications. The facilities used in this room consist of two speakers which are part of the Logitech Z906 5.1 surround system. The visual cues and joystick output are projected using a BenQ TH690ST projector on a white screen that has a width of 3.5 m and a height of 2.2 m. The subject is positioned in the middle of the room at an equal distance from both speakers as shown in fig. 5.1a and fig. 5.1b. The maximum attainable angle for stimulus presentation was 33 degrees (fig. 5.1a). This is slightly smaller compared to the preliminary experiments as the subject will be seated during the final experiments, imposing limitations on the distance to the screen.

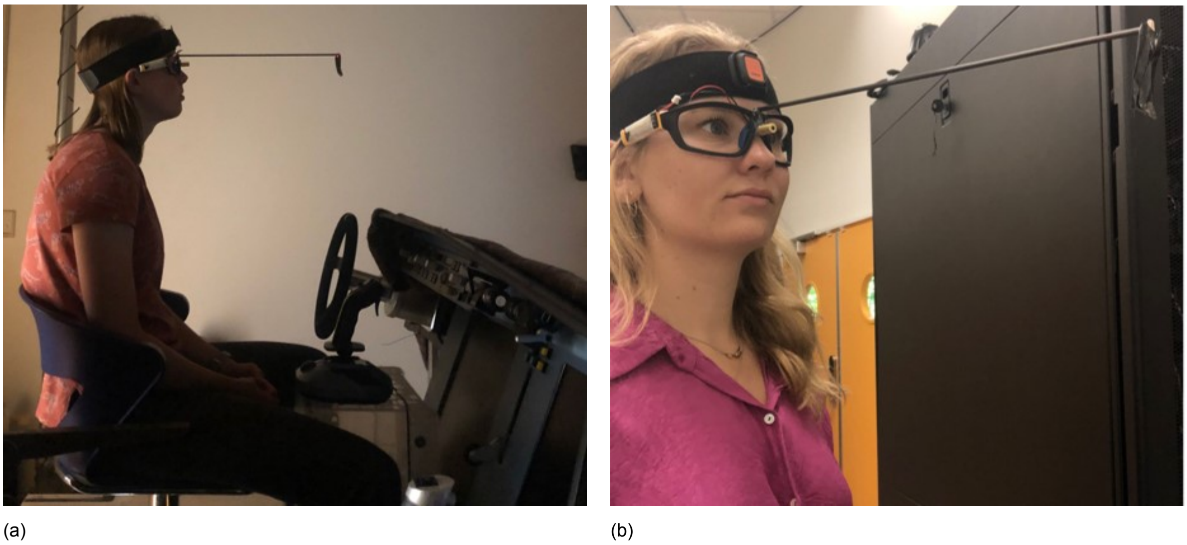


(a)

(b)

Figure 5.1: a) Top view of the setup used for the experiments including the two speakers. Maximum angles are 33 degrees. (b) Back view from setup including the two speakers and sizes of the screen.

During experiments, the subject will be seated in a high adjustable chair (maximum height of 1.35 m, minimum height 1.10 m) and the joystick will be positioned in front as shown in fig. 5.2a. At the start of each experiment, the height will be adjusted such that the aural axis is aligned with the speakers. The subject was seated behind a car bonnet with a height of 1.70 m and a depth of 1 m. Note that this bonnet is covered with a blanket to limit the reflections. During the experiments, the room will be completely darkened to limit distractions and improve localization performance. As described in chapter 3, head and eye movements are a common measure for localization performance. In this experiment, head movements will be measured using the Movella Xsens DOT head tracker (serial number:40195bea809f0098) which includes a 3-axis accelerometer, a gyroscope, and a magnetometer. The eye-in-head position is fixated using a glasses-mounted laser pointer shown in fig. 5.2b [8]. Subjects wore glasses with a mounted laser pointer that projects its red beam onto a small, frame-attached disk (diameter, 1cm) at 38 cm in front of the listener's nose. Arm movements are measured using a gain-controlled joystick (Thrustmaster TCA sidestick Airbus edition joystick) which is positioned in front of the subject as shown in fig. 5.2a. A deflection of the joystick directly relates to a proportionate position change of the projected output.


(a)

(b)

Figure 5.2: a) Side view of the subject during the experiment. Note that the joystick is positioned in the middle and under experimental conditions the room is completely dark. (b) The subject wearing the head tracker as well laser pointer being mounted to the glasses. Note that these glasses are only worn during head trials.

The experiments will run on 3 different Dell desktops with a Linux operation system (Ubuntu 18.04.2 LTSS) on the Delft University Environment for Communication and Activation (DUECA). DUECA is a module-based real-time software. A big advantage of this software is that it accounts for software timing latency between the modules (joystick, speakers, head tracker) and the simulations can easily be distributed over multiple computers. The different sounds are generated using the signal processing toolbox 9.1 of MATLAB. This toolbox is user-friendly and allows for easy visualization of the created signals. For the data analyses of the results, this toolbox will also be used.

5.3. Task description and experimental procedure

In fig. 5.3 an example of an arm trial is given. Each trial starts with a white fixation dot at an azimuth of 0 degrees and an elevation of -5 degrees. From the subject's position, the dot has a width of 0.64 degrees. The output of the gain-controlled joystick is projected on the screen with a small red dot (0.32 degrees). A deflection of the joystick directly relates to a proportionate position change. Next, a random delay between 0 and 1.0 seconds, in steps of 250 ms, is introduced to prevent predictability in the stimulus presentation. After this, the stimulus (audio, visual or audiovisual) is presented all having a duration of 100 ms. The visual stimuli are a small gray dot (0.2 degrees). The audio stimuli consist of modulated high-pass noise. Depending on the delay interval, the response interval varies between 1000 - 2000 ms, resulting in a total trial time of 5000 ms.

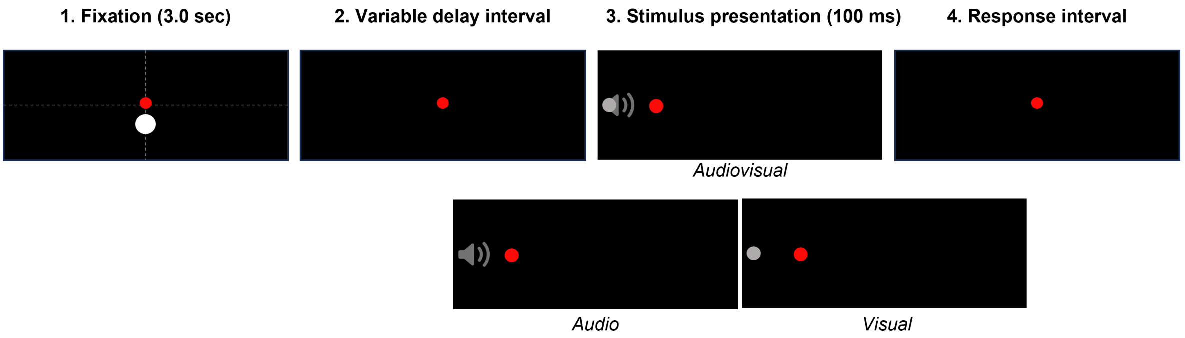


Figure 5.3: The trial structure for both head arm trials. Note that head trials are similar only the red joystick output is not shown 1) Fixation dot is positioned 5 degrees below the midline and shown at the start of each trial. Note that for arm trials the fixation of the eyes (white dot) is not aligned with the initial position of the joystick (red) in elevation 2) Random delay is introduced ranging between 0-1.0 seconds in steps of 250 ms 3) Stimulus (audio, visual or audiovisual) presentation in which the audio and visual stimuli are always presented at the same time and location (congruent) for audiovisual trials. 4) Depending on the delay interval, the response interval varies between 1000 - 2000 ms.

During the so-called arm trials, the subjects will be instructed to move the joystick output, represented by the red circle, as accurately and quickly as possible to the perceived location. This deliberate contradictory instruction aims to prevent subjects from focusing solely on one specific movement characteristic. Subjects will have the freedom to move the joystick in both elevation and azimuth. Head movements will also be measured during the arm trials, and subjects will be allowed to move their heads freely and no specific instructions regarding head or eye movements will be given. Also, no glasses will be used to fixate the eye-in-head position in this trial. To fixate the eye-in-head position during head trials, subjects will wear glasses with a mounted laser pointer and will be asked to fixate on the projected laser signal with their eyes. Task instructions will be the same as for the arm trials but now the subject needs to align the laser pointer with the perceived location. Similar to the arm trials, they will be instructed to perform the movement as accurately and quickly as possible.

Based on comparable experiments reported in the literature, the inclusion of a total of 10 subjects is aimed for [52] [32]. No specific baseline test will be performed, but practice runs will be conducted to familiarize the subjects with the tasks and to verify that the stimuli can be heard and seen by the subjects. The experiment procedure for each subject is depicted in fig. 5.4. A short briefing will initiate each experiment, followed by several blocks of measurements. At the onset of each experiment, the head tracker will be calibrated through a calibration experiment performed by the subject. Subsequently, the subject will engage in 10 practice trials for both arm and head movements to gain an understanding of the expected stimuli and become acquainted with the dynamics of the joystick. Each block will include 135 trials, consisting of joystick movements (arm trials) or head movements. The trials will involve a combination of audio, visual, and audiovisual stimuli within a single block. Each block will consist of 3 repetitions per stimulus condition per target location.

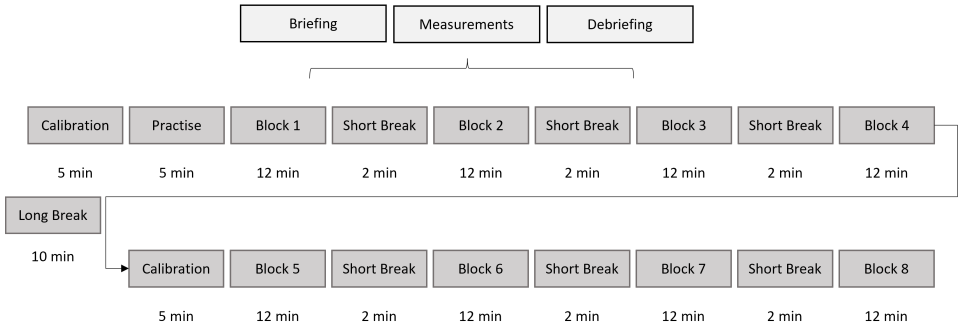


Figure 5.4: Schematic overview of the experiment procedure for the final experiments. During each block, either arm or head movements will be performed whilst audio, visual, and audiovisual trials are mixed within the blocks.

5.4. Conclusion

This research aims to analyze the response latency and orienting accuracy of 2D joystick movements in response to visual, auditory, and combined audio-visual targets, and compare them with the parameters of head movements in an echoic environment. It is hypothesized that audiovisual benefits will be present in the head movements as well as joystick movements. Benefits originate from benefits observed in the sensory processing such as increased accuracy and decreased reaction time for multimodal stimuli.

Bibliography

[1] FR Alex, Richurd A Peters, and RL Stapleford. *Experiments and a model for pilot dynamics with visual and motion inputs*. Tech. rep. NASA, 1969.

[2] Pavlo Bazilinskyy et al. “Blind driving by means of auditory feedback”. In: *IFAC-PapersOnLine* 49.19 (2016), pp. 525–530.

[3] Durand R Begault and Marc T Pittman. “Three-dimensional audio versus head-down traffic alert and collision avoidance system displays”. In: *The International Journal of Aviation Psychology* 6.1 (1996), pp. 79–93.

[4] Marian E Berryhill, Tanya Chiu, and Howard C Hughes. “Smooth pursuit of nonvisual motion”. In: *Journal of neurophysiology* 96.1 (2006), pp. 461–465.

[5] Paul Bertelson and Gisa Aschersleben. “Automatic visual bias of perceived auditory location”. In: *Psychonomic bulletin & review* 5 (1998), pp. 482–489.

[6] B Biguer, C Prablanc, and M Jeannerod. “The contribution of coordinated eye and head movements in hand pointing accuracy”. In: *Experimental brain research* 55.3 (1984), pp. 462–469.

[7] Jens Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[8] Peter Bremen, Marc M van Wanrooij, and A John van Opstal. “Pinna cues determine orienting response modes to synchronous sounds in elevation”. In: *Journal of Neuroscience* 30.1 (2010), pp. 194–204.

[9] Philip Joris Peter Bremen. *Woods Hole Summer course: ‘Biology of the Inner Ear - Systems part 3*. 2022.

[10] José A García-Uceda Calvo, Marc M van Wanrooij, and A John Van Opstal. “Adaptive response behavior in the pursuit of unpredictably moving sounds”. In: *Eneuro* 8.3 (2021).

[11] BD Corneil et al. “Auditory-visual interactions subserving goal-directed saccades in a complex scene”. In: *Journal of Neurophysiology* 88.1 (2002), pp. 438–454.

[12] Rachel Ege, A Opstal, and Marc M Van Wanrooij. “Accuracy-precision trade-off in human sound localisation”. In: *Scientific reports* 8.1 (2018), pp. 1–12.

[13] Elizabeth Fehrer and David Raab. “Reaction time to stimuli masked by metacontrast.” In: *Journal of experimental psychology* 63.2 (1962), p. 143.

[14] Maarten A Frens and A John Van Opstal. “Visual-auditory interactions modulate saccade-related activity in monkey superior colliculus”. In: *Brain research bulletin* 46.3 (1998), pp. 211–224.

[15] Stan CAM Gielen, Richard A Schmidt, and Pieter JM Van Den Heuvel. “On the nature of intersensory facilitation of reaction time”. In: *Perception & psychophysics* 34 (1983), pp. 161–168.

[16] Hieronymus HLM Goossens and A John Van Opstal. “Human eye-head coordination in two dimensions under different sensorimotor conditions”. In: *Experimental Brain Research* 114.3 (1997), pp. 542–560.

[17] Lawrence K Harrington and Carol K Peck. “Spatial disparity affects visual-auditory interactions in human sensorimotor processing”. In: *Experimental Brain Research* 122 (1998), pp. 247–252.

[18] Paul M Hofman, Jos GA Van Riswick, and A John Van Opstal. “Relearning sound localization with new ears”. In: *Nature neuroscience* 1.5 (1998), pp. 417–421.

[19] Jeannette D Hoit and Gary Weismer. *Foundations of speech and hearing: Anatomy and physiology*. Plural Publishing, 2016.

[20] Ruud Hosman and Henk Stassen. “Pilot’s perception and control of aircraft motions”. In: *IFAC Proceedings Volumes* 31.26 (1998), pp. 311–316.

[21] Howard C Hughes et al. “Visual-auditory interactions in sensorimotor processing: saccades versus manual responses.” In: *Journal of Experimental Psychology: Human Perception and Performance* 20.1 (1994), p. 131.

[22] Barbro Birgitta Johansson. “Multisensory stimulation in stroke rehabilitation”. In: *Frontiers in human neuroscience* 6 (2012), p. 60.

[23] Niilo Konttinen et al. “The effects of augmented auditory feedback on psychomotor skill learning in precision shooting”. In: *Journal of Sport and Exercise Psychology* 26.2 (2004), pp. 306–316.

[24] Ewan A Macpherson and John C Middlebrooks. “Listener weighting of cues for lateral angle: the duplex theory of sound localization revisited”. In: *The Journal of the Acoustical Society of America* 111.5 (2002), pp. 2219–2236.

[25] Richard Magill and David I Anderson. *Motor learning and control*. McGraw-Hill Publishing New York, 2010.

[26] James C Makous and John C Middlebrooks. “Two-dimensional sound localization by human listeners”. In: *The journal of the Acoustical Society of America* 87.5 (1990), pp. 2188–2200.

[27] M Alex Meredith, James W Nemitz, and Barry E Stein. “Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors”. In: *Journal of Neuroscience* 7.10 (1987), pp. 3215–3229.

[28] M Alex Meredith and Barry E Stein. “Spatial factors determine the activity of multisensory neurons in cat superior colliculus”. In: *Brain research* 365.2 (1986), pp. 350–354.

[29] Donald H Mershon and L Edward King. “Intensity and reverberation as factors in the auditory perception of egocentric distance”. In: *Perception & Psychophysics* 18 (1975), pp. 409–415.

[30] John C Middlebrooks. “Sound localization”. In: *Handbook of clinical neurology* 129 (2015), pp. 99–116.

[31] Jeff Miller. “Divided attention: Evidence for coactivation with redundant signals”. In: *Cognitive psychology* 14.2 (1982), pp. 247–279.

[32] Nynke Niehof et al. “Dynamic Auditory Localisation: Head Tracking of Virtual Moving Sounds”. In: *From the Editors-in-Chief* (2014), p. 33.

[33] Daan M Pool. “Objective evaluation of flight simulator motion cueing fidelity through a cybernetic approach”. In: (2012).

[34] FJ Praamstra et al. “Function of Attitude Perception in Human Control Behavior in Target Tracking Tasks”. In: *AIAA Modeling and Simulation Technologies Conference 2008*. Curran. 2008, pp. 953–978.

[35] William H Press et al. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[36] Cyril Rashbass. “The relationship between saccadic and smooth tracking eye movements”. In: *The Journal of physiology* 159.2 (1961), p. 326.

[37] Elden F Ray. “Industrial Noise Series, Part IV, Modeling Sound Propagation”. In: *June* 16 (2010), p. 2010.

[38] John William Strutt Baron Rayleigh. *The theory of sound*. Vol. 2. Macmillan, 1896.

[39] Do A Robinson. “The mechanics of human smooth pursuit eye movement.” In: *The Journal of Physiology* 180.3 (1965), p. 569.

[40] Vincent Roggerone et al. “Auditory motion perception emerges from successive sound localizations integrated over time”. In: *Scientific Reports* 9.1 (2019), pp. 1–9.

[41] Giulio Rosati et al. “Effect of task-related continuous auditory feedback during learning of tracking motion exercises”. In: *Journal of neuroengineering and rehabilitation* 9.1 (2012), pp. 1–13.

[42] Richard A Schmidt et al. *Motor control and learning: A behavioral emphasis*. Human kinetics, 2018.

[43] Jeffery A Schroeder. “Evaluation of simulation motion fidelity criteria in the vertical and directional axes”. In: *Journal of the American Helicopter Society* 41.2 (1996), pp. 44–57.

[44] Roland Sigrist et al. “Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review”. In: *Psychonomic Bulletin and Review* 20 (1 2013), pp. 21–53. ISSN: 10699384. DOI: `10.3758/s13423-012-0333-8`.

[45] Ronald Small et al. “A pilot spatial orientation aiding system”. In: *AIAA 5th ATIO and16th Lighter-Than-Air Sys Tech. and Balloon Systems Conferences*. 2005, p. 7431.

[46] Robert D Sorkin et al. “An exploratory study of the use of movement-correlated cues in an auditory head-up display”. In: *Human Factors* 31.2 (1989), pp. 161–166.

[47] Barry E Stein and M Alex Meredith. “Multisensory integration”. In: *Annals of the New York Academy of Sciences* 608.1 (1990), pp. 51–70.

[48] Barry E Stein and M Alex Meredith. *The merging of the senses*. MIT press, 1993.

[49] Ryan A Stevenson et al. “Identifying and quantifying multisensory integration: a tutorial review”. In: *Brain topography* 27 (2014), pp. 707–730.

[50] Michael H Thaut et al. “Rhythmic auditory stimulation in gait training for Parkinson’s disease patients”. In: *Movement disorders: official journal of the Movement Disorder Society* 11.2 (1996), pp. 193–200.

[51] John Van Opstal. *The auditory system and human sound-localization behavior*. Academic Press, 2016.

[52] Marc M Van Wanrooij, Peter Bremen, and A John Van Opstal. “Acquired prior knowledge modulates audiovisual integration”. In: *European Journal of Neuroscience* 31.10 (2010), pp. 1763–1771.

[53] Marc M Van Wanrooij and A John Van Opstal. “Contribution of head shadow and pinna cues to chronic monaural sound localization”. In: *Journal of Neuroscience* 24.17 (2004), pp. 4163–4171.

[54] Marc M Van Wanrooij and A John Van Opstal. “Relearning sound localization with a new ear”. In: *Journal of Neuroscience* 25.22 (2005), pp. 5413–5424.

[55] Marc M Van Wanrooij et al. “The effect of spatial–temporal audiovisual disparities on saccades in a complex scene”. In: *Experimental brain research* 198 (2009), pp. 425–437.

[56] Edward W. Vinje and Edward T. Pitkin. “Human Operator Dynamics for Aural Compensatory Tracking”. In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-2.4 (1972), pp. 504–512. DOI: `10.1109/TSMC.1972.4309160`.

[57] Simon Coenraad Fransiscus Vrouwenvelder. *A Cybernetic Approach To Point-To-Point Movements*. Tech. rep. Delft University of Technology, 2020.

[58] Mark T Wallace, M Alex Meredith, and Barry E Stein. “Multisensory integration in the superior colliculus of the alert cat”. In: *Journal of neurophysiology* 80.2 (1998), pp. 1006–1010.

[59] Frederic L Wightman and Doris J Kistler. “Headphone simulation of free-field listening. I: stimulus synthesis”. In: *The Journal of the Acoustical Society of America* 85.2 (1989), pp. 858–867.

[60] Frederic L Wightman and Doris J Kistler. “The dominant role of low-frequency interaural time differences in sound localization”. In: *The Journal of the Acoustical Society of America* 91.3 (1992), pp. 1648–1661.

[61] Robert S Woodworth and Harold Schlosberg. “Experimental psychology, rev”. In: (1954).

[62] Marcel P Zwiers, A John Van Opstal, and Gary D Paige. “Plasticity in human sound localization induced by compressed spatial vision”. In: *Nature neuroscience* 6.2 (2003), pp. 175–181.

[63] MP Zwiers, AJ Van Opstal, and JRM Cruysberg. “A spatial hearing deficit in early-blind humans”. In: (2001).

III

Paper Appendices

A
Endpoint Corrections Head Responses

A.1. Endpoint corrections

To validate the accuracy of angles measured by the head tracker, a calibration experiment was conducted. A visual target remained visible for 5 seconds, during which participants were instructed to move accurately. Endpoints were determined for individual responses, focusing on the average measured angle between 2 to 3 seconds following stimulus onset. As discrepancies existed between these measured endpoints and the intended target positions, a two-layer neural network was employed for correction. This neural network was trained based on the target location and the corresponding endpoints for each participant. Among the participants, 8 out of 12 performed the calibration twice as the long break took place in between the head response blocks.

In fig. A.1, the raw and calibrated endpoints are displayed for participant 1. In the lower two figures, the error for the calibrated endpoint is shown for azimuth and elevation. In the upper left graph, presenting endpoints from the raw data, it can clearly be observed that measured points for peripheral locations are located more inwards compared to the target locations. For peripheral locations, these errors are largest whilst they decrease for locations closer to the centerline. The calibrated endpoints of participant 1 showed the largest error. In contrast, participant 8 showed the lowest error, and results are depicted in fig. A.2.

In order to verify the applicability of the neural network, which was exclusively trained on endpoint data, to continuous traces, we applied the trained network to the time traces of the calibration experiment. Once more, examples of participant 1 (fig. A.3 and participant 8 (fig. A.4 are shown.

The neural network's adaptability to continuous traces already suggests a favorable response to untrained data points. To further test this observation, we partitioned the endpoints into training and testing sets. Subsequently, we calculated the minimum (min) and maximum (max) errors per target locations for all endpoints (E_{all}), the training set (E_{tr}), and the test set (E_{test}). Detailed outcomes from all calibration experiments are presented in table A.1.

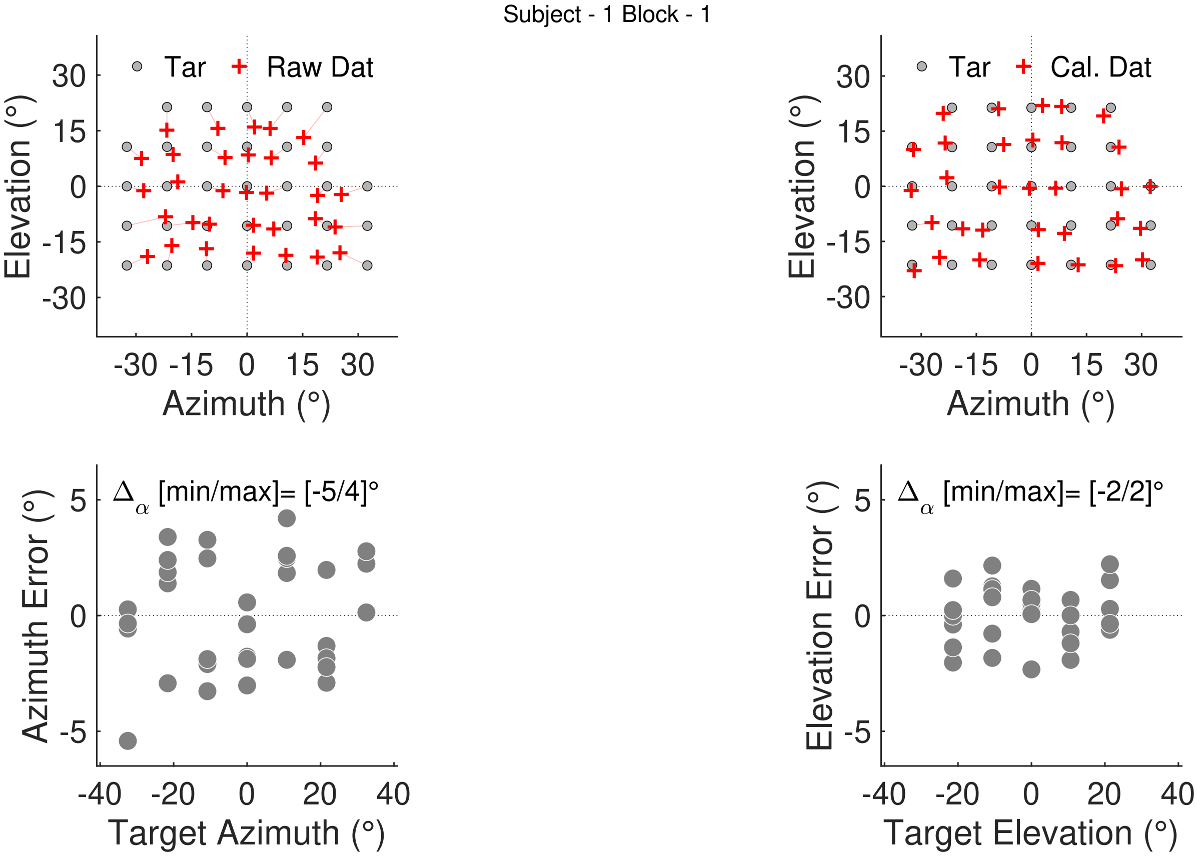


Figure A.1: Endpoint calibration of participant 1 is depicted, displaying both the unprocessed (Raw) and calibrated (Cal.) data alongside their respective targets (Tar). Note that the calibration of participant 1 demonstrated the largest errors when compared to all other participants.

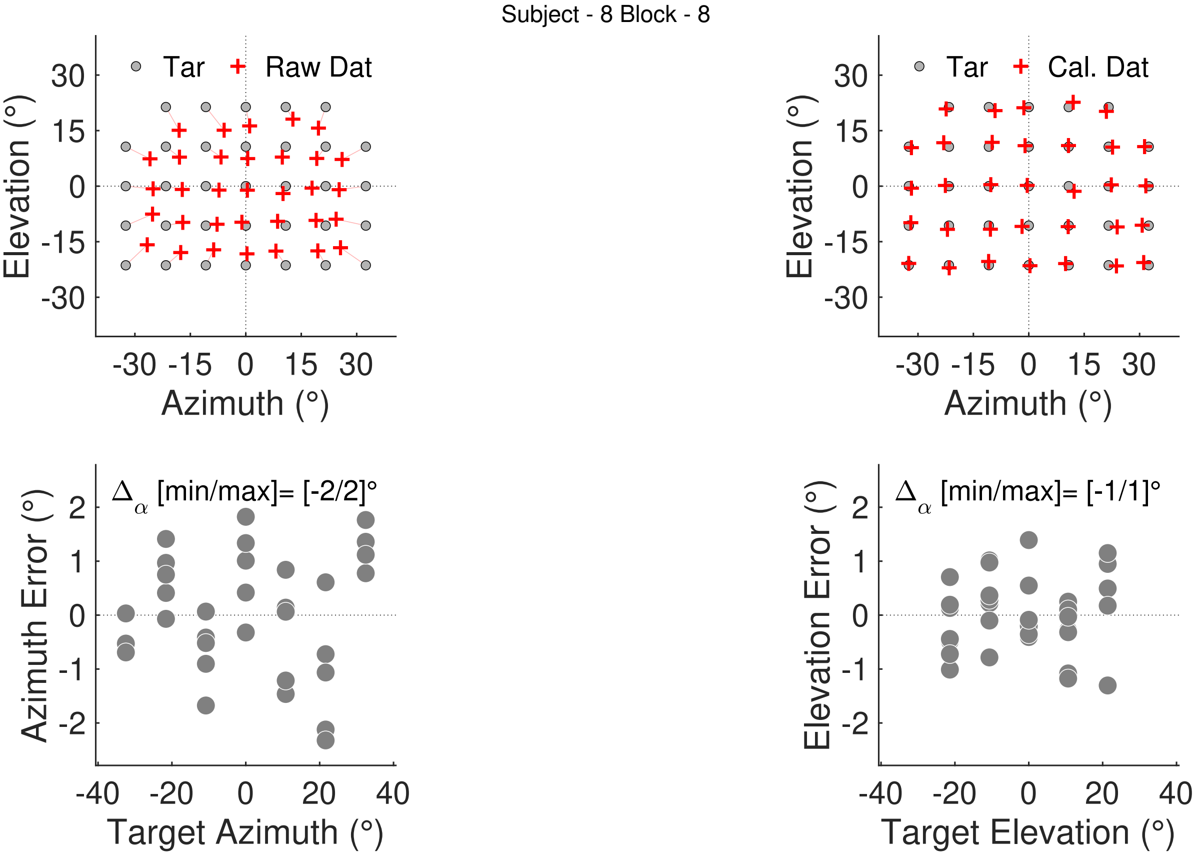


Figure A.2: Endpoint calibration of participant 8 is depicted, displaying both the unprocessed (Raw) and calibrated (Cal.) data alongside their respective targets (Tar). Note that the calibration of participant 8 demonstrated the smallest errors when compared to all other participants.

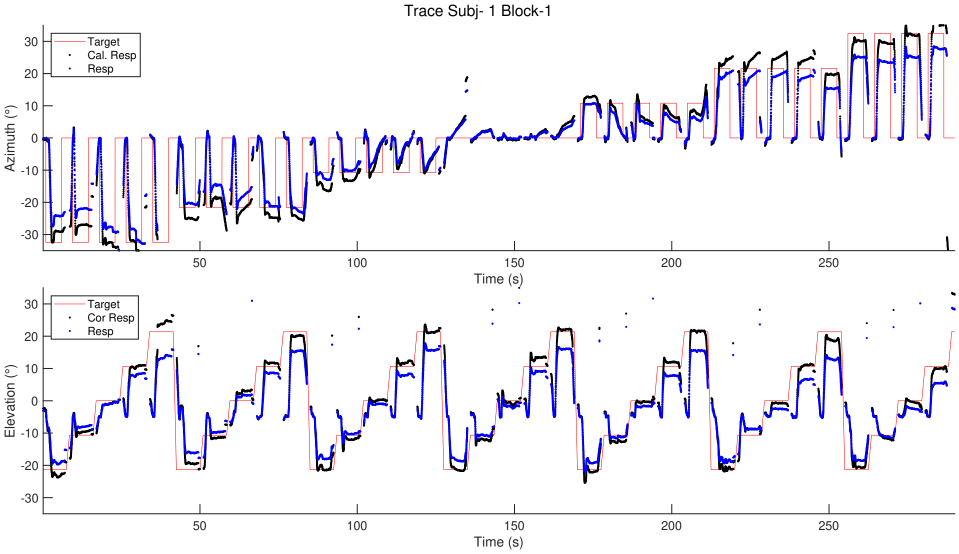


Figure A.3: Time trace of calibration experiment performed by participant 1

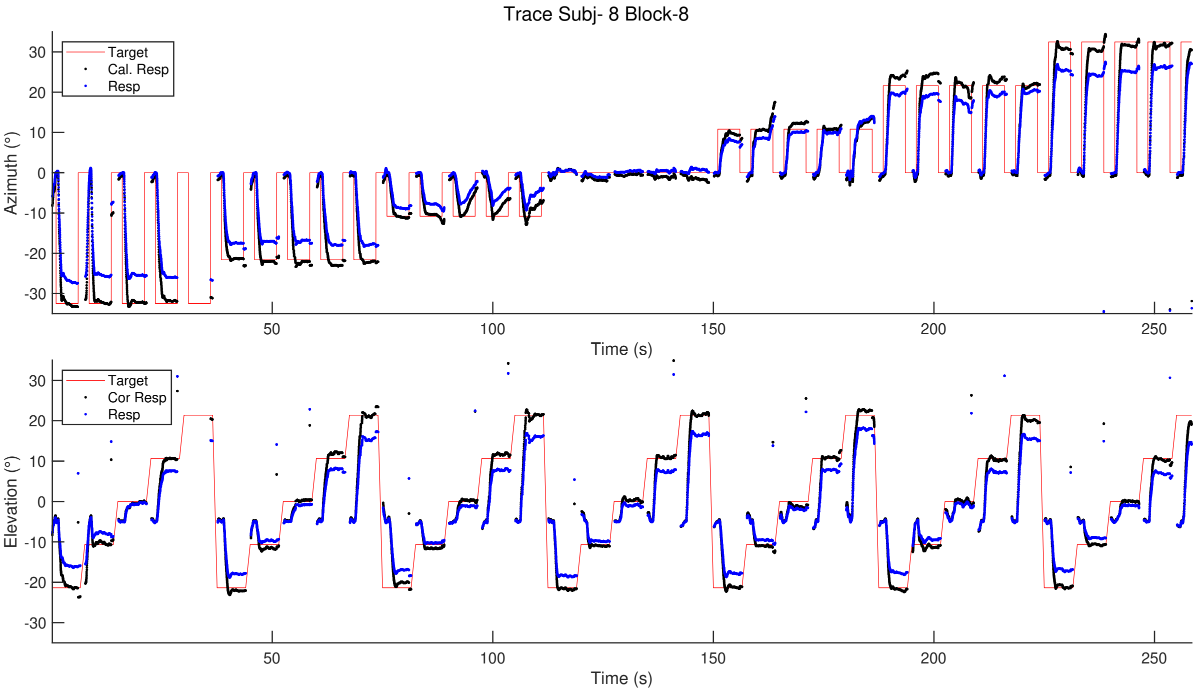


Figure A.4: Time trace of calibration experiment performed by participant 8

Table A.1: Specifications of azimuth and elevation errors for calibrated endpoints.

	Azimuth			Elevation		
	E_{all} min/max	E_{tr} min/max	E_{test} min/max	E_{all} min/max	E_{tr} min/max	E_{test} min/max
S1 - B1	[-5 / 4] °	[-3 / 3] °	[-7 / 4] °	[-2 / 2] °	[-2 / 2] °	[-2 / 1] °
S1 - B7	[-3 / 4] °	[-2 / 3] °	[-3 / 4] °	[-2 / 1] °	[-2 / 3] °	[-1 / 2] °
S2 - B1	[-3 / 3] °	[-1 / 2] °	[-7 / 3] °	[-2 / 4] °	[0 / 0] °	[-6 / 6] °
S2 - B7	-	-	-	-	-	-
S3 - B1	[-3 / 3] °	[-2 / 1] °	[-4 / 3] °	[-1 / 1] °	[-1 / 1] °	[-1 / 2] °
S3 - B7	[-3 / 3] °	[-2 / 3] °	[-2 / 3] °	[-1 / 1] °	[-1 / 1] °	[-1 / 1] °
S4 - B1	[-3 / 2] °	[-3 / 2] °	[-3 / 5] °	[-1 / 1] °	[-1 / 1] °	[-3 / 4] °
S4 - B7	[-4 / 4] °	[-1 / 2] °	[-4 / 7] °	[-5 / 2] °	[-2 / 2] °	[-6 / 2] °
S5 - B1	[-3 / 2] °	[-3 / 2] °	[-1 / 2] °	[-1 / 1] °	[-1 / 1] °	[-1 / 2] °
S5 - B7	[-4 / 2] °	[-6 / 4] °	[-1 / 1] °	[-3 / 2] °	[-4 / 1] °	[-1 / 2] °
S6 - B1	[-1 / 1] °	[-1 / 1] °	[-4 / 1] °	[-1 / 1] °	[-1 / 1] °	[0 / 2] °
S6 - B7	-	-	-	-	-	-
S7 - B1	[-2 / 2] °	[-1 / 1] °	[-2 / 6] °	[-3 / 2] °	[-1 / 0] °	[-5 / 0] °
S7 - B7	-	-	-	-	-	-
S8- B1	[-2 / 2] °	[-2 / 2] °	[-4 / 2] °	[-1 / 1] °	[-1 / 1] °	[-2 / 2] °
S8 - B7	-	-	-	-	-	-
S9 - B1	[-3 / 3] °	[-2 / 2] °	[-3 / 4] °	[-3 / 2] °	[-2 / 2] °	[-2 / 2] °
S9 - B7	[-2 / 3] °	[-2 / 3] °	[-1 / 1] °	[0 / 0] °	[0 / 0] °	[-1 / 1] °
S10 - B1	[-1 / 1] °	[-1 / 1] °	[-2 / 2] °	[-1 / 1] °	[0 / 1] °	[-1 / 2] °
S10 - B7	[-2 / 2] °	[0 / 0] °	[-3 / 3] °	[-1 / 1] °	[0 0] °	[-2 / 1] °
S11 - B1	[-3 / 2] °	[0 / 0] °	[-4 / 2] °	[-1 / 1]	[0 / 0] °	[-2 / 0] °
S11 - B7	[-3 / 2] °	[-3 / 2] °	[-3 / 1] °	[-1 / 1] °	[-1 / 1] °	[-1 / 1] °
S12 - B1	[-2 / 2] °	[-1 / 1] °	[-3 / 6] °	[-2 / 2] °	[-1 / 1] °	[-5 / 2] °
S12 - B7	[-5 / 3] °	[-1 / 1] °	[-7 / 3] °	[-3 / 2] °	[0 / 1] °	[-4 / 2] °

A.2. Time Corrections

In our experimental setup, head responses were measured using a head tracker. The initiation of measurement was governed by the DUECA software, which triggered the head tracker at the start of each trial. During the measurement process, the head tracker transmitted data via a Bluetooth connection, subsequently processed by the DUECA software. Notably, while the initiation signal for measurement was sent at the onset of each trial, there existed substantial variability in the temporal occurrence of the first data point received by the DUECA software across different trials. The maximal disparities, for each block per subject (S) are provided in Table A.2. Given this variance, it remains uncertain whether this delay primarily results from the delayed initiation of the measurements or from data buffering before transmission. For consistency, we designated the instant at which the first data point was received by DUECA as the official commencement of measurement, a choice that inherently introduces variability into the measured reaction times.

Furthermore, we have identified certain issues regarding the uniqueness of time steps. To provide a clear explanation, we establish a distinction between the "internal head tracker" clock and the "overall" clock, which is established by the DUECA software. These clocks function independently and are not synchronized, resulting in clock drift. Each output angle is associated with both a DUECA timestamp (representing the time of data reception) and a head tracker time step (representing the time of data measurement). We have noticed instances where different output angles share the same DUECA timestamp, while each output angle maintains a distinct head tracker timestamp. This observation could potentially suggest that the head tracker accumulates multiple data points before transmitting the information.

Considering that all the other modules in our software were synchronized with the DUECA clock, we decided to make use of the DUECA timestamp. In order to avoid reducing the sampling frequency by directly discarding data points or averaging measurements, we employed the time difference calculated by the internal clock to establish a distinctive DUECA timestamp. The DUECA timestamp for the "last" received location signal corresponded directly to its timestamp, while the timestamp for the "first" location signal was determined by subtracting the time difference in head tracker timing between the two points from the DUECA timestamp, resulting in unique timestamps.

Table A.2: Timing discrepancies between DUECA software and the head tracker measurements for all participants

Time difference (ms)

	Max B1	Max B2	Max B3	Max B4	25t perc.	75th perc.	Median
S1	127	177	127	257	123	143	133
S2	280	267	233	337	123	153	133
S3	197	217	303	253	123	137	133
S4	210	273	193	173	123	140	130
S5	210	183	267	60	130	143	130
S6	247	333	247	247	113	133	133
S7	90	207	187	80	123	143	133
S8	163	297	203	207	113	133	127
S9	250	243	320	327	123	143	133
S10	200	317	163	180	123	147	133
S11	173	230	107	207	123	133	133
S12	260	237	303	183	123	150	133

B

Supporting Results

B.1. Participant 1

B.1.1. Stimulus response plots

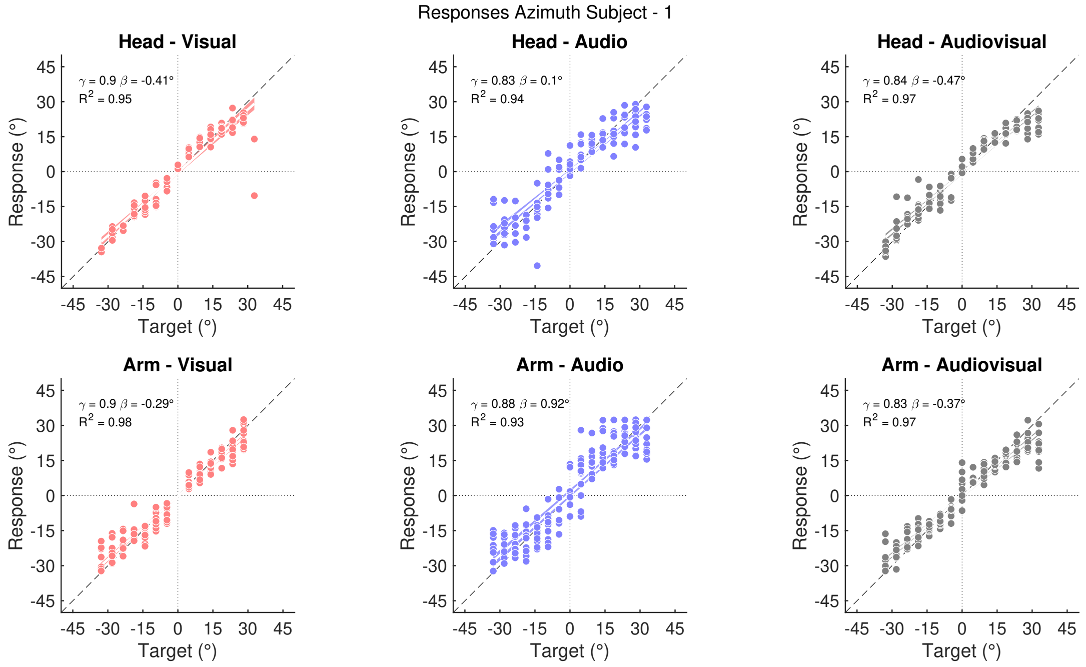


Figure B.1: Stimulus response plots of participant 1 for all modalities and effectors. The one-on-one relations between the target and response are presented by the black line while the optimal linear fit details are indicated in the top left corner. Note the differences between the modalities.

B.1.2. Reaction time compared to azimuth error

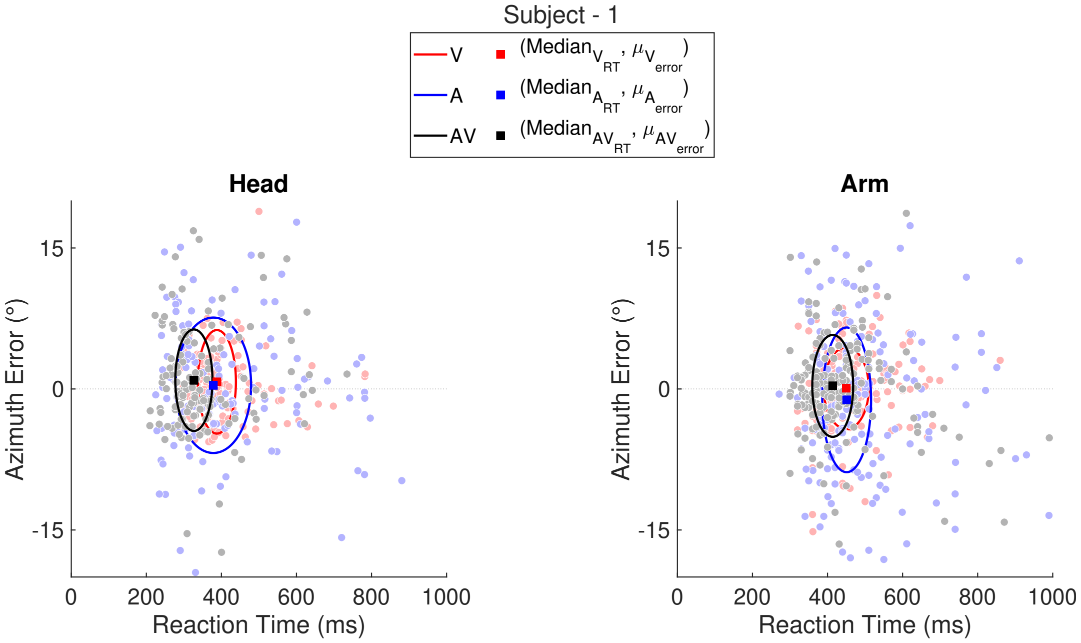


Figure B.2: Localization error as a function of reaction time for auditory (A, blue), visual (V, red), and audio-visual (AV, black) responses of participant 1. Squares indicate the median reaction time (x-axis) and mean localization error (y-axis). Ellipses depict the interquartile range (75th-25th percentile) of the reaction times (x-axis) and the standard deviation of the errors (y-axis). Note the difference in ellipse sizes for different modalities

B.1.3. Reaction time per experiment block

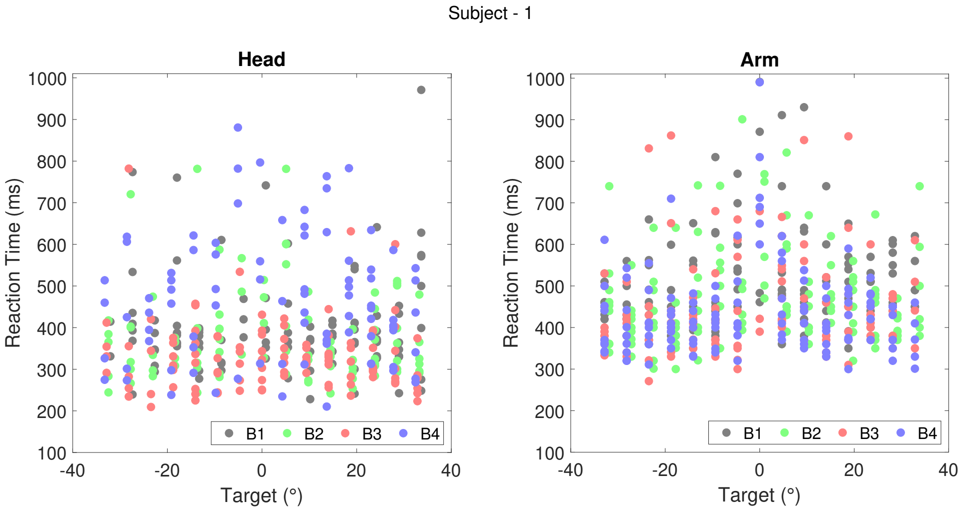


Figure B.3: Reaction times are plotted according to target locations for participant 1, with distinct colors denoting responses from various blocks (B).

B.1.4. Time discrepancies per trial

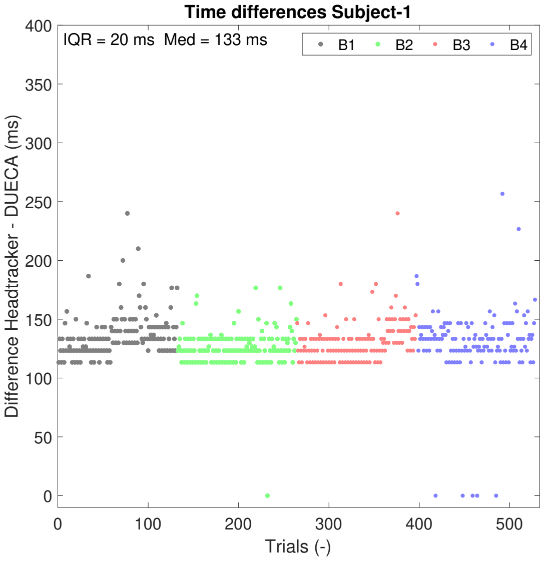


Figure B.4: Temporal discrepancies between the head tracker and DUECA software are plotted. Sequentially ordered trials are color-coded by block, with trial numbers cumulative across previous blocks.

B.2. Participant 2

B.2.1. Stimulus response plots

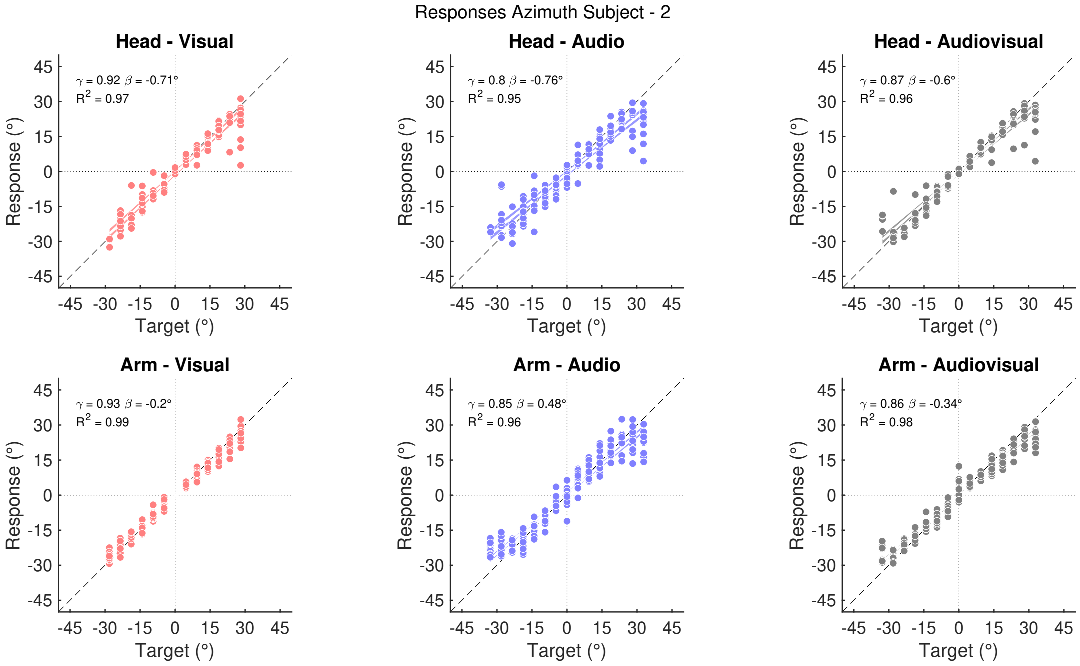


Figure B.5: Stimulus response plots of participant 2 for all modalities and effectors. The one-on-one relations between the target and response are presented by the black line while the optimal linear fit details are indicated in the top left corner. Note the differences between the modalities.

B.2.2. Reaction time compared to azimuth error

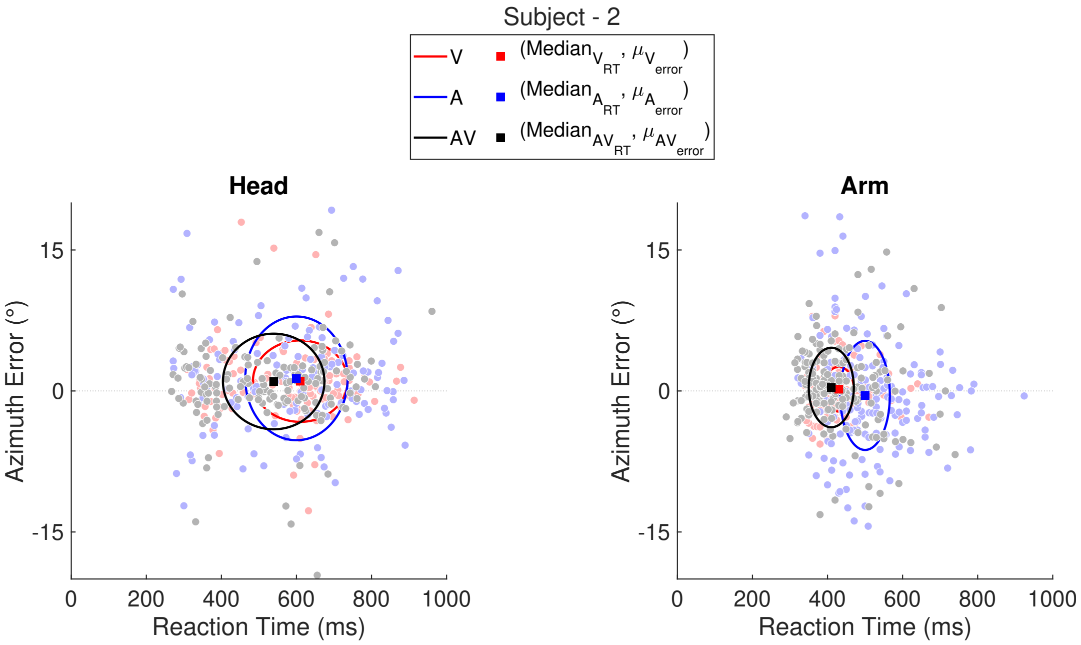


Figure B.6: Localization error as a function of reaction time for auditory (A, blue), visual (V, red), and audio-visual (AV, black) responses of participant 2. Squares indicate the median reaction time (x-axis) and mean localization error (y-axis). Ellipses depict the interquartile range (75th-25th percentile) of the reaction times (x-axis) and the standard deviation of the errors (y-axis). Note the difference in ellipse sizes for different modalities

B.2.3. Reaction time per experiment block

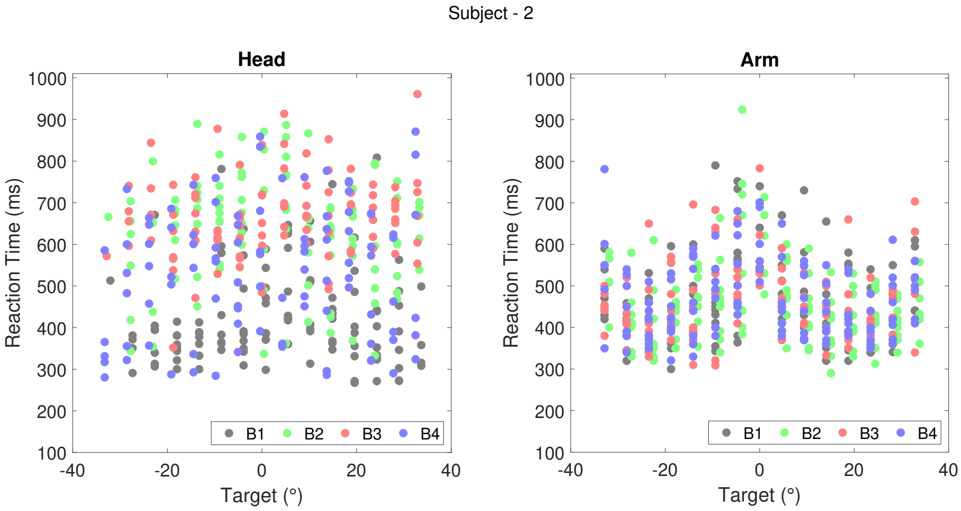


Figure B.7: Reaction times are plotted according to target locations for participant 2, with distinct colors denoting responses from various blocks (B).

B.2.4. Time discrepancies per trial

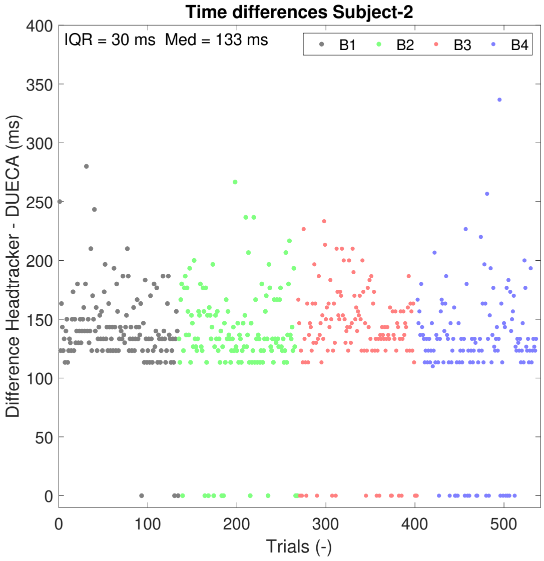


Figure B.8: Temporal discrepancies between the head tracker and DUECA software are plotted. Sequentially ordered trials are color-coded by block, with trial numbers cumulative across previous blocks.

B.3. Participant 3

B.3.1. Stimulus response plots

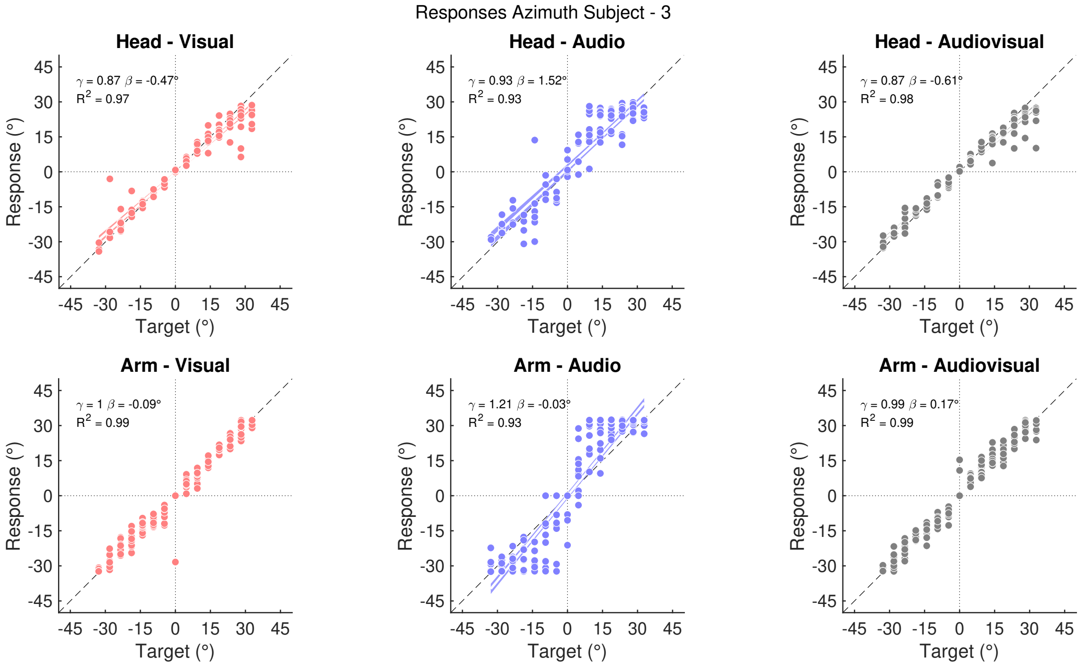


Figure B.9: Stimulus response plots of participant 3 for all modalities and effectors. The one-on-one relations between the target and response are presented by the black line while the optimal linear fit details are indicated in the top left corner. Note the differences between the modalities.

B.3.2. Reaction time compared to azimuth error

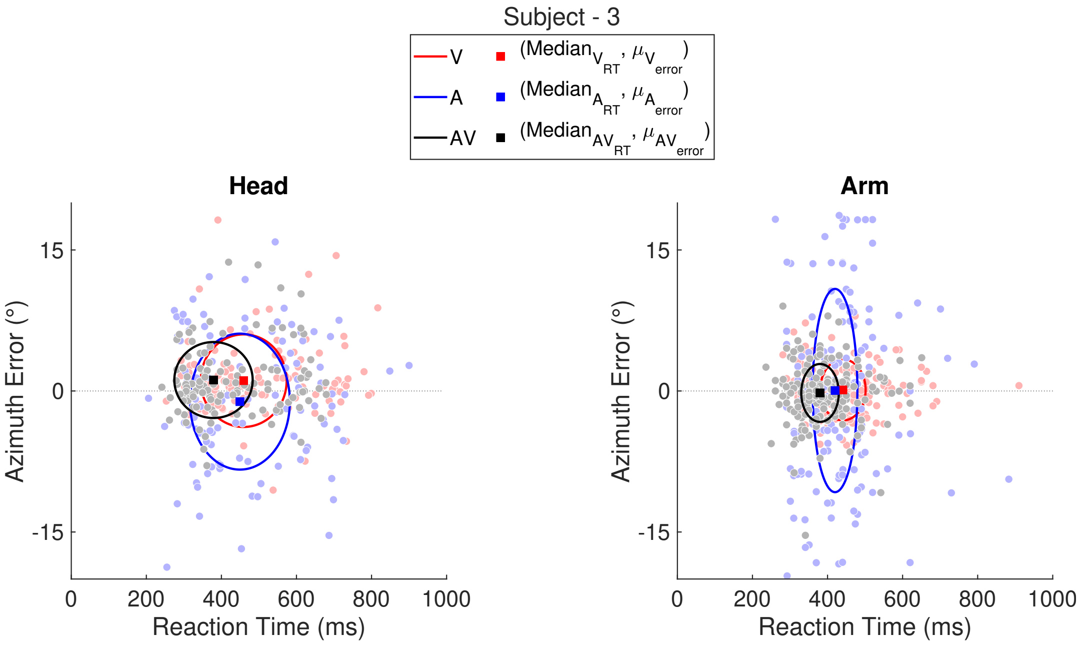


Figure B.10: Localization error as a function of reaction time for auditory (A, blue), visual (V, red), and audio-visual (AV, black) responses of participant 3. Squares indicate the median reaction time (x-axis) and mean localization error (y-axis). Ellipses depict the interquartile range (75th-25th percentile) of the reaction times (x-axis) and the standard deviation of the errors (y-axis). Note the difference in ellipse sizes for different modalities

B.3.3. Reaction time per experiment block

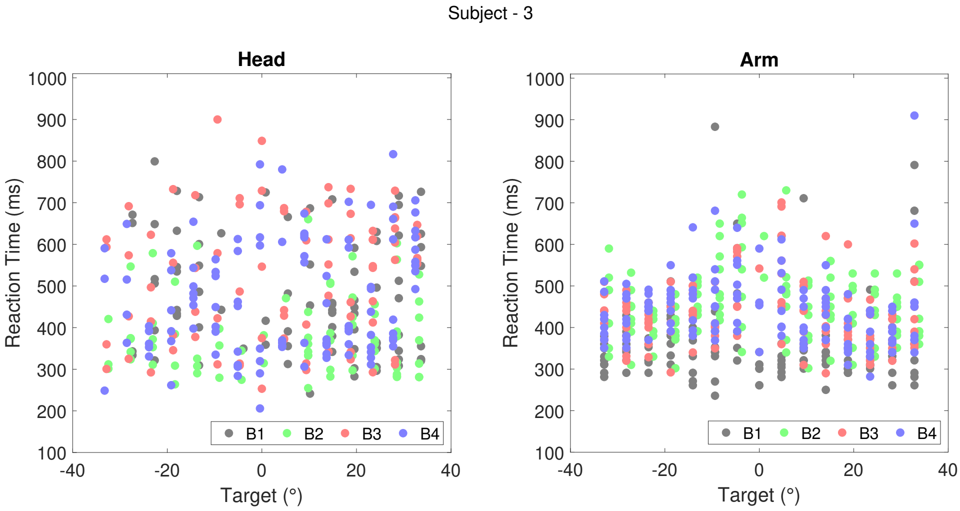


Figure B.11: Reaction times are plotted according to target locations for participant 3, with distinct colors denoting responses from various blocks (B).

B.3.4. Time discrepancies per trial

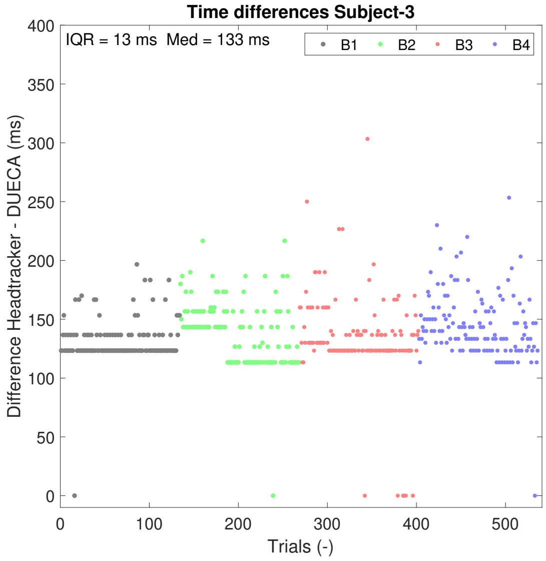


Figure B.12: Temporal discrepancies between the head tracker and DUECA software are plotted. Sequentially ordered trials are color-coded by block, with trial numbers cumulative across previous blocks.

B.4. Participant 4

B.4.1. Stimulus response plots

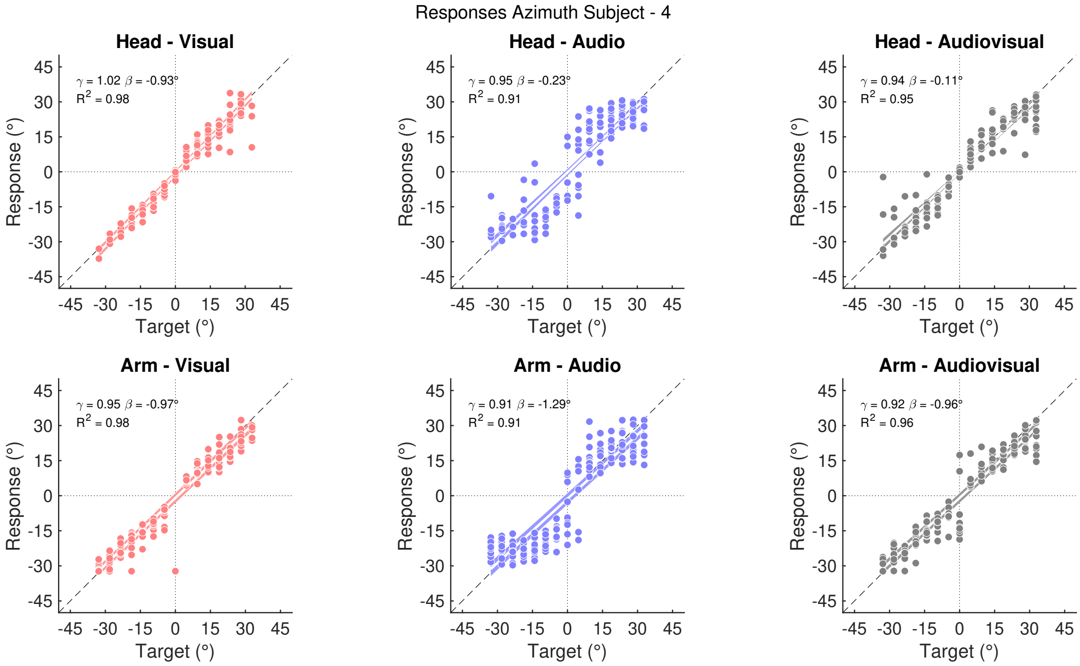


Figure B.13: Stimulus response plots of participant 4 for all modalities and effectors. The one-on-one relations between the target and response are presented by the black line while the optimal linear fit details are indicated in the top left corner. Note the differences between the modalities.

B.4.2. Reaction time compared to azimuth error

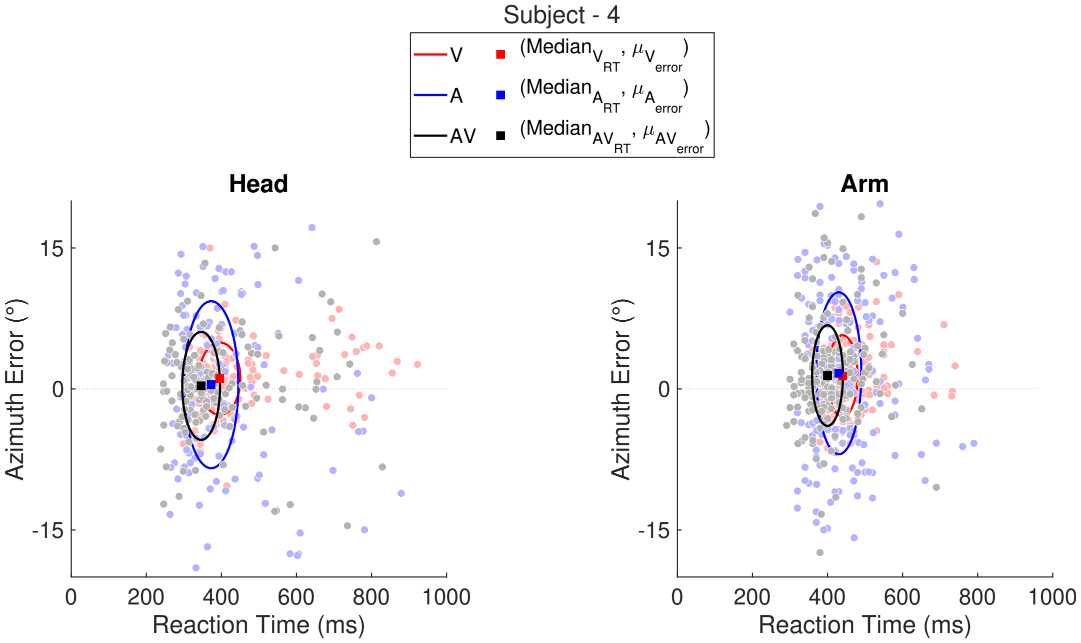


Figure B.14: Localization error as a function of reaction time for auditory (A, blue), visual (V, red), and audio-visual (AV, black) responses of participant 4. Squares indicate the median reaction time (x-axis) and mean localization error (y-axis). Ellipses depict the interquartile range (75th-25th percentile) of the reaction times (x-axis) and the standard deviation of the errors (y-axis). Note the difference in ellipse sizes for different modalities

B.4.3. Reaction time per experiment block

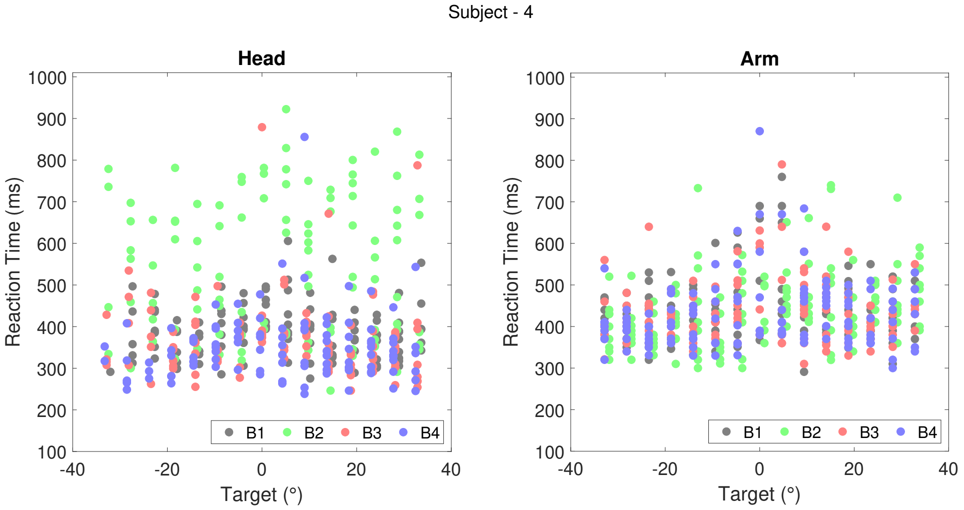


Figure B.15: Reaction times are plotted according to target locations for participant 4, with distinct colors denoting responses from various blocks (B).

B.4.4. Time discrepancies per trial

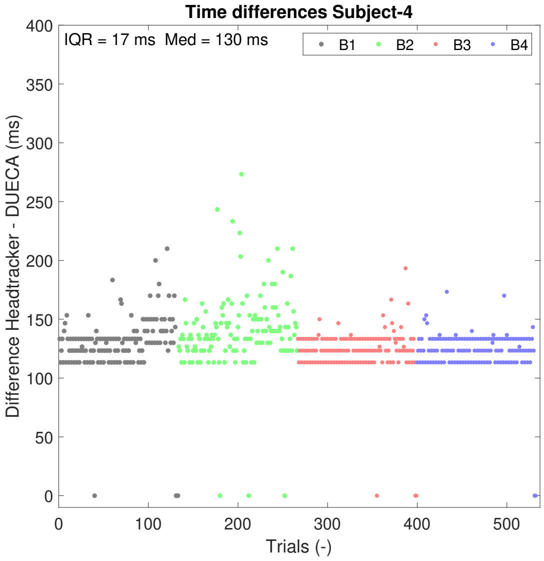


Figure B.16: Temporal discrepancies between the head tracker and DUECA software are plotted. Sequentially ordered trials are color-coded by block, with trial numbers cumulative across previous blocks.

B.5. Participant 5

B.5.1. Stimulus response plots

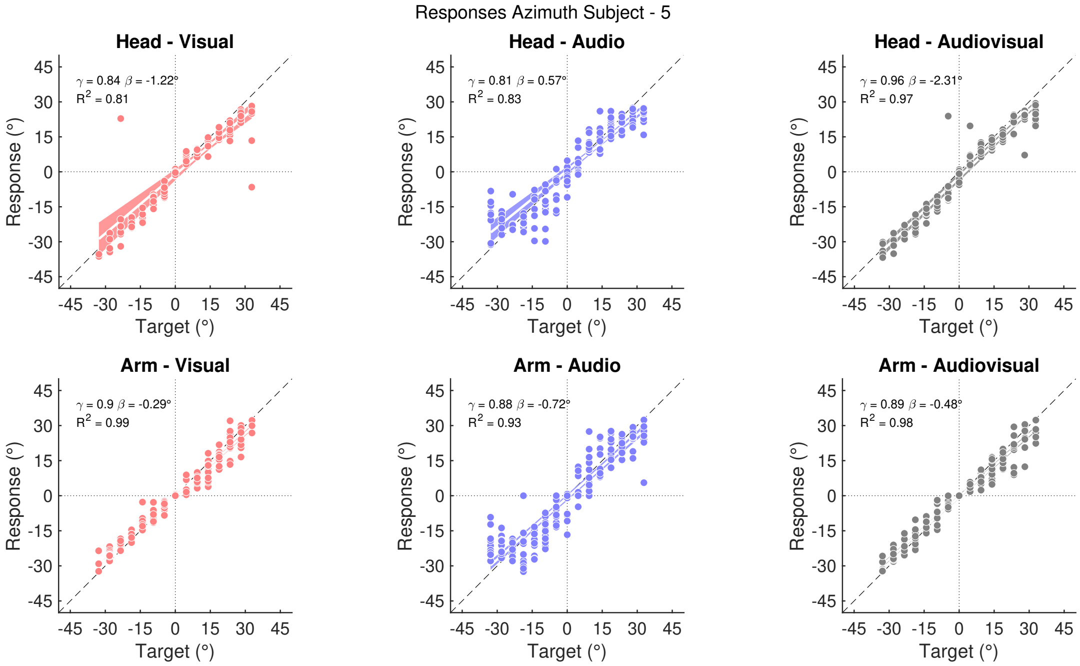


Figure B.17: Stimulus response plots of participant 1 for all modalities and effectors. The one-on-one relations between the target and response are presented by the black line while the optimal linear fit details are indicated in the top left corner. Note the differences between the modalities.

B.5.2. Reaction time compared to azimuth error

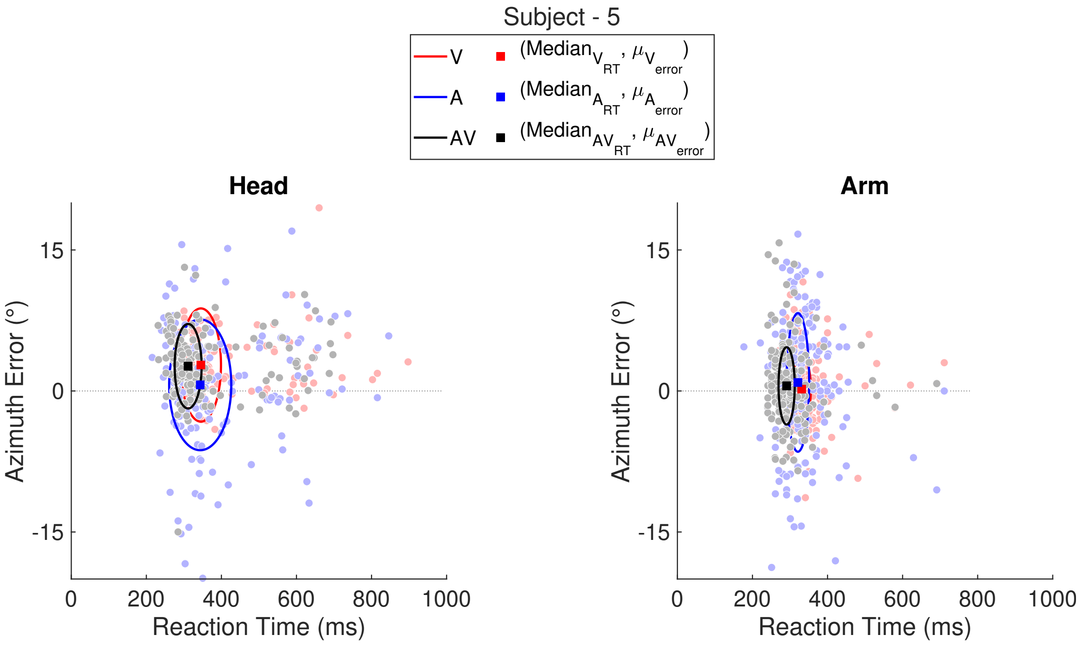


Figure B.18: Localization error as a function of reaction time for auditory (A, blue), visual (V, red), and audio-visual (AV, black) responses of participant 5. Squares indicate the median reaction time (x-axis) and mean localization error (y-axis). Ellipses depict the interquartile range (75th-25th percentile) of the reaction times (x-axis) and the standard deviation of the errors (y-axis). Note the difference in ellipse sizes for different modalities

B.5.3. Reaction time per experiment block

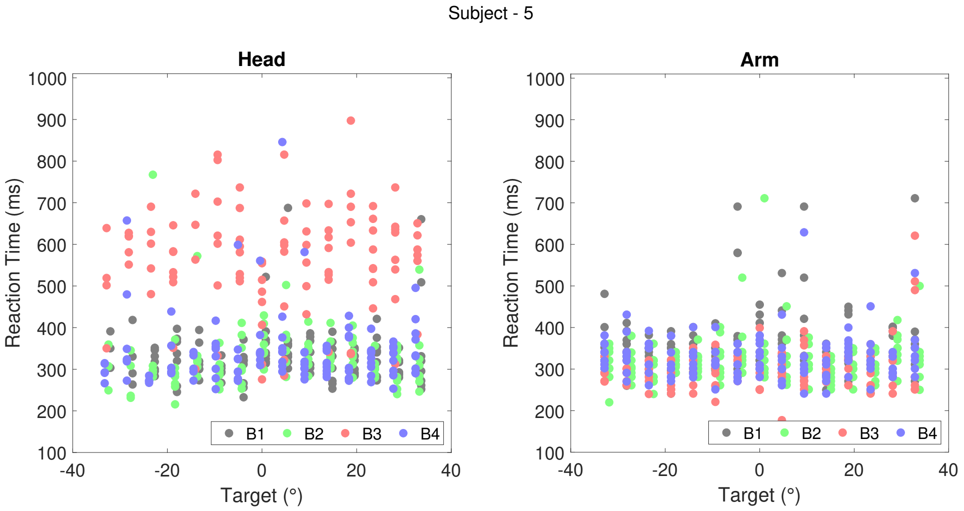


Figure B.19: Reaction times are plotted according to target locations for participant 5, with distinct colors denoting responses from various blocks (B).

B.5.4. Time discrepancies per trial

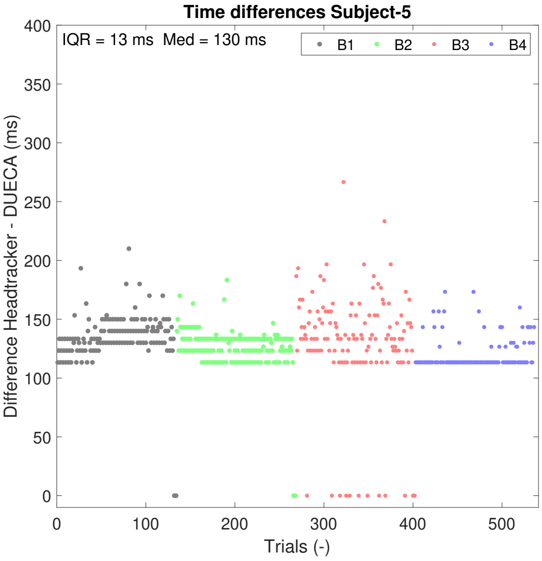


Figure B.20: Temporal discrepancies between the head tracker and DUECA software are plotted. Sequentially ordered trials are color-coded by block, with trial numbers cumulative across previous blocks.

B.6. Participant 6

B.6.1. Stimulus response plots

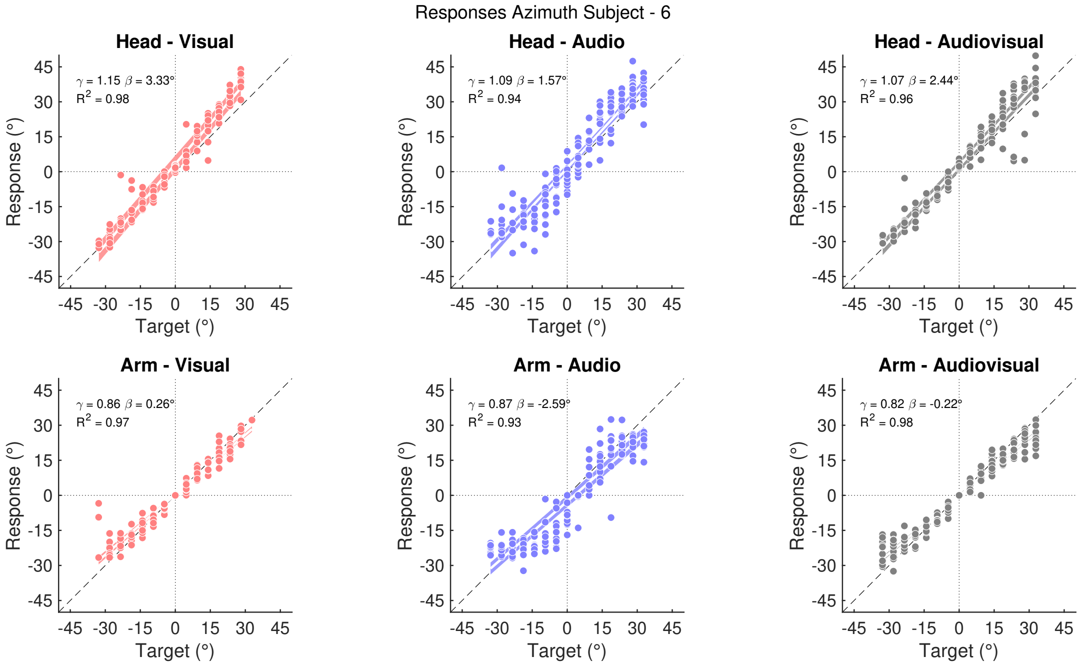


Figure B.21: Stimulus response plots of participant 6 for all modalities and effectors. The one-on-one relations between the target and response are presented by the black line while the optimal linear fit details are indicated in the top left corner. Note the differences between the modalities.

B.6.2. Reaction time compared to azimuth error

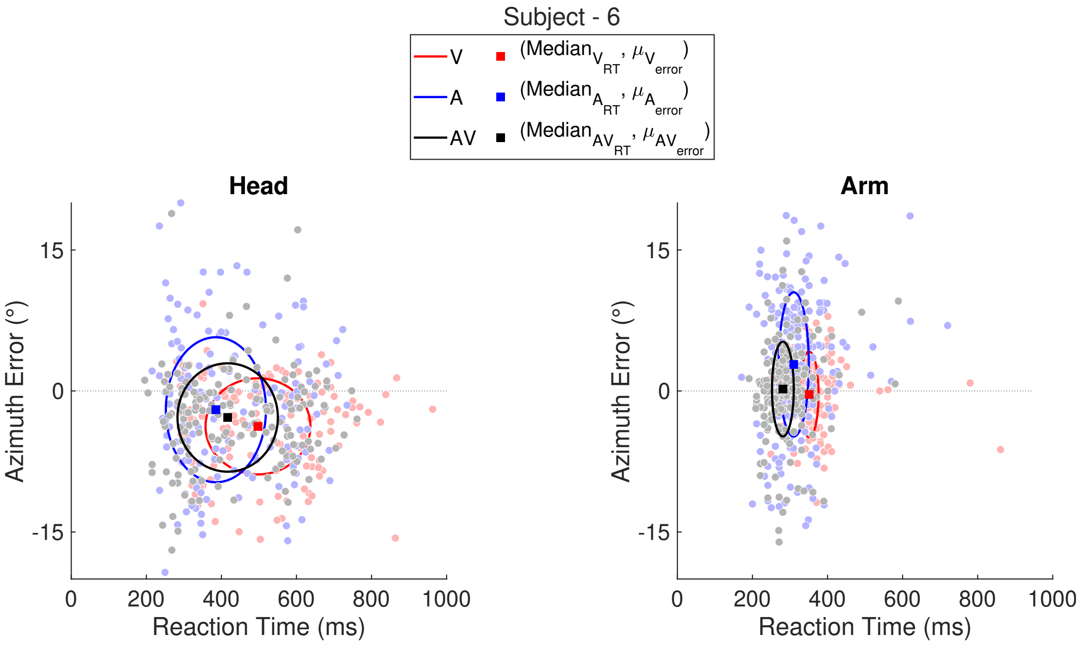


Figure B.22: Localization error as a function of reaction time for auditory (A, blue), visual (V, red), and audio-visual (AV, black) responses of participant 6. Squares indicate the median reaction time (x-axis) and mean localization error (y-axis). Ellipses depict the interquartile range (75th-25th percentile) of the reaction times (x-axis) and the standard deviation of the errors (y-axis). Note the difference in ellipse sizes for different modalities

B.6.3. Reaction time per experiment block

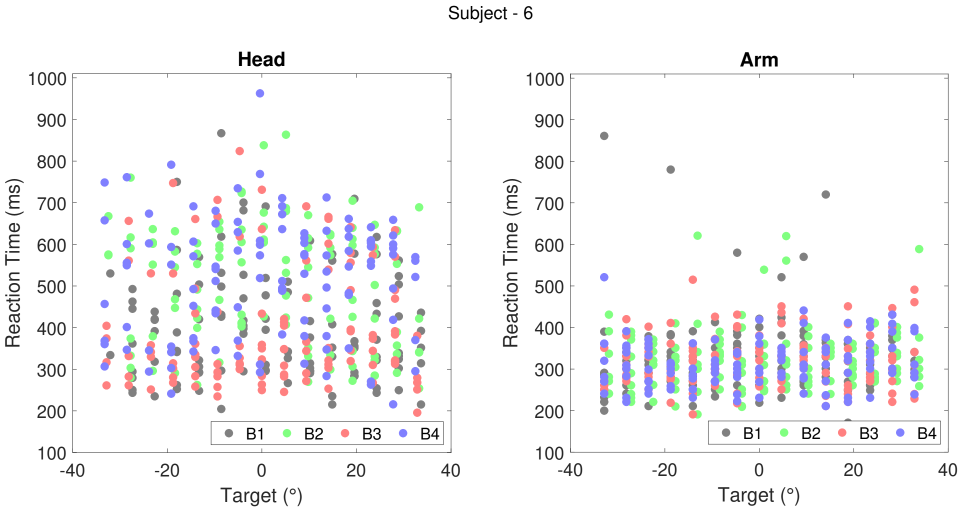


Figure B.23: Reaction times are plotted according to target locations for participant 6, with distinct colors denoting responses from various blocks (B).

B.6.4. Time discrepancies per trial

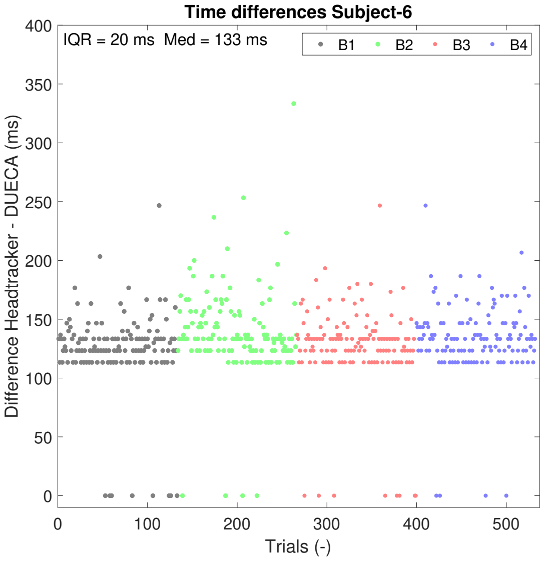


Figure B.24: Temporal discrepancies between the head tracker and DUECA software are plotted. Sequentially ordered trials are color-coded by block, with trial numbers cumulative across previous blocks.

B.7. Participant 7

B.7.1. Stimulus response plots

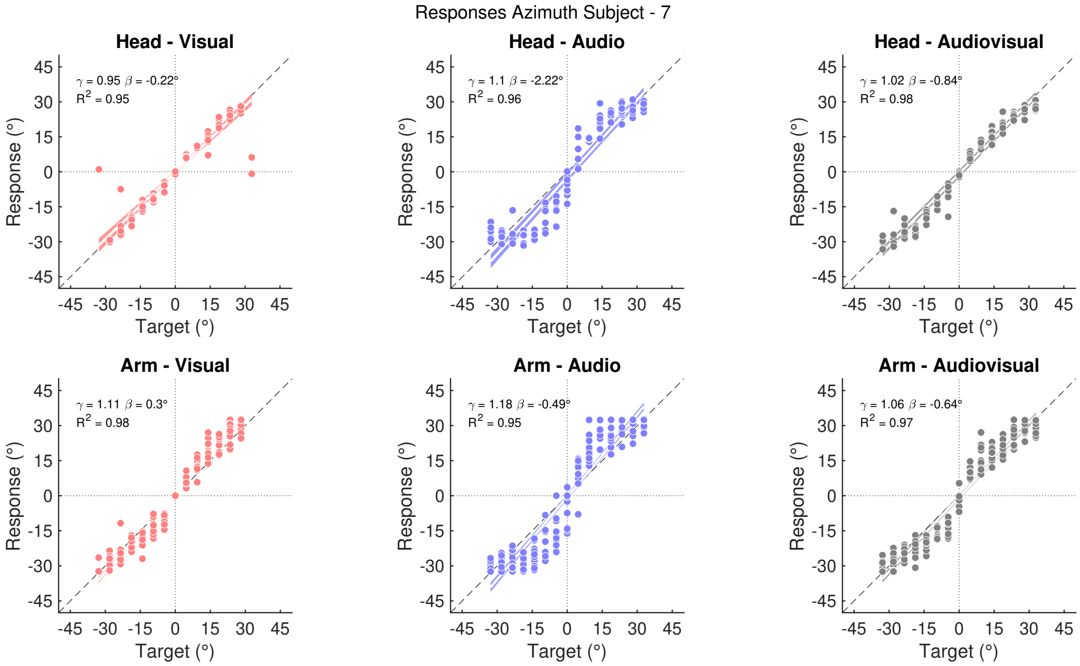


Figure B.25: Stimulus response plots of participant 7 for all modalities and effectors. The one-on-one relations between the target and response are presented by the black line while the optimal linear fit details are indicated in the top left corner. Note the differences between the modalities.

B.7.2. Reaction time compared to azimuth error

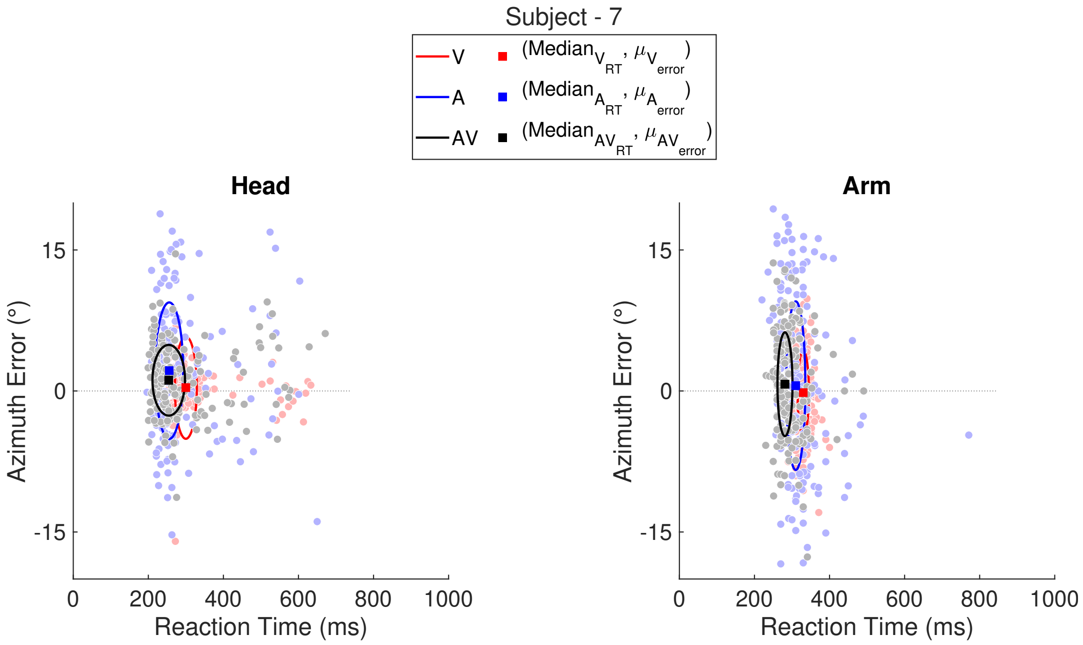


Figure B.26: Localization error as a function of reaction time for auditory (A, blue), visual (V, red), and audio-visual (AV, black) responses of participant 7. Squares indicate the median reaction time (x-axis) and mean localization error (y-axis). Ellipses depict the interquartile range (75th-25th percentile) of the reaction times (x-axis) and the standard deviation of the errors (y-axis). Note the difference in ellipse sizes for different modalities

B.7.3. Reaction time per experiment block

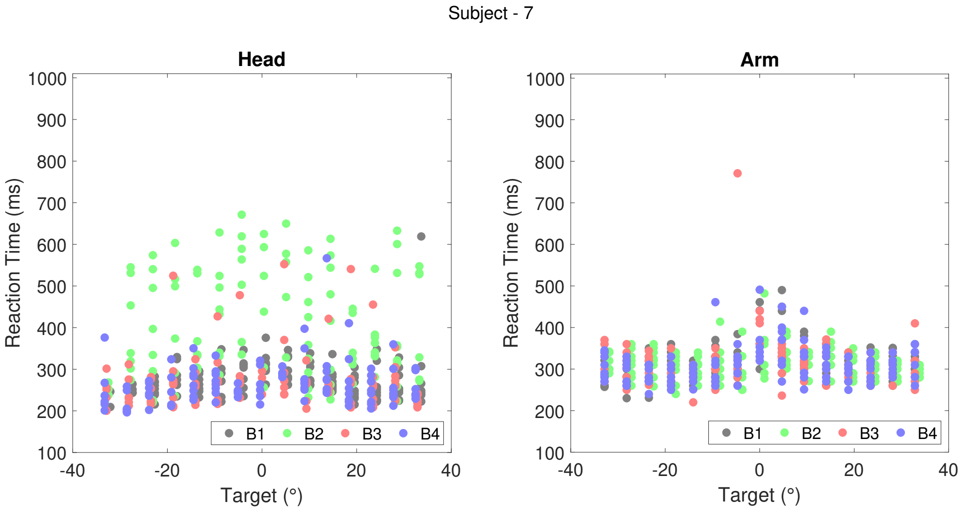


Figure B.27: Reaction times are plotted according to target locations for participant 7, with distinct colors denoting responses from various blocks (B).

B.7.4. Time discrepancies per trial

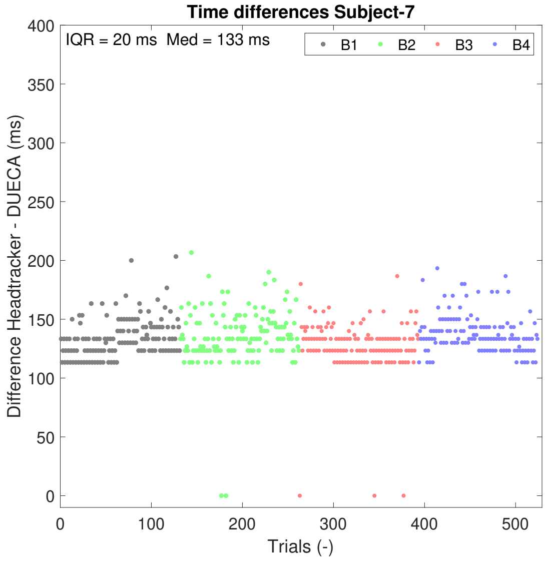


Figure B.28: Temporal discrepancies between the head tracker and DUECA software are plotted. Sequentially ordered trials are color-coded by block, with trial numbers cumulative across previous blocks.

B.8. Participant 8

B.8.1. Stimulus response plots

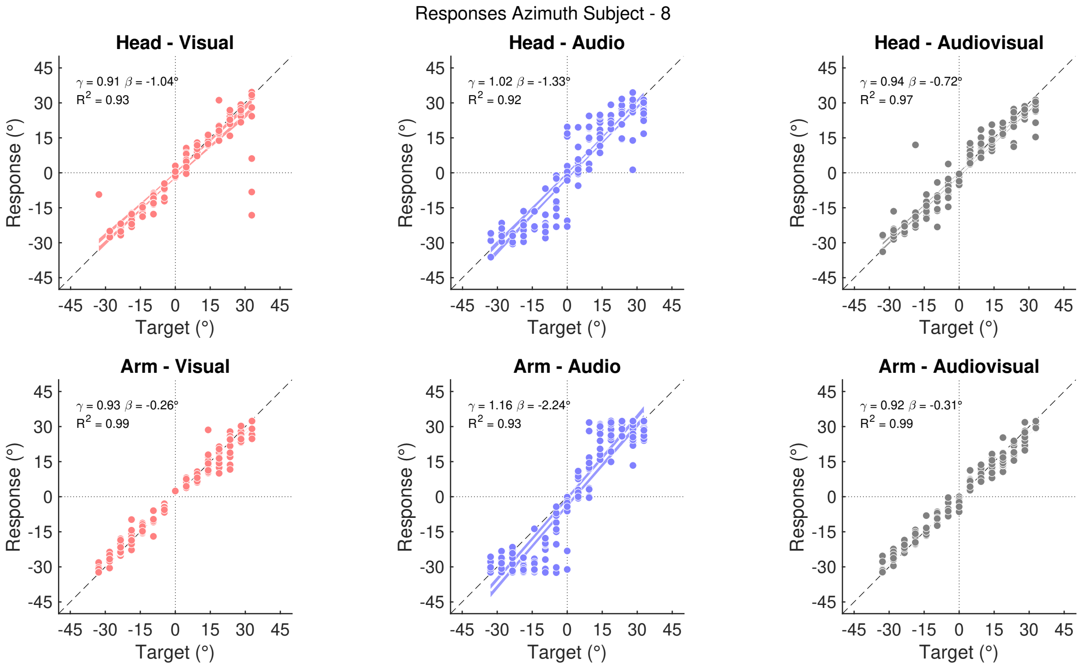


Figure B.29: Stimulus response plots of participant 8 for all modalities and effectors. The one-on-one relations between the target and response are presented by the black line while the optimal linear fit details are indicated in the top left corner. Note the differences between the modalities.

B.8.2. Reaction time compared to azimuth error

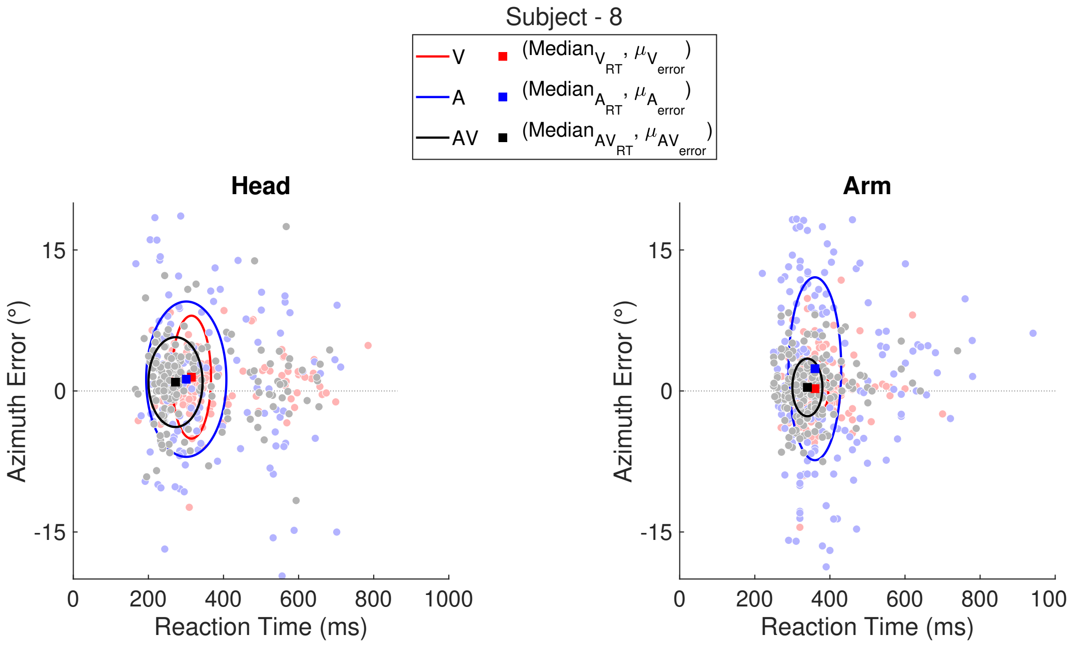


Figure B.30: Localization error as a function of reaction time for auditory (A, blue), visual (V, red), and audio-visual (AV, black) responses of participant 8. Squares indicate the median reaction time (x-axis) and mean localization error (y-axis). Ellipses depict the interquartile range (75th-25th percentile) of the reaction times (x-axis) and the standard deviation of the errors (y-axis). Note the difference in ellipse sizes for different modalities

B.8.3. Reaction time per experiment block

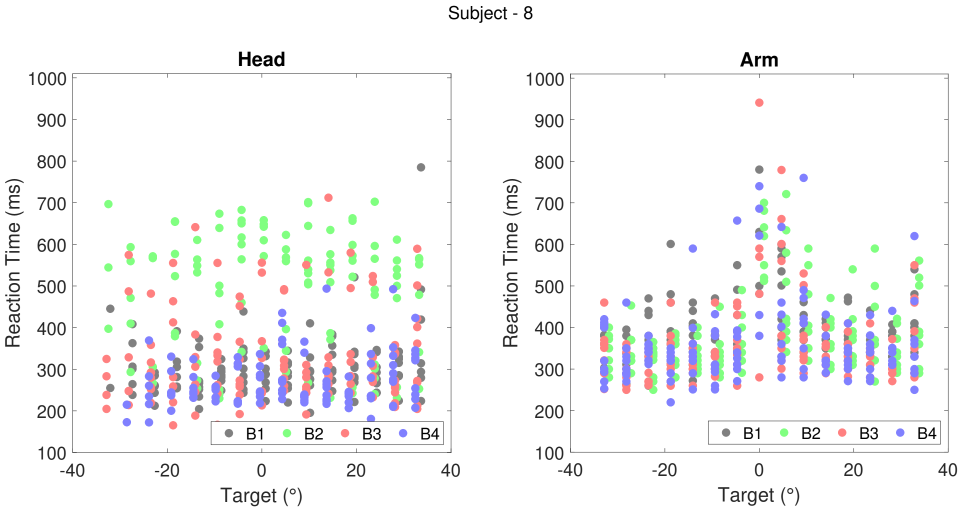


Figure B.31: Reaction times are plotted according to target locations for participant 8, with distinct colors denoting responses from various blocks (B).

B.8.4. Time discrepancies per trial

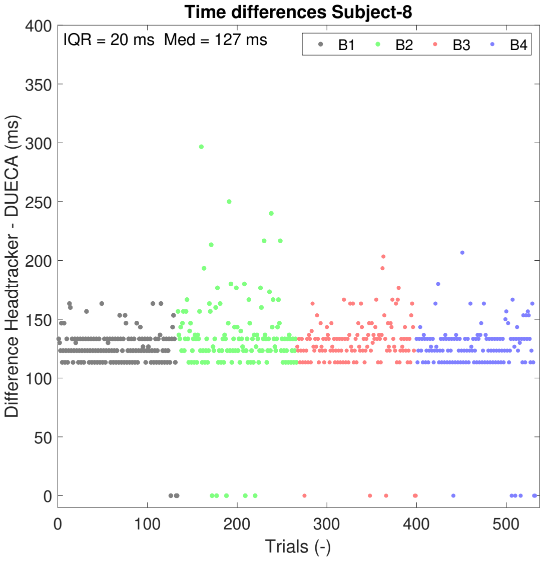


Figure B.32: Temporal discrepancies between the head tracker and DUECA software are plotted. Sequentially ordered trials are color-coded by block, with trial numbers cumulative across previous blocks.

B.9. Participant 9

B.9.1. Stimulus response plots

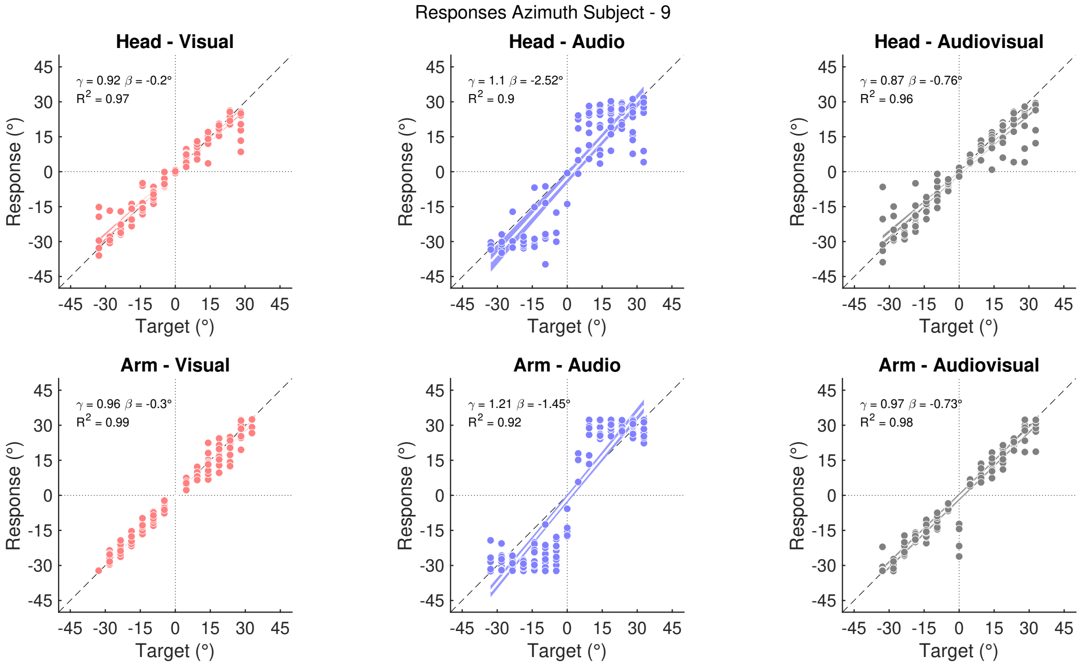


Figure B.33: Stimulus response plots of participant 9 for all modalities and effectors. The one-on-one relations between the target and response are presented by the black line while the optimal linear fit details are indicated in the top left corner. Note the differences between the modalities.

B.9.2. Reaction time compared to azimuth error

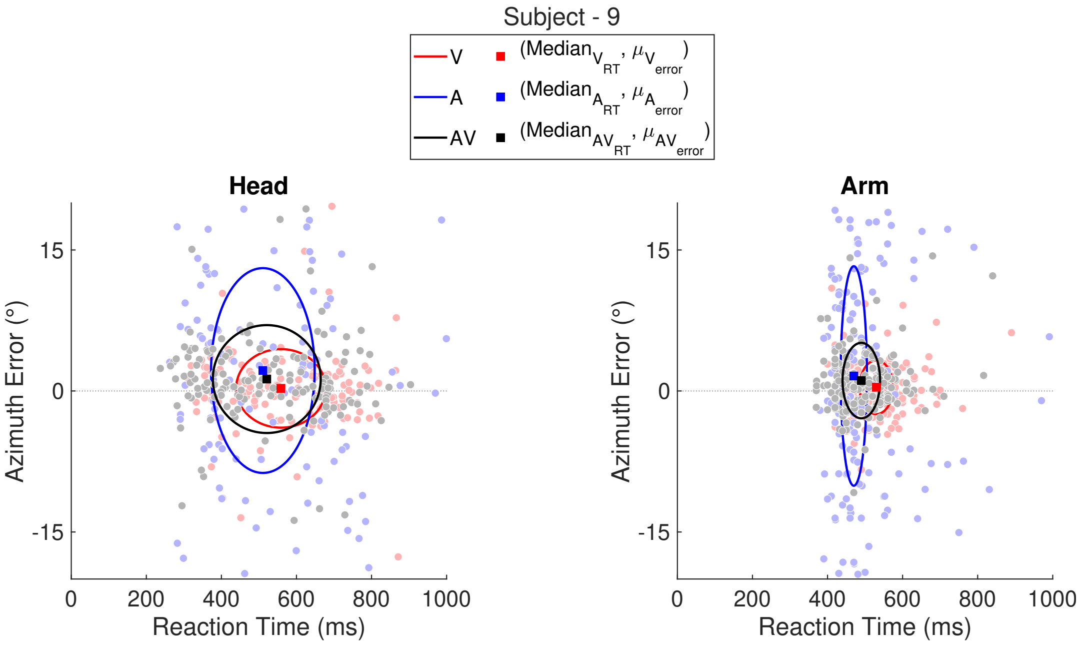


Figure B.34: Localization error as a function of reaction time for auditory (A, blue), visual (V, red), and audio-visual (AV, black) responses of participant 9. Squares indicate the median reaction time (x-axis) and mean localization error (y-axis). Ellipses depict the interquartile range (75th-25th percentile) of the reaction times (x-axis) and the standard deviation of the errors (y-axis). Note the difference in ellipse sizes for different modalities

B.9.3. Reaction time per experiment block

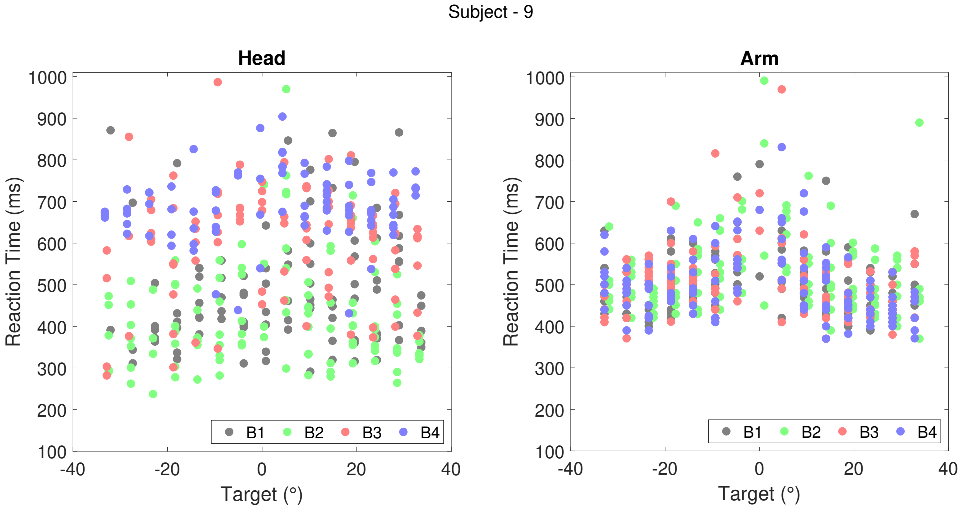


Figure B.35: Reaction times are plotted according to target locations for participant 9, with distinct colors denoting responses from various blocks (B).

B.9.4. Time discrepancies per trial

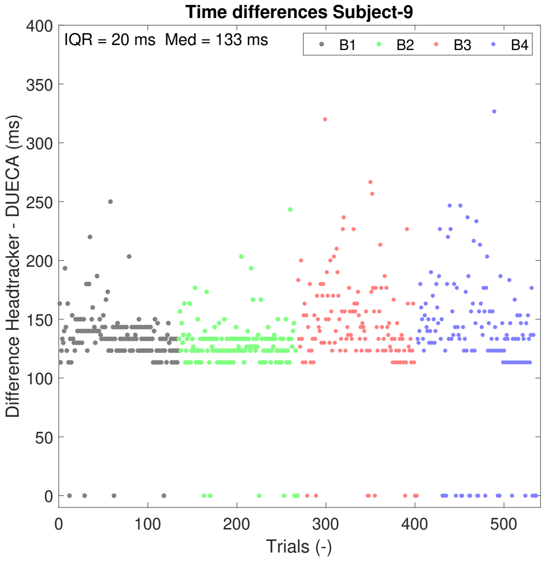


Figure B.36: Temporal discrepancies between the head tracker and DUECA software are plotted. Sequentially ordered trials are color-coded by block, with trial numbers cumulative across previous blocks.

B.10. Participant 10

B.10.1. Stimulus response plots

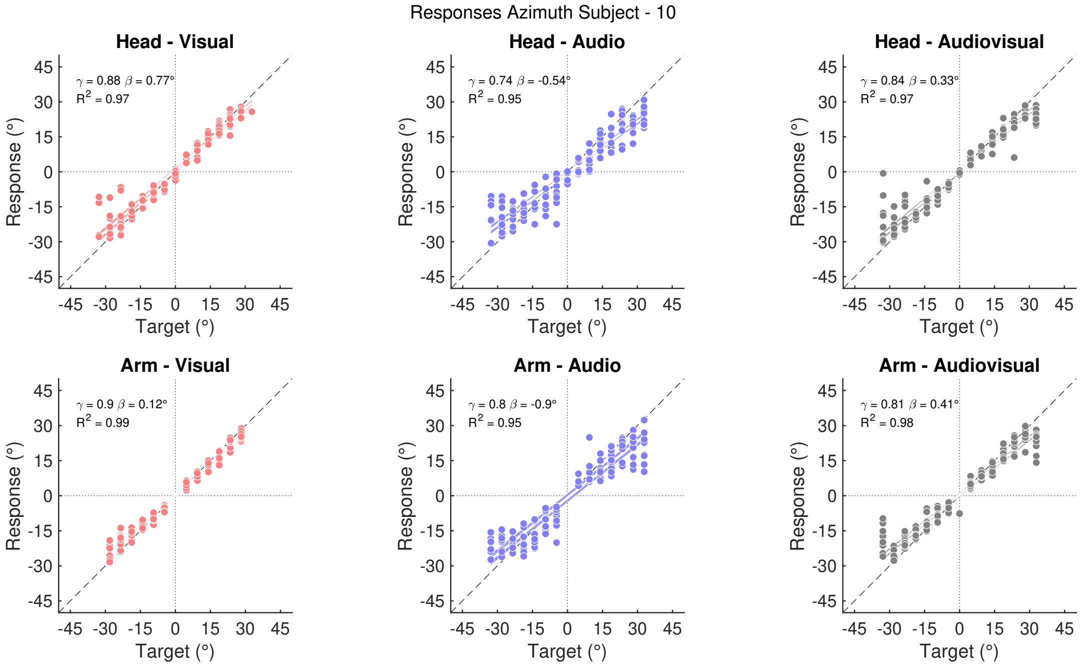


Figure B.37: Stimulus response plots of participant 10 for all modalities and effectors. The one-on-one relations between the target and response are presented by the black line while the optimal linear fit details are indicated in the top left corner. Note the differences between the modalities.

B.10.2. Reaction time compared to azimuth error

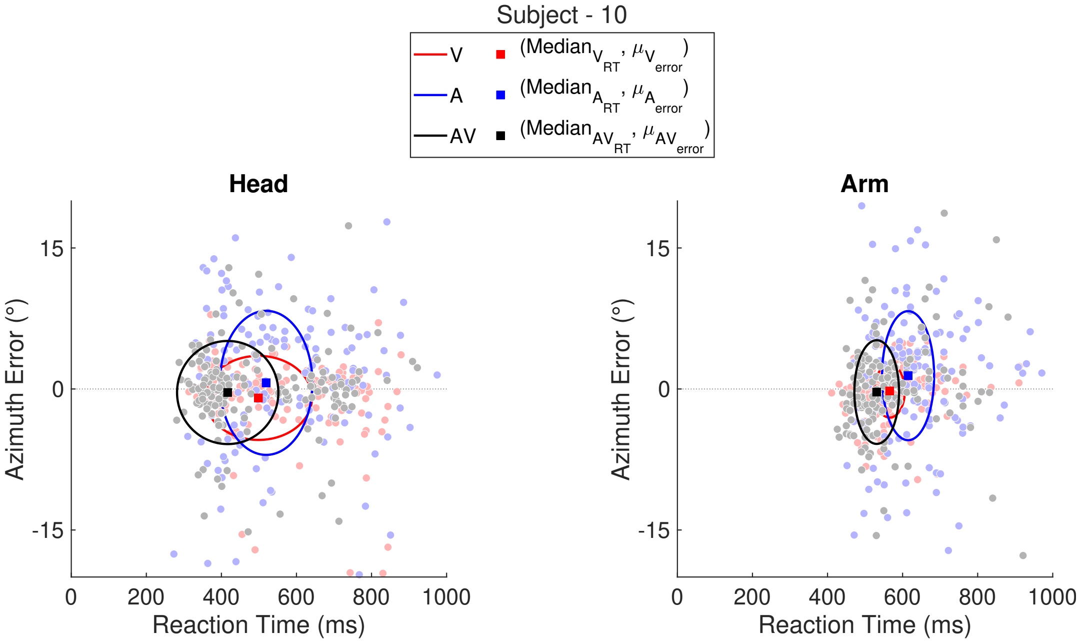


Figure B.38: Localization error as a function of reaction time for auditory (A, blue), visual (V, red), and audio-visual (AV, black) responses of participant 10. Squares indicate the median reaction time (x-axis) and mean localization error (y-axis). Ellipses depict the interquartile range (75th-25th percentile) of the reaction times (x-axis) and the standard deviation of the errors (y-axis). Note the difference in ellipse sizes for different modalities

B.10.3. Reaction time per experiment block

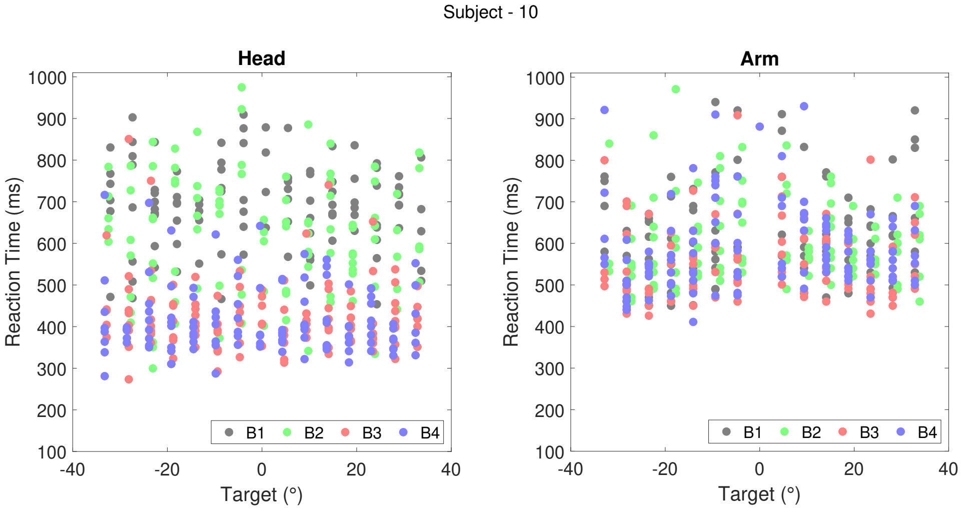


Figure B.39: Reaction times are plotted according to target locations for participant 10, with distinct colors denoting responses from various blocks (B).

B.10.4. Time discrepancies per trial

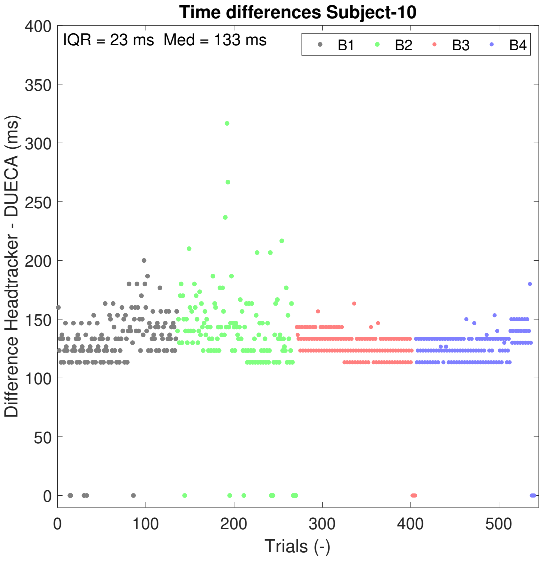


Figure B.40: Temporal discrepancies between the head tracker and DUECA software are plotted. Sequentially ordered trials are color-coded by block, with trial numbers cumulative across previous blocks.

B.11. Participant 11

B.11.1. Stimulus response plots

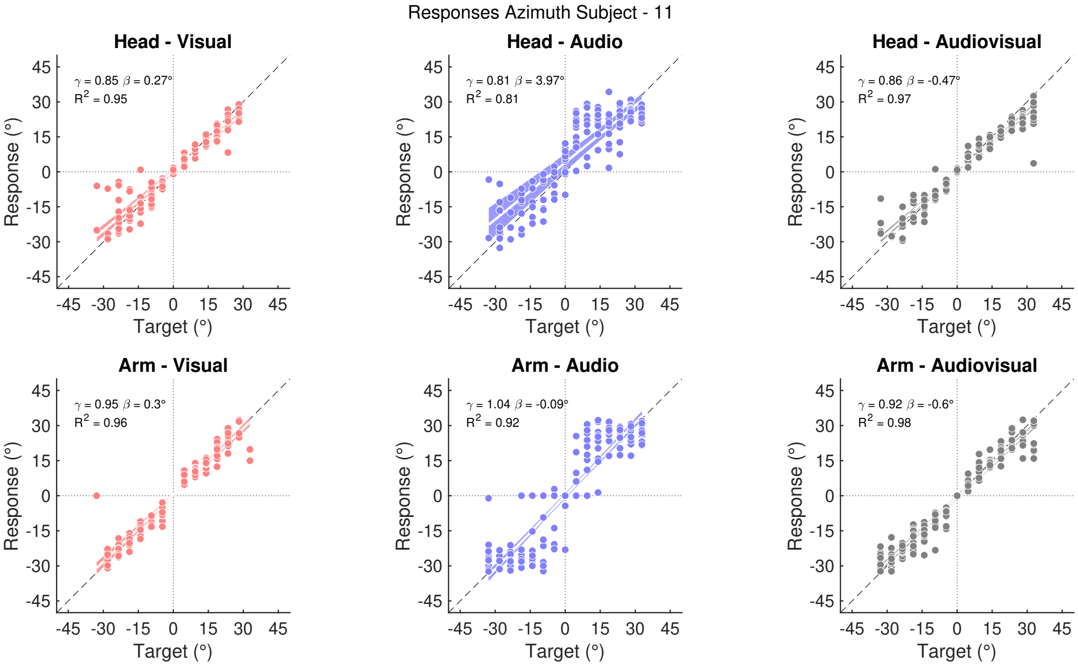


Figure B.41: Stimulus response plots of participant 11 for all modalities and effectors. The one-on-one relations between the target and response are presented by the black line while the optimal linear fit details are indicated in the top left corner. Note the differences between the modalities.

B.11.2. Reaction time compared to azimuth error

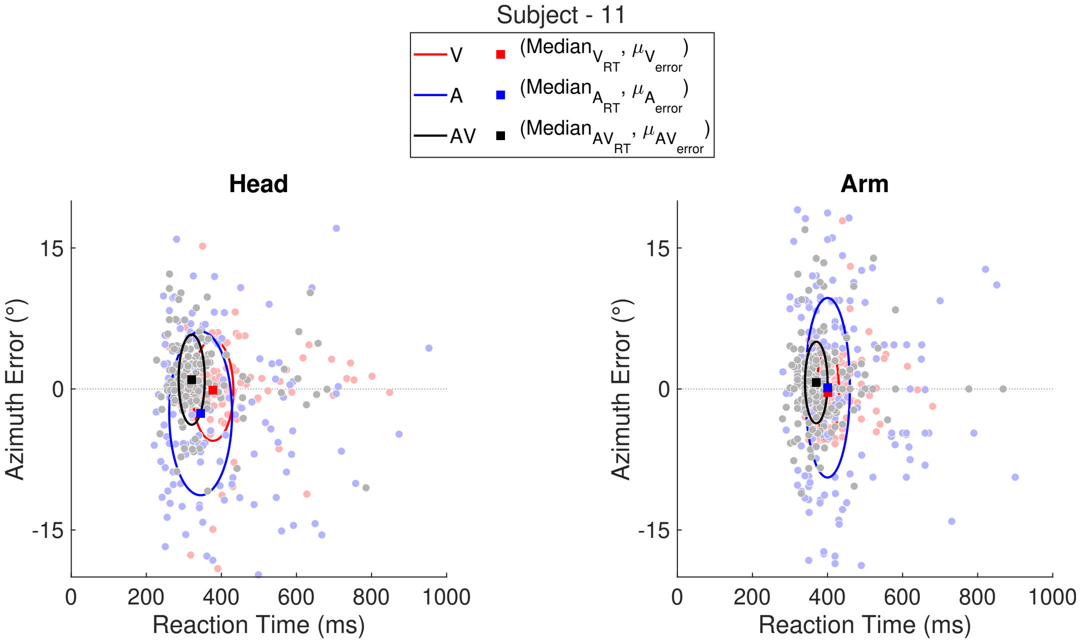


Figure B.42: Localization error as a function of reaction time for auditory (A, blue), visual (V, red), and audio-visual (AV, black) responses of participant 11. Squares indicate the median reaction time (x-axis) and mean localization error (y-axis). Ellipses depict the interquartile range (75th-25th percentile) of the reaction times (x-axis) and the standard deviation of the errors (y-axis). Note the difference in ellipse sizes for different modalities

B.11.3. Reaction time per experiment block

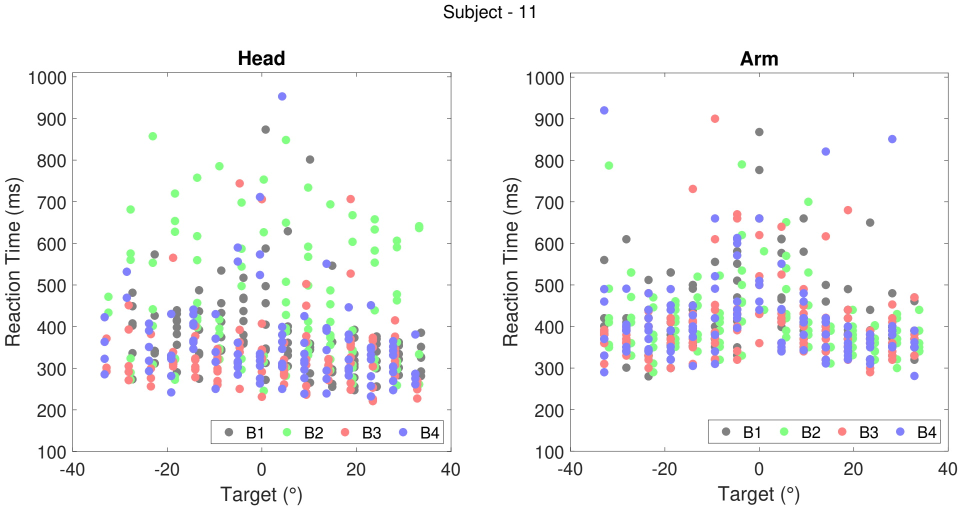


Figure B.43: Reaction times are plotted according to target locations for participant 11, with distinct colors denoting responses from various blocks (B).

B.11.4. Time discrepancies per trial

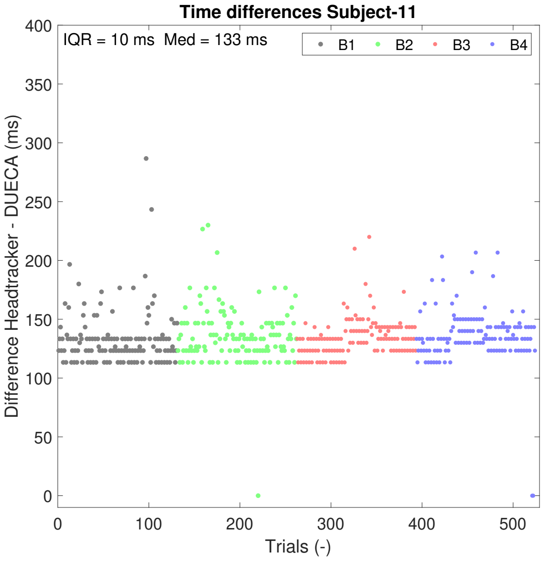


Figure B.44: Temporal discrepancies between the head tracker and DUECA software are plotted. Sequentially ordered trials are color-coded by block, with trial numbers cumulative across previous blocks.

B.12. Participant 12

B.12.1. Stimulus response plots

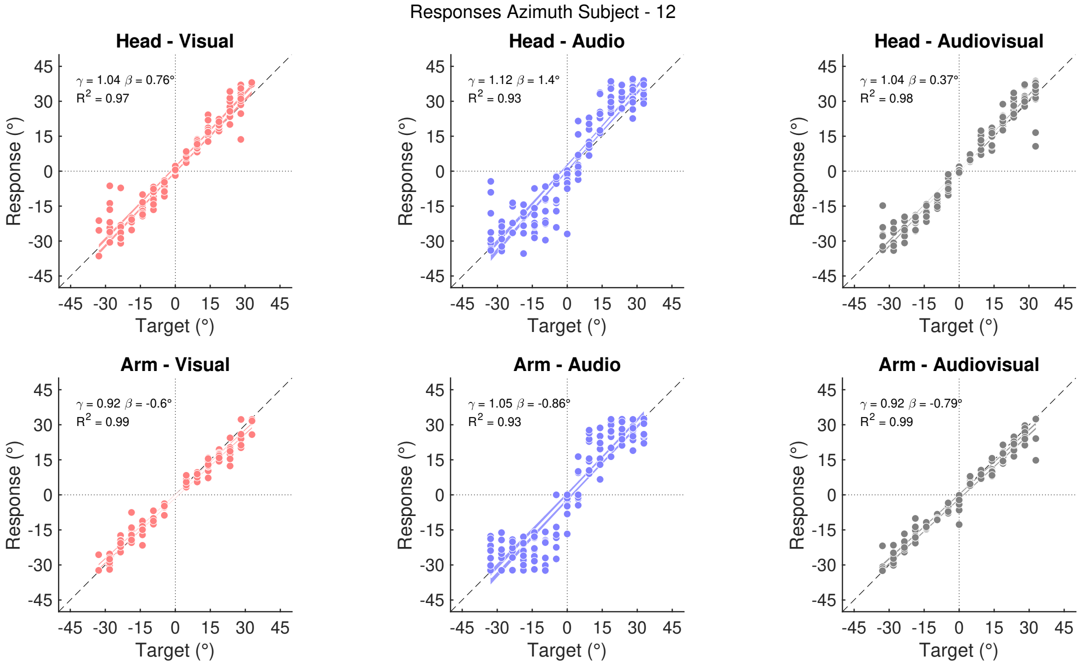


Figure B.45: Stimulus response plots of participant 12 for all modalities and effectors. The one-on-one relations between the target and response are presented by the black line while the optimal linear fit details are indicated in the top left corner. Note the differences between the modalities.

B.12.2. Reaction time compared to azimuth error

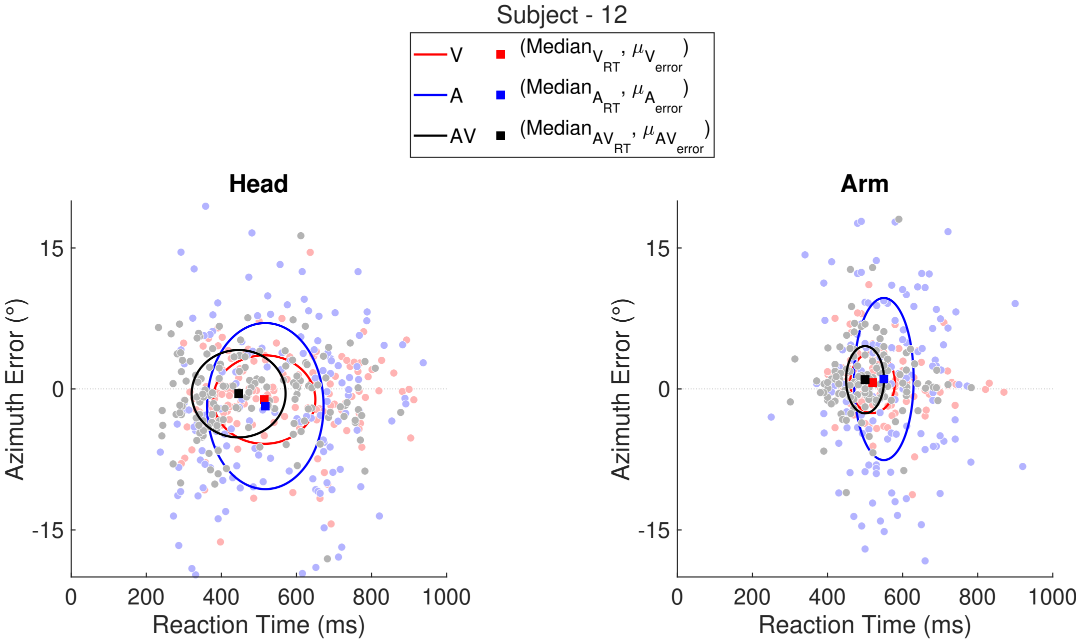


Figure B.46: Localization error as a function of reaction time for auditory (A, blue), visual (V, red), and audio-visual (AV, black) responses of participant 12. Squares indicate the median reaction time (x-axis) and mean localization error (y-axis). Ellipses depict the interquartile range (75th-25th percentile) of the reaction times (x-axis) and the standard deviation of the errors (y-axis). Note the difference in ellipse sizes for different modalities

B.12.3. Reaction time per experiment block

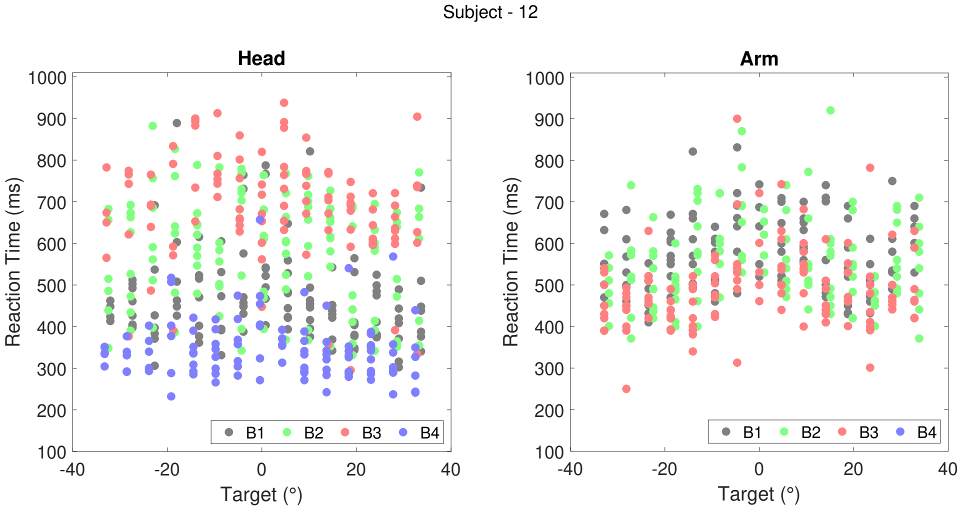


Figure B.47: Reaction times are plotted according to target locations for Participant 12, with distinct colors denoting responses from various blocks (B).

B.12.4. Time discrepancies per trial

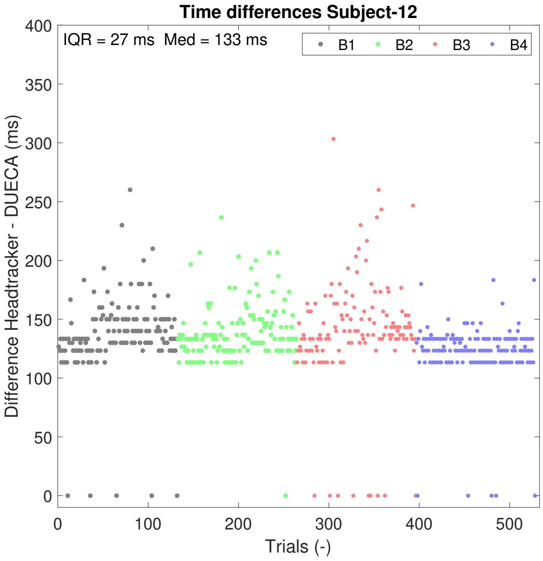


Figure B.48: Temporal discrepancies between the head tracker and DUECA software are plotted. Sequentially ordered trials are color-coded by block, with trial numbers cumulative across previous blocks.

B.13. Average Stimulus Response Plot

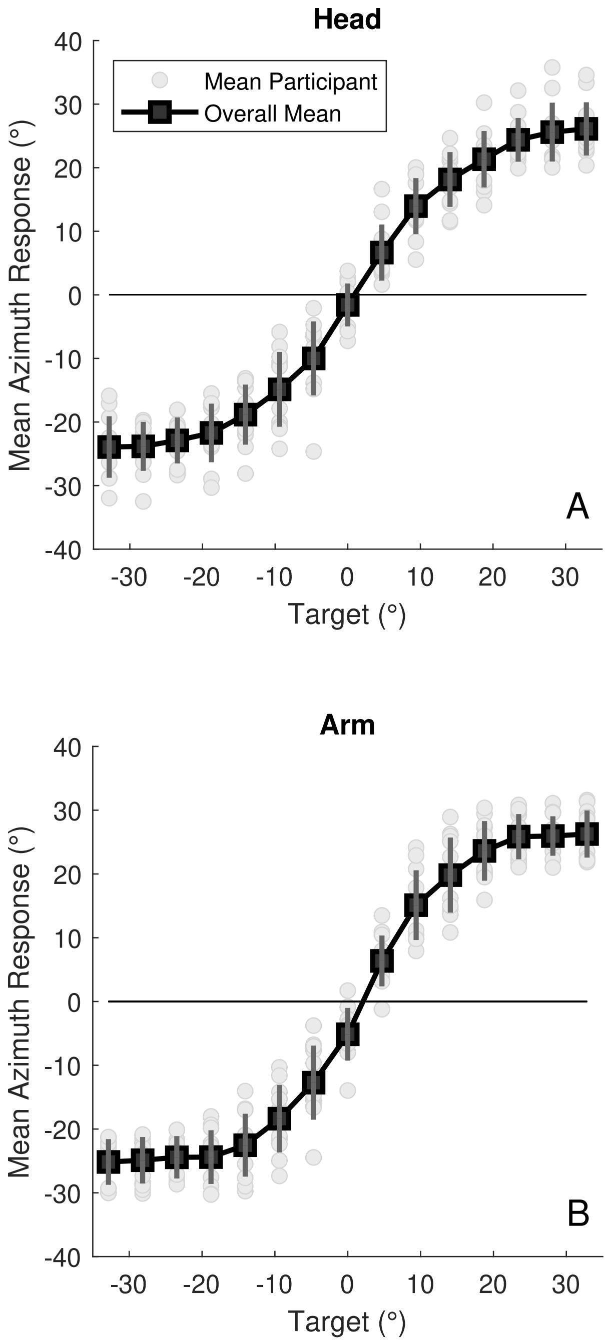


Figure B.49: Stimulus response plot of average response endpoints in azimuth for all participants.

C

Movement characterisctics

In this appendix, we asses the modality dependency of movement characteristics, defined as correction saccades and peak velocity.

C.1. Corrections saccades

We observed that participants made multiple movements before reaching a constant end position. Correction saccades were pooled over all participants and target locations and are shown in fig. C.1. Both corrections for head movements (H, left) and arm movements (A, right) are displayed. Each correction movement is shown with a small blue, red, or black dot representing corrections for audio(A), visual (V), and audiovisual (AV) trials. The large blue squares indicate the median of all correction movement per modality and effector. The median of the amount of correction made is not modality dependent, however, for arm movements more corrections are made most likely caused by the corrections for overshoot caused by the high gain of the joystick.

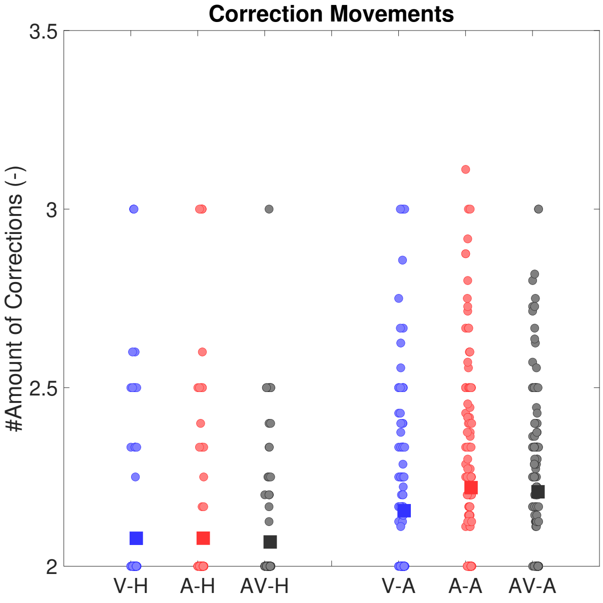


Figure C.1: Correction saccades pooled over all participants and target locations. Responses are categorized for each effector, arm (A), and head (H), and each modality, audio (A), visual (V) and audiovisual (AV)

C.2. Peak Velocity

To evaluate the modality dependency of the peak velocity in more detail we plotted the unimodal responses against the multimodal peak velocities in fig. C.2. Different unimodal responses are depicted as red and blue squares, visual and audio respectively. The solid red and blue lines indicate the best linear fit for the visual and audio responses of which the specifications are shown in the upper left corner. Both for head and arm movements, the audio and visual line have a high coefficient of determination (R^2), indicating a good fit to the line. Next, the bias (β) and gain (γ) are close to 1 and 0, respectively, indicating similarity between the peak velocity of both unimodal and multimodal responses.

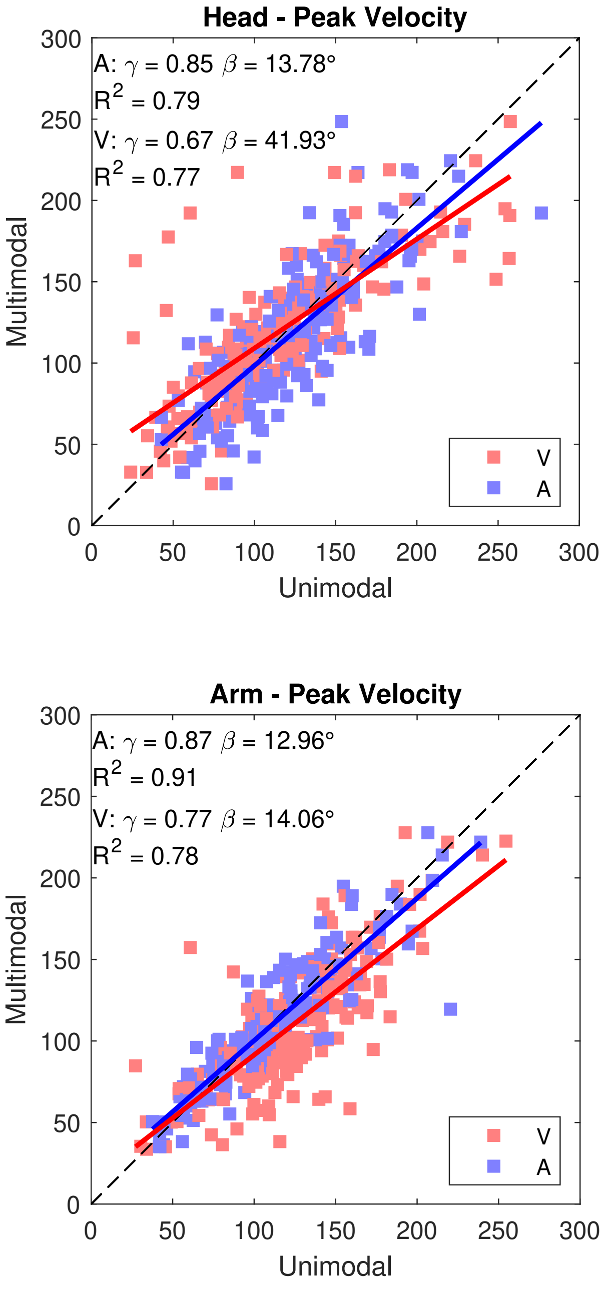


Figure C.2: Unimodal and multimodal peak velocities plotted for head (top) and arm (bottom) responses. The linear fit for visual responses (red) and audio responses (blue) are illustrated using solid lines. Details about the characteristics of the fit are provided in the upper-left corner. Notably, a resemblance is observed between the two lines.

C.3. Discussion

With regard to correction saccades, no distinct contrast was apparent between head and arm movements. Nonetheless, the cumulative count of correction saccades was notably greater for all modalities in arm movements compared to head movements. As previous research on audiovisual integration mainly focused on arm-pointing tasks, we introduced the position-controlled controller due to its close approximation to natural arm movements. However, the controller's high gain setting, coupled with the need for rapid responses, often resulted in overshooting the target locations. This aligns with the increased count of correction saccades during arm movements, a trend also reflected in the feedback provided by participants. Therefore, the augmented occurrence of correction saccades can be attributed to the characteristics of the controller's dynamics

Both for head and arm movements no difference was observed between audio, visual, and audio-visual peak velocities over all participants. Hence, peak velocity is not modality dependent.

While these two variables offer only preliminary insights, their findings hold considerable significance for subsequent investigations. Given the extensive body of research focused on manual control tasks utilizing visual and haptic cues, these outcomes propose the potential applicability of these models to auditory tracking tasks. Future studies should be conducted involving varied controlled element dynamics to evaluate other motor behavior characteristics

D

Experiment Table

Table D.1: Block order experiments for all participants

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Block	A	E	B	C	D	E	F	F	A	B	D	C
	Head	Arm	Arm	Head	Arm	Arm	Head	Head	Head	Arm	Arm	Head
	Arm	Arm	Head	Head	Arm	Arm	Head	Head	Arm	Head	Arm	Head
	Head	Arm	Arm	Arm	Head	Arm	Head	Head	Head	Arm	Head	Arm
	Arm	Arm	Head	Arm	Head	Arm	Head	Head	Arm	Head	Head	Arm
	Head	Head	Arm	Head	Arm	Head	Arm	Arm	Head	Arm	Arm	Head
	Arm	Head	Head	Head	Arm	Head	Arm	Arm	Arm	Head	Arm	Head
	Head	Head	Arm	Arm	Head	Head	Arm	Arm	Head	Arm	Head	Arm
	Arm	Head	Head	Arm	Head	Head	Arm	Arm	Arm	Head	Head	Arm

E

Participant Consent Form

This appendix includes the consent from signed by all participants prior to taking part in the experiment.

Experiment Consent Form

Measuring audiovisual benefits in a manual discrete localization task

I hereby confirm, by ticking the box, that

1. I volunteer to participate in the experiment conducted by the researcher (**Tessa Mennink**) under the supervision of **dr.ir. Daan Pool,** from the Faculty of Aerospace Engineering of TU Delft. I understand that my participation in this experiment is voluntary and that I may withdraw (“opt-out”) from the study at any time, for any reason. □

2. I have read the briefing document and understand the experiment instructions and have had all remaining questions answered to my satisfaction. □

3. I understand that taking part in the experiment involves performing manual discrete localization tasks in the HMILab simulator at TU Delft based on audio, visual, and audiovisual stimuli. I understand that only the pseudonymized recorded time traces of the tracking tasks and head tracking data are saved and used for data analysis. □

4. I confirm that the researcher has provided me with detailed safety and operational instructions for the HMILab simulator (simulator setup, electro-hydraulic side stick, emergency procedures) used in the experiment. Furthermore, I understand the researcher’s instructions for guaranteeing the experiment’s compliance with current COVID-19 guidelines, and that this experiment shall at all times follow these guidelines. □

5. I understand that the researcher will not identify me by name in any reports or publications that will result from this experiment, and that my confidentiality as a participant in this study will remain secure. Specifically, I understand that any demographic information I provide (gender, handedness, age range, ***see next page***) will only be used for reference and always presented in aggregate form in scientific publications. □

6. I understand that this research study has been reviewed and approved by the TU Delft Human Research Ethics Committee (HREC). To report any problems regarding my participation in the experiment, I know I can contact the researchers using the contact information below. □

My Signature

Date

My Printed Name

Signature of researcher

Contact information researcher:

Contact information research supervisor:

Participant Demographic Information

Measuring audiovisual benefits in a manual discrete localization task

Age: __________

Handedness:

- ○ Left handed
- ○ Right handed
- ○ Ambidextrous

Gender: __________

Participant number: __________
(filled out by the researcher)

Contact information researcher: Contact information research supervisor:

F

Briefing Document

In this appendix, the informational document for the briefing is showcased. All individuals were provided with this informative document prior to engaging in the experiments. Subsequently, both the participant and the researcher collectively reviewed the instructions once more to ensure an accurate understanding of all provided information.

Experiment Briefing

Audiovisual benefits in a manual discrete localization task

Thank you for your participation. This experiment is part of an MSc thesis research project that aims to measure audiovisual benefits in a manual localization task. The experiment is performed in the Human-Machine Interaction Laboratory (HMILab) at TU Delft's Faculty of Aerospace Engineering. This briefing will give an overview of the experiment and explains what is expected from the participants. Please read this document carefully. Should any questions or comments remain, always feel free to discuss these with the researcher conducting the experiment.

Experiment Objective

Combined audiovisual stimuli could solve many challenges in human-machine interactions. This experiment is meant to collect responses to audio, visual, and audiovisual stimuli in an arm-orienting and a head-orienting task.

Experiment Set-up

The "HMILab" (Fig. 1), a fixed-base simulator set-up at TU Delft's Faculty of Aerospace Engineering, is used to investigate the interaction between human operators and controlled elements. You are asked to take place in the middle chair, where you can control the joystick with your right hand. Throughout the experiment, your joy stick output and head movements will be recorded with a head tracker (Moving dot see Fig. 2) to objectively detect reaction times and endpoint accuracy. Next to that, a laser projection will be used to fix the eye movements. These glasses are shown in Figure 2.

During the experiment short audio, visual and audiovisual stimuli will be presented. The visual stimuli will consist of a small grey dot projected on the big screen in front. The audio stimuli presented are directional high-pass frequency signals (4 kHz -20 kHz).

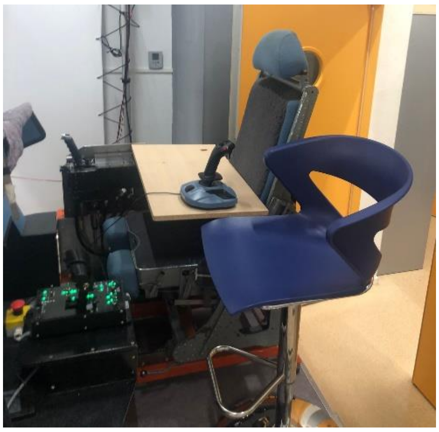


Figure 1: Illustration of HMI Lab. The participant will be sitting in the seat in the middle (blue) and controls the side stick. The participant will have to look at the screen in front.

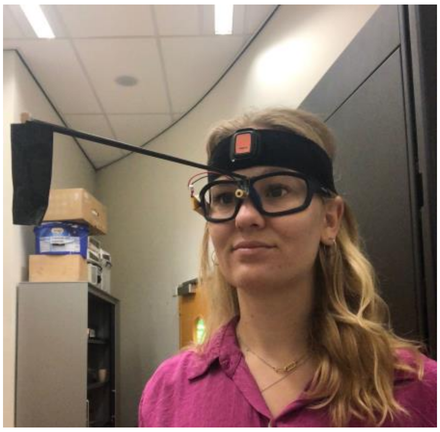


Figure 2:Moving dot head tracker and glass-mounted laser pointer

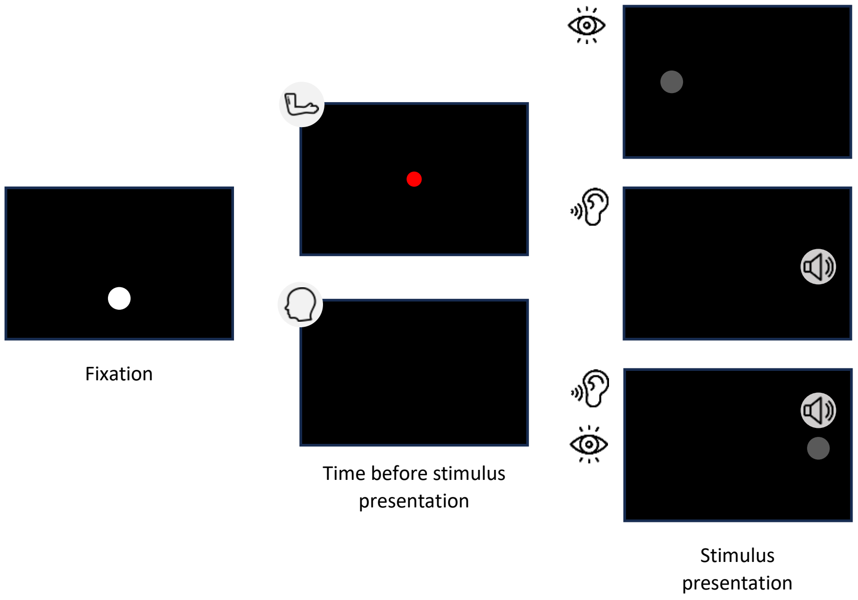


Figure 3: Example of trial lay out. Head and arm trials are separated whilst stimuli conditions are mixed among trials.

Experiment procedure

During the experiment, you are asked to determine the location of the audio, visual, or audiovisual stimuli presented. Throughout the experiments, you need to keep your eyes focused on the laser point project in front of you. The localization is performed either by head movements or arm movements, the trial type will be communicated before the start of each experiment. During the arm trials, you can move your head freely and you do not need to wear the glasses. During the head trials, you need to align the projected laser with the perceived location.

You need to perform this localization task as fast and accurately as possible. No updates on performance are given throughout the experiment. You will be asked to complete a number of repeated trials and the experimenter will notify you when sufficient data has been collected. Periodically, you will be asked to take a short (i.e., 20-minute) break to avoid fatigue. Should more breaks be required, you can request them at any moment. Prior to data collection, you will perform some practice trials to get familiar with the tasks. Moreover, a short period of time will be needed for the calibration of the head tracker. Conducting the full experiment takes approximately 2-3 hours.

Your Rights & Consent

Experiment participation is voluntary. Should you feel uncomfortable, you can decide to stop your participation at any time. By participating in the experiment you agree that the collected data may be published. Your personal data will remain confidential and anonymous, only the researcher can link the collected data to a specific participant. To ensure you understand and comply with the conditions of the experiment, you will be asked to sign an informed consent form.

Contact information researcher:	Contact information research supervisor

Thank you again for participating!