2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) | 979-8-3315-1647-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISMAR62088.2024.00022

First or Third-Person Hearing? A Controlled Evaluation of Auditory Perspective on Embodiment and Sound Localization Performance

Yi Fei Cheng*
Human-Computer Interaction Institute, Carnegie Mellon University

Laurie M. Heller†
Department of Psychology, Carnegie Mellon University

Stacey Cho‡
Human-Computer Interaction Institute, Carnegie Mellon University

David Lindlbauer§
Human-Computer Interaction Institute, Carnegie Mellon University

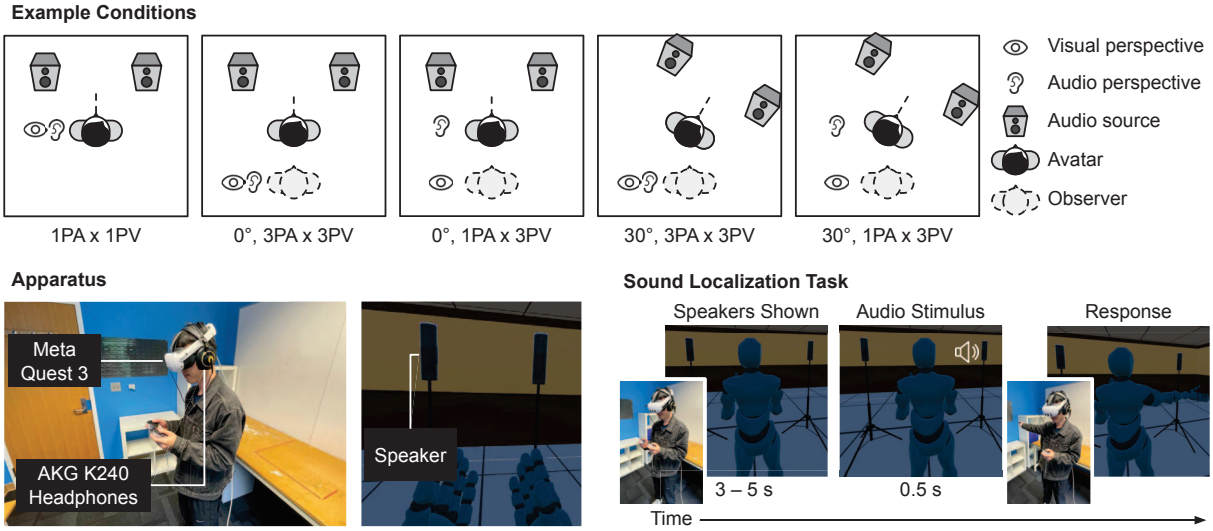

Figure 1: We present the results of a 24-participant empirical study investigating the effects of auditory perspective on sound localization performance and subjective experience. Participants performed a sound localization task in three conditions: while they controlled a virtual avatar from a first-person perspective (*1PA x 1PV*), from a third-person perspective (*3PA x 3PV*), and while they heard from the avatar's first-person auditory perspective but controlled it from a third-person visual perspective (*1PA x 3PV*). We additionally studied how the effects of perspective are mediated by a rotational misalignment introduced between the auditory and visual perspectives ($0°$ *versus* $30°$). The experiment stimuli were presented with a pair of AKG K240 headphones and a Meta Quest 3 VR headset. In the sound localization task, participants were first visually exposed to two speaker locations. One of the speakers then emitted a burst of white noise. Participants indicated which omnidirectional, cylindrical speaker they perceived as the source of the stimulus by extending their left arm for the leftmost speaker or their right arm for the rightmost speaker.

Abstract

Virtual Reality (VR) allows users to flexibly choose the perspective through which they interact with a synthetic environment. Users can either adopt a first-person perspective, in which they see through the eyes of their virtual avatar, or a third-person perspective, in which their viewpoint is detached from the virtual avatar. Prior research has shown that the visual perspective affects different interactions and influences core experiential factors, such as the user's sense of embodiment. However, there is limited understanding of how auditory perspective mediates user experience in immersive virtual environments. In this paper, we conducted a controlled experiment ($N = 24$) on the effect of the user's auditory perspective on their performance in a sound localization task and their sense of embodiment. Our results showed that when viewing a virtual avatar from a third-person visual perspective, adopting the auditory perspective of the avatar may increase agency and self-avatar merging, even when controlling for variations in task difficulty caused by shifts in auditory perspective. Additionally, our findings suggest that differences in auditory perspective generally have a smaller effect than differences in visual perspective. We discuss the implications of our empirical investigation of audio perspective for designing embodied auditory experiences in VR.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Empirical Studies in HCI

1 Introduction

Virtual Reality (VR) enables experiences that transcend the "ordinary rules of physical reality" [66]. Users can interact with immersive environments as someone with a different identity [23] or body morphology [10]. Moreover, VR allows individuals to adopt physically implausible viewpoints. Instead of a **first-person perspective (1PP)**, from which we operate in the real world, people can alternatively adopt a **third-person perspective (3PP)**, where they control a virtual avatar of themselves from an external vantage point.

Over the years, researchers have shown that these altered embodied states can offer a pathway to expand our understanding of human cognition and perception (e.g., [8]), improve empathy levels [61], and manage mental health conditions [16]. Interactions

*e-mail: yifeic2@andrew.cmu.edu
†e-mail: lauriehelle r@cmu.edu
‡email: staceycho@gse.harvard.edu
§e-mail: davidlindlbauer@cmu.edu

979-8-3315-1647-5/24/$31.00 ©2024 IEEE
DOI 10.1109/ISMAR62088.2024.00022

through different perspectives, in particular, have been a focal point of research. Most current VR applications are designed for 1PP, since this point of view generally enhances presence and embodiment [13]. In contrast, 3PP enables a wide range of interesting experiences. 3PP usage has been shown to improve situational awareness of the environment [36], while maintaining high levels of presence [18] and reducing motion sickness [46]. Additionally, 3PP can be beneficial for training scenarios [59] and help guide motor control [11]. Thus, understanding how different perspectives affect our experience of VR offers valuable insights for future applications.

Numerous works have explored the differences between engaging with VR using 1PP or 3PP. For instance, prior studies have found that 1PP leads to higher body ownership [63] and immersion [13] than 3PP. Most of these investigations, however, focus on visual or visuo-tactile cues (e.g., [14]), neglecting the auditory sense. Audio, however, serves an important role in VR experiences, similarly mediating core experiential factors in virtual environments such as one's sense of presence [31]. This raises a question on how different **auditory perspectives** affect people's experience of VR. Consider the scenario of controlling a virtual avatar from a third-person visual perspective. Based on embodiment research literature, as with visual perspective, audio perspective should align with the avatar to encourage embodiment [21]. However, this also introduces conflicting reference frames between the third-person observer's visual perspective and the first-person auditory perspective that is aligned with the avatar. This may be detrimental to spatial cognition, especially if the avatar's and the observer's reference frames are *rotationally misaligned* [37], which can frequently occur in applications like gaming [13] and teleoperation [49].

In this paper, we designed and conducted a controlled user study to investigate the following research questions:

RQ1 How does the user's audio perspective in VR influence their ability to localize sounds and their sense of embodiment?

RQ2 How does the rotational misalignment between the auditory and visual perspective reference frames mediate these effects?

RQ3 What is the relationship between task performance and perceived embodiment?

In our study, 24 participants performed a sound localization task in which they associated auditory stimuli with potential sources while embodying a virtual avatar from different auditory and visual perspectives (Figure 1). We assessed participants' sense of embodiment using questionnaires and evaluated sound localization performance with quantitative metrics such as response time and errors. In a subset of the conditions, we introduced a rotational misalignment between audio and visual perspectives.

Our results show that audio perspective influences perceived embodiment in scenarios where there is a rotational misalignment between the audio and visual perspectives. Specifically, when avatars are viewed and controlled from a third-person perspective, aligning the audio with the avatar is beneficial for users' sense of agency and embodiment. Notably, this is not purely due to variations in task difficulty introduced by the spatial arrangement differences between the perspective conditions. Finally, our results show that the effect of audio perspective on embodiment is less pronounced than that of visual perspective, which aligns with prior research [63].

Taken together, we contribute an empirical study on the differences between first- and third-person audio and visual perspective for localizing sound in VR. Our results offer guidance for the choice of auditory perspective in VR applications.

2 Background and Related Work

Our study builds upon prior research on embodiment of virtual avatars, interactions in both first- and third-person perspectives, how people localize sounds, and spatial audio usage in VR.

2.1 Embodiment of virtual avatars

The sense of "embodiment" toward a virtual avatar represents a core experiential aspect of VR, referring to the feeling of being located inside the avatar (self-location), in control of it (agency), and having ownership over it (body ownership) [33]. Embodiment is associated with enhanced presence [20] and immersion [62], and has been shown to reduce implicit biases [51] and foster behavioral change [26]. Many factors contribute to how much one feels embodied in a virtual avatar, including the avatar appearance [39], the synchronicity of movements between the avatar and the user [50], and, most relevant to our work, the perspective from which the body is viewed [12]. Our work seeks to further understand the role of perspective, particularly auditory, in virtual embodiment.

2.2 First- versus third-person perspective

In immersive environments, first-person perspective (1PP) refers to interacting with the virtual scene from the avatar's viewpoint, while third-person perspective (3PP) refers to control from a point of view outside the avatar [63]. According to Hoppe [28], VR experiences are predominantly designed for 1PP engagement, since this perspective promotes immersion, presence, embodiment, and identification with the virtual character. Indeed, these benefits have been substantiated by a variety of studies (e.g., [63, 13, 21, 12]). Denisova and Cairns [13], for instance, demonstrated that people felt more immersed in a game with 1PP, regardless of their perspective preference. Similarly, an experiment by Fribourg et al. [17] showed that it serves as a more important contributor to embodiment than avatar appearance.

On the other hand, according to Gorisse et al. [21], 3PP may enhance spatial awareness by offering a wider field of view, enabling the user to monitor peripheral elements around the avatar. Salamin et al. [59] suggest that 3PP is preferred for displacement actions, interactions with moving objects, and assisting users in evaluating distances. One caveat is that the 3PP introduces multiple, potentially misaligned spatial reference frames, whose implications should be carefully considered for tasks involving spatial cognition [37]. There is also extensive literature (cf., [34]) describing methodologies for potentially inducing comparative levels of embodiment in these experiences. For example, the canonical rubber-hand illusion demonstrates that, through synchronized visuo-tactile stimuli, subjects can be induced to perceive a fake rubber hand as if it were their own [9]. Several experiments (e.g., [14]) further demonstrate that illusions of embodiment can be induced for full humanoid bodies that are not collocated with oneself.

However, prior comparisons of 1PP and 3PP are predominantly oriented around visual (e.g., [13]) or visuo-tactile (e.g., [14]) stimuli. In contrast, our work examines how auditory perspective affects embodiment and performance in a sound localization task.

2.3 Sound localization

The human auditory system relies on a rich set of perceptual processes for localizing sounds [38]. To localize sounds in the horizontal plane, for instance, people primarily rely on two binaural cues: interaural time difference (ITD) and interaural level difference (ILD) [47]. ITD and ILD refer to the difference between the times and pressure level of a sound signal reaching the two ears, respectively. The human auditory system translates these differences into an approximation of the sound's azimuthal angle. Aside from ITD and ILD differences, sound localization is also affected by the *head-related transfer function (HRTF)* [68]. HRTF refers to how sound waves are filtered through interactions with the anatomical features of the listener (e.g., the shape of their head and ears) before perception. The HRTF further provides important cues for determining the elevation of a sound relative to the listener [68].

There is a significant body of literature characterizing human sound localization capabilities. Paradigms often involve sequen-

tially presenting sounds from spatially distinct positions, such as from different speakers in a loudspeaker array inside an anechoic chamber, and asking participants to report the position of the stimulus (e.g., [2]). Research suggests that people are generally most accurate in determining the azimuthal direction of a sound and less accurate at distinguishing elevation and distance [68]. Past experiments have shown that human localization accuracy is further mediated by a variety of factors, including the visual availability of the source [2, 29], prior experience and training [65], whether head motion was allowed [32, 53], and method of reporting [42].

First, we draw inspiration from previous research in our experimental design and variables (e.g., task [54], focus on azimuthal localization [69]). Second, our experiment uniquely studies conditions in which the listener's audio perspective is separated from their visual perspective, a scenario which is nearly impossible in physical reality without an elaborate apparatus, yet trivial to set up and more commonly adopted in virtual environments.

2.4 Spatial audio in VR

In virtual environments, the spatial experience of audio is synthesized by simulating the binaural cues of ITDs, ILDs, and HRTFs. This typically involves modeling the physics of the virtual environments and the listener's HRTF, and computing the expected stimulus to deliver to each of the listener's ears [73, 43]. Since HRTFs depend on an individual's anatomical structures and are thus difficult to personalize [45], most current audio spatializers in commercial digital devices rely on a generic HRTF model [19]. While this approach may be less accurate, Berger et al. [7] suggest that since people's representation of acoustic space is highly plastic, a generic HRTF may suffice for auditory spatial localization in VR. The audio localization study we conducted uses the HRTF spatial audio from the Meta XR Audio SDK version 59.0.0, representing the state-of-the-art at the time of the study.

Spatial audio synthesis is now a commonplace feature on digital devices, including VR headsets. It has been used in the past to support information presentation [60] and create richer, more immersive experiences [6]. In immersive virtual environments, spatial audio has been shown to increase presence [25], social presence [58], and psychological immersion [55]. Prior empirical work on spatial audio usage in VR, however, mainly focuses on comparing spatial audio with no spatial audio (e.g., [24]). In our work, we expand upon this knowledge by investigating how experiences of spatial audio may differ based on perspective.

3 Experiment

In our study, participants performed a sound localization task under nine different perspective conditions (illustrated in Figure 3). The conditions differed in terms of auditory perspective (hearing from the perspective of the avatar or an external observer), visual perspective (seeing from the perspective of the avatar or an external observer), and rotational misalignment (aligned or rotated).

3.1 Apparatus

Participants performed all tasks in a designated experimental space (Figure 1, bottom left). They were equipped with a Meta Quest 3 headset and AKG Pro K240 Studio over-ear headphones, which delivered the virtual visual and auditory stimuli, respectively. The experiment ran on an Intel Core i7-12700H CPU 2.30 GHz computer with 16 GB of RAM, supported by an NVIDIA GeForce RTX 3060 GPU. The virtual scene was developed using Unity 2021.3.32f1.[1]

All sounds in the experiment were spatialized using state-of-the-art HRTF technology (Meta XR Audio SDK 59.0.0). We defined the spatial audio environment in Unity by specifying the positions and orientations of the audio listener (where the user's hearing is located) and the sound sources (where the audio stimulus is located). For instance, in third-person audio conditions (e.g., 3PA x 3PV), the listener was positioned where the user observed the avatar. We empirically verified that the audio was correctly rendered.

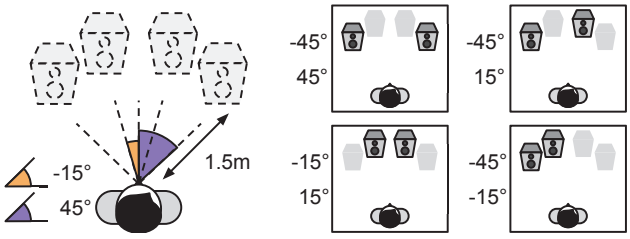

Figure 2: (Left) In each trial, pairs of speakers are placed in two of four possible locations relative to the virtual avatar, which are $1.5m$ away, at azimuths of $-45°$, $-15°$, $15°$, and $45°$. The $-15°$ and $45°$ locations are highlighted in orange and purple, respectively. (Right) Several potential speaker pair placements.

We chose to conduct our experiment using a Quest 3, as it is one of the most popular lines of commercial VR headsets [40]. We decided to use virtual auditory stimuli, presented through over-ear headphones, as it is representative of how users typically engage with spatial audio in digital devices [56].

In VR, participants controlled a full-body avatar, which was situated in the center of a $9m \times 9m \times 3m$ (length $\times$ width $\times$ height) virtual room. The avatar's movements were controlled by state-of-the-art Inverse Kinematics (IK) algorithms (RootMotion's Final IK package [57], as employed by prior work [1]), driven by the participant's head and handheld controllers' tracked poses. We used an abstract humanoid avatar, which previous research suggests provides a comparable sense of agency to a realistic representation when interacting with virtual environments [3]. Furthermore, prior work has shown that embodiment could be elicited with avatars of different appearances [48, 71]. The audio sources in our task are visually represented as tripod-mounted cylindrical speakers, designed to emit sound omnidirectionally along the horizontal plane.

3.2 Sound Localization Task

Participants performed a sound localization task in the study (Figure 1, bottom right), inspired by prior research (e.g., [65, 2, 67, 56]). The task followed a forced-choice procedure. In each trial, participants are first visually exposed to two speaker locations. After a $3-5$ s interval, one speaker emits a 0.5 s burst of white noise. Participants are then tasked with identifying which speaker produced the sound as rapidly and accurately as possible. Participants were instructed to respond by extending their left arm to the side to indicate they perceived the sound as coming from the left-most speaker of the pair, relative to the avatar, and by extending their right arm to indicate perception from the right-most speaker. Responses were counted when participants extended their arms past 0.5 m from their body. We provided audio feedback to confirm that the response was accepted, rather than indicating correctness, to prevent bias in participants' subjective evaluations of the different conditions. Participants proceeded to the next trial by returning their hands to the center position. Throughout the task, participants are instructed to refrain from moving their heads, to control for localization effects of head movement [44].

We used a forced-choice procedure instead of asking participants to directly report the position of the sound stimuli, such as by verbally indicating the apparent spatial position in terms of angles [69] or manually pointing [52], to minimize the effect of the response approach on our results. Unlike prior work, our experiment includes conditions where the locations from which participants see and hear

[1] https://augmented-perception.org/publications/2024-Auditory-Embodiment-Study.html

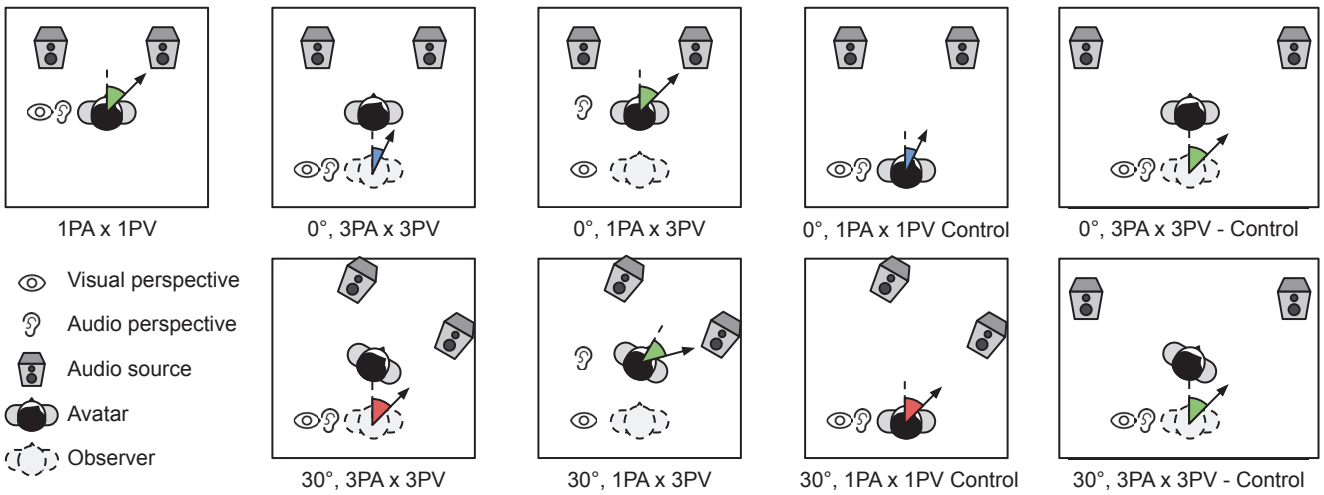

Figure 3: Participants' visual and audio perspectives in the nine conditions evaluated in our experiment. Colors are used to indicate differences in the azimuthal angles of the auditory stimuli between conditions (e.g., equivalent between the *1PA x 1PV* and *1PA x 3PV* conditions but different between the *1PA x 1PV* and *3PA x 3PV* conditions).

are separated. This introduces translational, and, in some cases, rotational misalignments between the reference frames of perception and judgment, which is known to introduce difficulties for spatial cognition [37] We designed the response approach to involve participants extending their arms, notably, instead of responding via button press, to introduce an experience of visuomotor synchrony, which prior work suggests is critical for inducing a feeling of embodiment of a virtual avatar [35].

3.2.1 Speaker Placement

In each trial, pairs of speakers are placed in two of four possible locations relative to the virtual avatar (Figure 2). The vertical position of each speaker is always aligned with participants' ears (i.e., 0° elevation), such that they are only distinguished by their relative azimuthal (i.e., horizontal) angle. We then set the azimuthal angle and distance of the speaker placements to maximize the angles between the positions of the four speakers while ensuring all speakers remain visible regardless of the visual perspective condition in our experiment. This was to control for potential effects of ventriloquism [64] in cases where only one source was visible. Considering the Quest 3 offers a horizontal field of view of 110°, we set the potential speaker locations to be at azimuths of $-45°$, $-15°$, 15°, and 45°, 1.5 *m* away from the avatar. Participants experienced all potential speaker placement pairs four times. In two out of the four instances, the sound stimulus was played from the left-most speaker, and vice versa for the remaining two. We excluded trials where speaker pairs were placed at azimuths of $-45°$ and 45°, as early pilot studies indicated that this condition was trivial. Here, it is important to note that, in some cases (e.g., when the speaker pair is placed at 15° and 45°), both speakers could end up on the same side of the avatar. Hence, in our instructions, we emphasized to participants that their task was to identify the leftmost or rightmost speaker "of the visible speakers" rather than the speaker on the auditory left or right. Participants completed 20 trials per condition.

3.3 Perspective

The experiment compared nine conditions that varied in PERSPECTIVE and the ROTATION introduced between the observer and the avatar's reference frames. All conditions are illustrated in Figure 3. Conditions are named as *Audio* × *Visual*, e.g., *1PA* × *3PV* denotes *1st Person Audio* and *3rd Person Visual*.

1PA x 1PV: Participant experience the virtual world through the avatar's eyes and ears. This first-person perspective is most commonly used in VR applications.

3PA x 3PV: Participants operate the avatar from the point of view of an external observer. Both the visual and auditory stimuli are rendered as though perceived from the perspective of a third-person observer. Participant are situated 1 *m* behind the avatar, which mimics their behaviors (e.g., changes in arm position or head rotation).

1PA x 3PV Participants *see* and control the avatar from a third-person perspective of an observer (*3PV)*, but *hear* from the avatar's perspective (*1PA)*. This leads to a virtual world that is perceived as visually and auditorily disparate.

3.3.1 Control Conditions

One of our core research questions of interest is whether auditory perspective affects embodiment and sound localization performance. Differences in auditory perspective are primarily reflected between *3PA x 3PV* and *1PA x 3PV*. However, *3PA x 3PV* differs from *1PA x 3PV* not only in terms of audio perspective but also in sound localization task difficulty. This is because adopting the external observer's audio perspective in *1PA x 3PV* that is situated 1 *m* behind the avatar reduces the azimuthal angle between audio sources, which may make differentiating between them more challenging. Moreover, the additional distance between the avatar and the speaker will reduce the volume of the audio stimulus, adding to the difficulty of distinguishing between the audio cues.

To control for these factors, we first disabled distance attenuation effects in the spatial audio synthesis to keep volume levels constant across all conditions. This design decision is further motivated by Zahorik et al. [72], who suggest that people's ability to estimate the distance of a sound source is significantly less accurate than their ability to estimate its angular direction anyhow.

We additionally introduce two additional control conditions to isolate task difficulty effects in our analysis:

1PA x 1PV - Control: We introduce *1PA x 1PV - Control* to examine whether task difficulty affects embodiment and performance when the participant fully embodies the avatar both visually and auditorily. Therefore, *1PA x 1PV - Control* sets the avatar location to the external observer's position as in *3PA x 3PV* and *1PA x 3PV*, making the task more challenging while maintaining the same perspective settings as *1PA x 1PV*.

3PA x 3PV - Control: We introduce *3PA x 3PV - Control* as a condition where the participant sees, hears, and controls a virtual avatar from the perspective of an external observer, but with the same auditory stimulus as the *1PA x 1PV* and *1PA x 3PV* conditions. We achieve this by adjusting the position of the sound sources. We adjust the potential speaker locations to be at azimuths of $-45°$, $-15°$, $15°$, and $45°$ from the external observer instead of the virtual avatar. We set the speaker distances to match the *1PA x 3PV* condition as closely as possible, resulting in distances of 1.5 *m* and 2.5 *m* for the speakers positioned at the $15°$ and $45°$ locations, respectively.

3.4 Rotation

In applications that involve third-person perspective control (e.g., [49, 13]), the observer may not always be situated directly behind the avatar, but rather at an angle, which presents a scenario of rotationally misaligned reference frames [37]. Therefore, we also explore whether introducing such a rotational misalignment influences participants' localization performance and perceived sense of embodiment. Specifically, in addition to PERSPECTIVE, we investigate the effect of a $0°$ versus $30°$ ROTATION. We chose $30°$ to test the maximal effect of rotation while keeping both relevant speakers visible within the headset, thus avoiding visibility confounds [64]. While finer angular offsets may also be interesting to investigate, we only tested $30°$ to keep the experiment length manageable.

3.5 Measures

We analyzed the following metrics:

Self-reports After each condition, participants responded to a survey that included a subset of Gonzalez-Franco and Peck's avatar embodiment questionnaire [20], addressing aspects of body ownership, agency and motor control, location of the body, and Aron et al.'s Inclusion of Other in the Self (*IOS*) scale [4] (see supplementary material). All items were evaluated using 7-point Likert scales. We followed the approach detailed in Gonzalez-Franco and Peck's [20] to compute cumulative *ownership*, *agency*, *location*, and *total embodiment* metrics using the responses we collected.

Localization performance We recorded average *response time* and *error count* for each condition to measure sound localization performance. *Response time* refers to the time between when the white noise is presented and when participant's response was recorded. *Error count* refers to the number of stimuli from the right or left speaker that the participant misidentified as coming from the left or right, respectively.

3.6 Experimental Design

Our experiment effectively examines the impact of two independent variables: PERSPECTIVE (with five levels: *1PA x 1PV*, *1PA x 3PV*, *3PA x 3PV*, *1PA x 1PV - Control*, *3PA x 3PV - Control*) and ROTATION ($0°$ and $30°$), on participants' sound localization performance and sense of embodiment. Since the angle parameter represents the rotational offset between the avatar and external observer, it is not applicable to *1PA x 1PV*. Therefore, the $0°$ and $30°$ cases in *1PA x 1PV* collapse into a single condition. As a result, we employed a 9 condition within-subject design.

To mitigate ordering effects, we counterbalanced the condition order using a Latin Square, resulting in 18 possible orders. For the first 18 participants, we used all 18 possibilities. First the last six participants, we randomly selected six out of the existing possibilities. We analyzed ordering effects (9 levels, i.e. condition orders, between-subject) and did not find a main effect of order on any of our dependent variables (all $p > .05$). The order of the speaker placements was randomized for each condition.

3.7 Experimental Procedure

After obtaining informed consent from the participant, the experimenter first introduced them to the study, the equipment involved, and the data we recorded (which was anonymized). Participants then filled out the pre-questionnaire.

Afterwards, participants proceeded through the conditions of our study. Before performing the sound localization task in each condition, participants completed three training trials to become accustomed to the perspective settings. The three trials consisted of distinguishing between pairs of speakers placed at $-15°$ and $15°$, $15°$ and $45°$ (or $-15°$ and $-45°$), and $-15°$ and $45°$ (or $-45°$ and $15°$). We excluded the $-45°$ and $45°$ case from our training trials because participants consistently achieved perfect accuracy in pilot tests. We repeated the training until participants correctly responded to two of the three trials. In the first training instance, we also asked participants to adjust the volume to a comfortable level.

Participants then proceeded to the recorded trials. After each condition, participants reported on several subjective metrics in a post-condition questionnaire. Between conditions, participants were allowed to rest for as long as they preferred. Overall, the procedure took 75 *min* per participant.

3.8 Power and Experimental Participants

Prior to conducting our study, we ran an a priori power analysis using G* Power 3.1 [15] to determine an appropriate sample size. We chose two effect sizes, $f = 0.25$ and $f = 0.5$, corresponding to small and medium effect sizes, respectively, to determine the appropriate range for the sample size. We set an alpha error probability of $\alpha = 0.05$ and a power of $\beta = 0.8$. Since each condition consists of 20 trials, we set the number of measurements given to G* Power as 180 (20 per condition). Some measures are observed once per trial, so we also tested entering 9 as the number of measurements. The number of groups was dependent upon the within-subject factors, which, in the case of our experiment, was 9. Finally, the correlation among repeated measures was left at the default value of 0.5. The power analysis revealed that we would need 18 participants to obtain a medium effect size. We also considered prior experiments on perspective effects on interaction tasks and embodiment (e.g., [12, 63]), which had a similar number of participants.

We recruited 24 participants via snowball sampling starting from university message groups and social networks. Participants had to be 18-70 years old, without significant auditory or visual impairments that would disrupt their experience of a VR application with spatial audio. Participants received $25 as gratuity for their time.

In the pre-questionnaire, we asked participants to report their demographic information, prior experience with VR (7-point Likert scale, from 1-none to 7-expert), frequency of playing video games (from 1-never to 5-at least once a week), and level of alertness using the Stanford Sleepiness Scale [27] (from 0-asleep to 7-active, vital, alert, or wide awake). All participants (age: $M = 25$, $SD = 3$; 12 female, 12 male) reported normal or corrected-to-normal vision and normal hearing. Participants' median responses were VR experience = 2, gaming frequency = 2, and level of alertness = 6.

4 Results

First, we compared PERSPECTIVE and ROTATION in terms of subjective ratings and localization performance. Then, we analyzed relationships between dependent variables using Spearman correlations.

For the effect analysis, ordinal data (questionnaire ratings) was analyzed using an Aligned Rank Transform (ART) ANOVA [70]. Interval data (*response time*, *error count*) was analyzed using a two-factor repeated-measures ANOVA. For each data value, the *participant* was considered as a random factor, with PERSPECTIVE (4 levels: *1PA x 3PV*, *3PA x 3PV*, *1PA x 1PV - Control*, *3PA x 3PV - Control*) and ROTATION (2 levels: $0°$, $30°$) treated as within-subject independent variables. We excluded the *1PA x 1PV* condition from our initial analysis because our primary focus is on the effect of the audio perspective, and since the condition was only observed

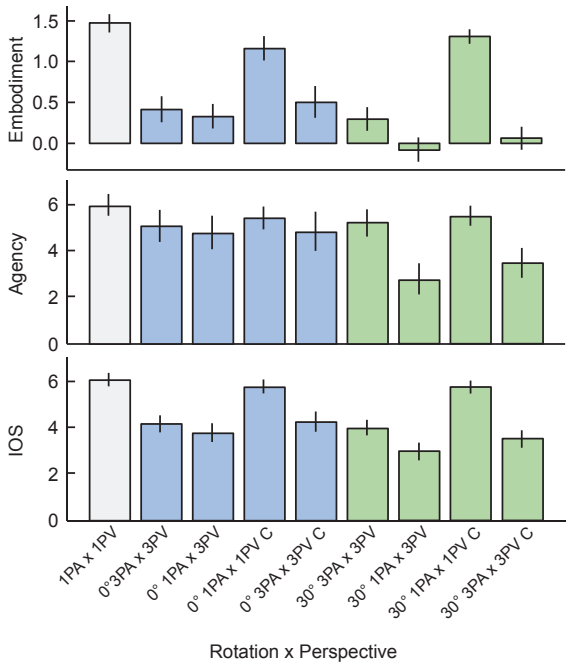

Figure 4: Effect of ROTATION and PERSPECTIVE on *embodiment* (top), *agency* (middle), and *IOS* (bottom). Error bars indicate Standard Error.

once across both angles, it does not allow for a cross-factorial design. Instead, we subsequently report pairwise comparisons between the *1PA x 1PV* condition and the conditions included in our cross-factorial analysis in Section 4.3. When the assumption about the normality of residuals and homogeneity was violated (Shapiro-Wilk test $p < .05$), we analyzed them using the ART. When the equal variances assumption was violated (Mauchley's test $p < .05$), we corrected degrees of freedom using Greenhouse-Geisser. When needed, pairwise post-hoc tests (Bonferroni adjusted p-values) were performed. For the sake of concision, we focus on reporting statistically significant main effects and interactions. The statistical analysis was performed using IBM SPSS 29 [30]. Detailed results are provided in the supplementary materials.

4.1 Subjective Ratings

In summary, the analysis showed no significant main effects of PERSPECTIVE or ROTATION on any of the collected subjective ratings. However, significant interaction effects suggest that when the participant's visual perspective was set to that of the external observer (i.e., *3PV)*, adopting the avatar's audio perspective potentially (i.e., *1PA)* creates a stronger sense of embodiment and self-avatar merging. Results additionally indicate that participants felt the highest sense of embodiment, agency, and self-avatar merging when adopting the visual perspective of the avatar they were controlling (i.e., *1PV*). We summarize *embodiment*, *agency*, and *IOS* results in Figure 4. Significant post-hoc results are listed in Figure 5. We report only the significant interaction effects and post-hoc analyses below.

Embodiment: The analysis showed an interaction effect between PERSPECTIVE and ROTATION ($F_{1,3} = 6.74$, $p < .001$, $\eta_p^2 = .23$). Post-hoc tests showed that participants reported a higher sense of embodiment in the *1PA x 1PV - Control* conditions (i.e., for both 30°and 0°) compared to all conditions where participants viewed the avatar from a third-person perspective (all $p < .01$; i.e., *1PA x 1PV - Control* higher than *1PA x 3PV*, *3PA x 3PV* and *3PA x 3PV - Control*).

Agency: The analysis showed an interaction effect between PERSPECTIVE and ROTATION ($F_{1,3} = 3.05$, $p = .03$, $\eta_p^2 = .12$). Post-hoc tests suggest that participants reported a significantly higher sense of agency in 30° *1PA x 3PV* compared to 30° *3PA x 3PV* ($p < .05$). Furthermore, participants reported greater agency in 0° *1PA x 1PV - Control* compared to 30° *3PA x 3PV* ($p < .01$) and 30° *3PA x 3PV - Control* ($p < .05$), as well as the 30° *1PA x 1PV - Control* compared to 30° *3PA x 3PV* ($p < .01$).

IOS: The analysis showed an interaction effect between PERSPECTIVE and ROTATION ($F_{1,3} = 4.86$, $p = .004$, $\eta_p^2 = .18$). Post-hoc tests show that participants reported a significantly higher IOS rating in 30° *1PA x 3PV* compared to 30° *3PA x 3PV* ($p = .03$). IOS ratings were also significantly higher in 0° *1PA x 1PV - Control* compared to all conditions where participants viewed the avatar from a third-person perspective (all $p < .01$). 30° *1PA x 1PV - Control* similarly showed higher IOS ratings than all third-person visual perspective conditions, besides 0° *3PA x 3PV - Control* (all $p < .05$).

4.2 Sound Localization Performance

In summary, participants were faster in *1PA x 3PV* and *3PA x 3PV - Control* than in *3PA x 3PV*. *Error count* was not significantly affected by either PERSPECTIVE or ROTATION.

Response time: An ANOVA showed a main effect of PERSPECTIVE on response time ($F_{1,3} = 4.74$, $p = .005$, $\eta_p^2 = .17$), but not ROTATION or any interaction effects. Post-hoc tests suggest that participants were significantly faster in *1PA x 3PV* ($p = .04$), and *3PA x 3PV - Control* ($p < .01$), than in *3PA x 3PV*. The average difference, however, was less than 100 *ms*. We summarize response times by PERSPECTIVE in Figure 6.

Error count: The ART analysis showed no effect of PERSPECTIVE or ROTATION on *error count*.

4.3 Comparison with 1PA x 1PV

We conducted a series of pairwise comparisons (Bonferroni adjusted p-values) between *1PA x 1PV* and the eight conditions included in our cross-factorial analysis.

Participants were significantly *faster* in *1PA x 1PV* than in 0° *3PA x 3PV* ($p < .001$), and made fewer *errors* in *1PA x 1PV* compared to 0° *3PA x 3PV* ($p = .01$), 30° *1PA x 1PV - Control* ($p < .001$), and 30° *3PA x 3PV* ($p < .001$). They also reported a significantly higher sense of *agency* in *1PA x 1PV* compared to 30° *3PA x 3PV* and 30° *3PA x 3PV - Control* (both $p < .001$). Lastly, in they reported a significantly higher sense of *IOS*, *embodiment*, *location*, and *ownership* in *1PA x 1PV* compared to all conditions where participants viewed the avatar from a third-person perspective (all $p < .001$). There were no significant differences between *1PA x 1PV* and *1PA x 1PV - Control* on any measures.

4.4 Correlation Analysis

We additionally analyzed how *IOS*, *embodiment*, *ownership*, *agency*, *location*, and localization performance correlated with each other using Spearman correlation. Overall, all subjective ratings showed a positive correlation with each other (all $\rho > .32$, $p < .001$). *Response time* was negatively associated with *error count* ($\rho = -.22$, $p = .001$). We did not observe any significant correlations between any of the performance metrics and the subjective ratings ($p > .05$).

5 Discussion

In this paper, we conducted an empirical study with 24 participants to investigate the effect of PERSPECTIVE and ROTATION on embodiment and sound localization performance.

5.1 First- versus Third-person Audio Perspective

1PA x 3PV and *3PA x 3PV* represent the two main conditions in our experiment in which the visual perspective is controlled, while the audio perspective is set to the first- and third-person perspectives, respectively. A direct comparison between the *1PA x 3PV* and *3PA x 3PV* conditions indicates that when viewing an avatar

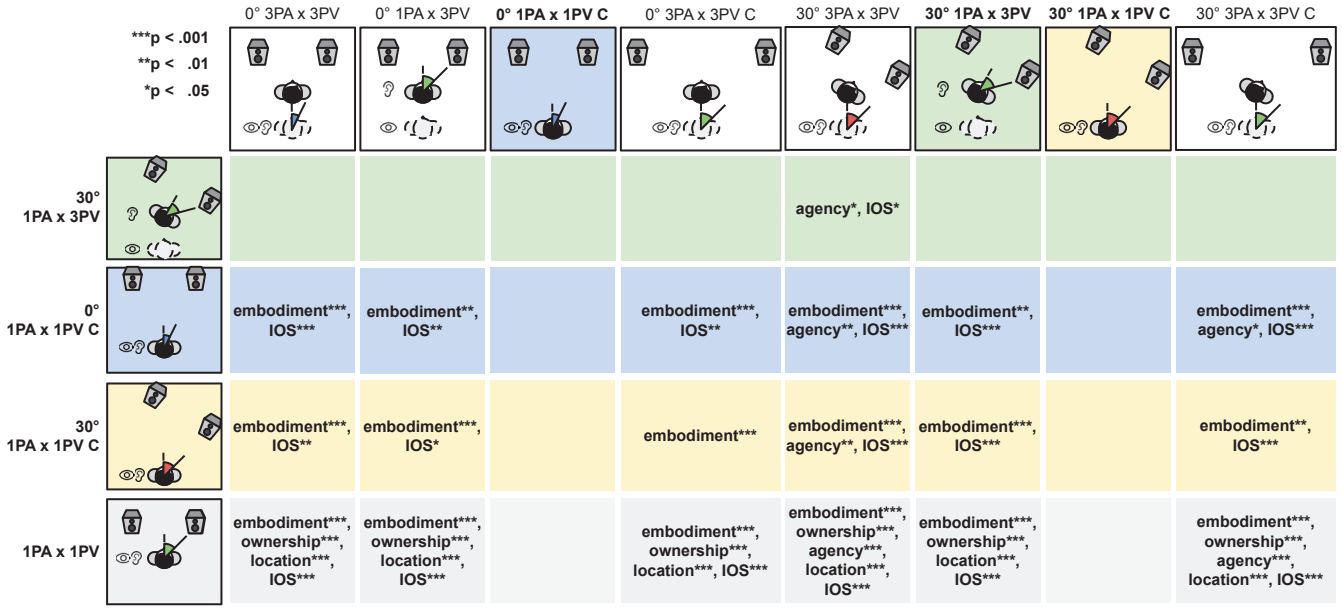

Figure 5: Significant post-hoc results for the subjective ratings (*embodiment*, *ownership*, *agency*, *location*, *IOS*).

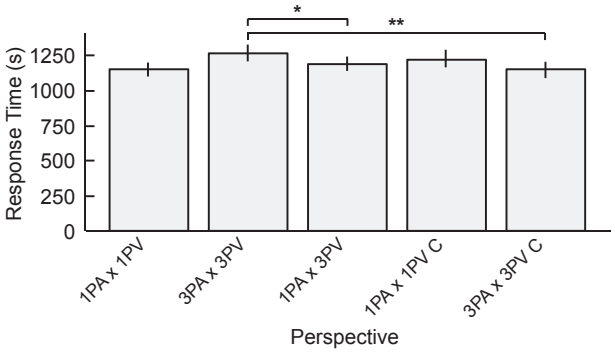

Figure 6: Effect of PERSPECTIVE on *response time*. Significance levels: ** $p < .01$, * $p < .05$. Error bars indicate Standard Error.

from an external point of view, adopting the audio perspective of the avatar increases perception of agency and self-avatar merging when there is a rotational misalignment between the avatar and observer's frames of reference. This suggests that the audio perspective may indeed influence users' perceived sense of embodiment (RQ1) and that its effects are mediated by the extent to which the audio and visual reference frames are misaligned (RQ2).

However, comparing the *1PA x 3PV* and *3PA x 3PV* conditions is confounded by speaker locations, which are more perceptually difficult to distinguish from the third-person perspective. The backward translational offset introduced to the audio perspective in *3PA x 3PV* narrows the space between speakers. In 30° conditions, the rotational offset further skews the speaker locations to one side relative to the observer. Our results further indicate that participants were approximately 80*ms* faster in *1PA x 3PV* than in *3PA x 3PV*, without incurring more errors, which corroborates the idea that *3PA x 3PV* represents a more perceptually difficult task.

Yet, we provide evidence that perceptual difficulty is not solely responsible for the differences we observed in perceived agency and self-avatar merging. To isolate task difficulty effects in our analysis, we included *1PA x 1PV - Control* as a parallel to *1PA x 1PV* allowing us to examine whether task difficulty affects embodiment and performance when the participant embodies the avatar both visually and auditorily. The visual perspective is controlled between *1PA x 1PV - Control* and *1PA x 1PV*, while the auditory stimulus is identical to *3PA x 3PV* and *1PA x 3PV*, respectively. Our results indicate that *1PA x 1PV - Control* and *1PA x 1PV* do not differ significantly in terms of performance and embodiment-related self-reported metrics, suggesting that task difficulty does not fully account for differences between *1PA x 3PV* and *3PA x 3PV*.

We additionally introduced *3PA x 3PV - Control* as a condition in which the participant controls a virtual avatar from the audio and visual perspectives of an external observer, with the auditory stimulus equivalent to that of the *1PA x 1PV* and *1PA x 3PV* conditions. Participants reported a higher sense of agency in the *1PA x 1PV* condition than in the 30° *3PA x 3PV - Control* condition. They also reported a higher sense of IOS, embodiment, location, and ownership in *1PA x 1PV* than all *3PA x 3PV - Control* conditions. Since perceptual difficulty was controlled for in this comparison, the subjective differences between *1PA x 1PV* and *3PA x 3PV - Control* can more likely be attributed to their audio-visual perspective differences. On the other hand, we observed no significant differences in the self-reported metrics between *1PA x 3PV* and *3PA x 3PV - Control*. Like *1PA x 3PV* and *3PA x 3PV*, *1PA x 3PV* and *3PA x 3PV - Control* represent first- and third-person perspective audio conditions, respectively; however, in contrast to *3PA x 3PV*, *3PA x 3PV - Control* aligns the speaker locations with the observer's spatial reference frame. This increases the similarity between the avatar and observer's spatial reference frames, which may have supported identification with the avatar regardless of whether both audio and visual perspectives were aligned with the observer. Indeed, the results for *embodiment*, *IOS*, and *agency* indicate that the *3PA x 3PV - Control* falls in between *3PA x 3PV* and *1PA x 3PV*. This may serve as additional support for the mediating role of the misaligned reference frames on the effect of audio perspective.

Lastly, our correlation analysis indicates that there is no significant correlation between task performance and participants' subjective ratings (RQ3), providing further support for the effect of audio perspective, rather than purely task difficulty, on embodiment.

Our results show that the choice of audio perspective will influence the perceptual difficulty of localizing audio cues. This

is, again, suggested by differences in *response time* in our results (i.e., between *3PA x 3PV* and *1PA x 1PV*and *1PA x 1PV - Control*). While our controls show that this does not account for our differences in perceived embodiment, it is an important consideration nonetheless for all other applications.

In summary, our study results indicate that the audio perspective can influence embodiment and its effects are mediated by differences between the audio and visual reference frames. Aligning the audio perspective with the avatar increases perceived agency and self-avatar merging, particularly in scenarios where the user's visual perspective is misaligned with that of the avatar. Misalignment may also result in perceptual difficulties for localization, although this does not fully account for the aforementioned differences.

These results generally align with effects of first- versus third-person visual perspective reported in prior work [12], wherein the first-person perspective generally yields increased identification with and agency over the avatar. Effects of perspective on localization performance also align with literature on perception and spatial reference frames [37]. Ultimately, reduced azimuthal angles and spatial reference frame misalignments will adversely impact the user's localization capabilities.

5.2 Visual versus Auditory Perspective

While our results indicate that first-person audio supports embodiment, its effects are notably less pronounced than those of visual perspective. Across our embodiment-related metrics, conditions that aligned the participant's visual perspective with that of the avatar consistently yielded higher ratings than those in which their visual perspective was set to that of an external observer. These findings are consistent with the results of previous research, which have similarly demonstrated visual dominance over other aspects such as synchrony of visuo-tactile feedback [63]. The comparatively small effect of the audio perspective indicates that using third-person audio may be an acceptable design decision, despite its incongruence and adverse effect on embodiment-related metrics, especially if the task entails localization performance requirements that are better addressed with the third-person audio perspective. However, it remains possible that other spatial setups may find larger effects of audio perspective.

5.3 Implications for Design

Based on our findings, we have gained the following insights into the selection of audio perspective for VR applications:

- Aligning the audio perspective with the avatar when using a third-person visual perspective can increase agency and self-avatar merging, especially with rotationally misaligned reference frames.
- The choice of audio perspective may affect localization task difficulty and performance, but we found no relationship between audio perspective and embodiment within our scenario.
- The effect of audio perspective is minor compared to visual perspective, making the adverse effect of third-person audio on embodiment potentially acceptable.

5.4 Limitations and Future Work

In this work, we demonstrate that when viewing an avatar from a third-person perspective, adopting the avatar's auditory perspective may increase embodiment in scenarios where the third-person's spatial reference frame is rotationally misaligned from that of the avatar. Our experiment was intentionally designed to be highly controlled, particularly in terms of task difficulty, visual access to task targets, and response approach. However, these variables may also affect embodiment and localization performance.

In our experiment, we set the relative positions between the avatar, speaker, and external observer to maximize the azimuthal angle between speakers while ensuring that across all our conditions, the speakers are visible. We opted for a spatial arrangement that ensured all speakers were visible to account for potential ventriloquism effects [64] that could distort how participants attribute sounds to visual sources, especially in cases where only one audio source is visible. However, this constrained the potential arrangement of speakers we evaluated, as well as the rotational misalignment we introduced. In contrast, audio cues are often used in digital interfaces to support out-of-view awareness [22]. Scenarios involving larger rotational misalignments in perspectives, such as in teleoperation [41], are also common. Moreover, while we only evaluated two ROTATION conditions (0° and 30°), the effect of more granular rotations remains unclear. Future work could therefore investigate alternative spatial arrangements, including different rotational misalignments both less than and beyond 30°.

To similarly control for the effect of the response approach [5] on our results, we restricted participants' movements throughout the experiment to only extending their arms. However, the permitted movement may also have mediated aspects such as embodiment and location. Whether the auditory perspective effects will translate to more dynamic scenarios, such as game-play with full-body movement, poses an interesting question for follow-up work.

Last but not least, our experiment currently relies on virtually synthesized spatial audio based on a generic HRTF. While this is representative of current VR experiences, future replication studies may consider using a personalized HRTF instead to provide a more accurate auditory experience. Additionally, we disabled distance attenuation to control for volume effects across conditions; however, it would be interesting to test whether this feature influences the auditory experience and, in turn, perceived embodiment. Finally, seeing how our results compare to those captured from an apparatus conventionally used for sound localization studies (i.e., loudspeaker arrays) would provide valuable insights into the applicability of our findings in VR to real-world settings.

6 Conclusion

We presented the results of an empirical study with 24 participants investigating the effects of auditory perspective on user sound localization performance and subjective experience. We observed that when viewing a virtual avatar from a third-person perspective, adopting the auditory perspective of the avatar may increase agency and self-avatar merging. In addition, our results provide empirical support for the larger effect of visual perspective relative to the effect of auditory perspective.

Our results suggest that when sense of agency or self-avatar merging is a desirable experiential factor, aligning the auditory perspective with the avatar is advantageous even though it introduces an incongruency between the user's visual and auditory perspectives in cases of third-person viewing. However, if visual and auditory congruence is a priority, setting the auditory perspective to align with the third-person may also be acceptable as it has a comparatively lower effect on these experiential factors than the visual perspective. We believe that our insights will provide valuable guidance for optimizing the design of immersive experiences.

Acknowledgments

We thank all involved peers, participants, and anonymous reviewers, especially Tiffany Luong, Portia Wang, Conrad Borchers, and Alexander Wang for their input throughout the project. This work was supported in part by the Croucher Foundation.

References

[1] P. Abtahi, M. Gonzalez-Franco, E. Ofek, and A. Steed. I'm a giant: Walking in large virtual environments at high speed gains. In *Proceedings of the 2019 CHI Conference on Human Factors in Comput-*

ing Systems, CHI '19, p. 1–13. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3290605.3300752 3

[2] A. Ahrens, K. D. Lund, M. Marschall, and T. Dau. Sound source localization with varying amount of visual information in virtual reality. *PloS one*, 14(3):e0214603, 2019. 3

[3] F. Argelaguet, L. Hoyet, M. Trico, and A. Lecuyer. The role of interaction in virtual embodiment: Effects of the virtual hand representation. In *2016 IEEE Virtual Reality (VR)*, pp. 3–10, 2016. doi: 10.1109/VR.2016.7504682 3

[4] A. Aron, E. N. Aron, and D. Smollan. Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of personality and social psychology*, 63(4):596, 1992. 5, 1

[5] H. Bahu, T. Carpentier, M. Noisternig, and O. Warusfel. Comparison of different egocentric pointing methods for 3d sound localization experiments. *Acta acustica united with Acustica*, 102(1):107–118, 2016. 8

[6] P. Bala, R. Masu, V. Nisi, and N. Nunes. "when the elephant trumps": A comparative study on spatial audio for orientation in 360º videos. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 1–13. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3290605.3300925 3

[7] C. C. Berger, M. Gonzalez-Franco, A. Tajadura-Jiménez, D. Florencio, and Z. Zhang. Generic hrtfs may be good enough in virtual reality. improving source localization through cross-modal plasticity. *Frontiers in neuroscience*, 12:21, 2018. 3

[8] O. Blanke, T. Landis, L. Spinelli, and M. Seeck. Out-of-body experience and autoscopy of neurological origin. *Brain*, 127(2):243–258, 2004. 1

[9] M. Botvinick and J. Cohen. Rubber hands 'feel'touch that eyes see. *Nature*, 391(6669):756–756, 1998. 2

[10] A. Cheymol, R. Fribourg, A. Lécuyer, J.-M. Normand, and F. Argelaguet. Beyond my real body: Characterization, impacts, applications and perspectives of "dissimilar" avatars in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 29(11):4426–4437, 2023. doi: 10.1109/TVCG.2023.3320209 1

[11] A. Covaci, A.-H. Olivier, and F. Multon. Third person view and guidance for more natural motor behaviour in immersive basketball playing. In *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology*, VRST '14, p. 55–64. Association for Computing Machinery, New York, NY, USA, 2014. doi: 10.1145/2671015.2671023 2

[12] H. G. Debarba, E. Molla, B. Herbelin, and R. Boulic. Characterizing embodied interaction in first and third person perspective viewpoints. In *2015 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 67–72. IEEE, 2015. 2, 5, 8

[13] A. Denisova and P. Cairns. First person vs. third person perspective in digital games: do player preferences affect immersion? In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pp. 145–148, 2015. 2, 5

[14] H. H. Ehrsson. The experimental induction of out-of-body experiences. *Science*, 317(5841):1048–1048, 2007. 2

[15] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang. Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, 41(4):1149–1160, 2009. 5

[16] D. Freeman, S. Reeve, A. Robinson, A. Ehlers, D. Clark, B. Spanlang, and M. Slater. Virtual reality in the assessment, understanding, and treatment of mental health disorders. *Psychological medicine*, 47(14):2393–2400, 2017. 1

[17] R. Fribourg, F. Argelaguet, A. Lécuyer, and L. Hoyet. Avatar and sense of embodiment: Studying the relative preference between appearance, control and point of view. *IEEE transactions on visualization and computer graphics*, 26(5):2062–2072, 2020. 2

[18] H. Galvan Debarba, S. Bovet, R. Salomon, O. Blanke, B. Herbelin, and R. Boulic. Characterizing first and third person viewpoints and their alternation for embodied interaction in virtual reality. *PloS one*, 12(12):e0190109, 2017. 2

[19] W. G. Gardner and K. D. Martin. Hrtf measurements of a kemar. *The Journal of the Acoustical Society of America*, 97(6):3907–3908, 1995. 3

[20] M. Gonzalez-Franco and T. C. Peck. Avatar embodiment. towards a standardized questionnaire. *Frontiers in Robotics and AI*, 5, 2018. doi: 10.3389/frobt.2018.00074 2, 5, 1

[21] G. Gorisse, O. Christmann, E. A. Amato, and S. Richir. First-and third-person perspectives in immersive virtual environments: presence and performance analysis of embodied users. *Frontiers in Robotics and AI*, 4:33, 2017. 2

[22] U. Gruenefeld, A. Löcken, Y. Brueck, S. Boll, and W. Heuten. Where to look: Exploring peripheral cues for shifting attention to spatially distributed out-of-view objects. In *Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, AutomotiveUI '18, p. 221–228. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3239060.3239080 8

[23] B. S. Hasler, B. Spanlang, and M. Slater. Virtual race transformation reverses racial in-group bias. *PloS one*, 12(4):e0174965, 2017. 1

[24] C. Hendrix and W. Barfield. Presence in virtual environments as a function of visual and auditory cues. In *Proceedings Virtual Reality Annual International Symposium'95*, pp. 74–82. IEEE, 1995. 3

[25] C. Hendrix and W. Barfield. The sense of presence within auditory virtual environments. *Presence: Teleoperators & Virtual Environments*, 5(3):290–301, 1996. 3

[26] H. E. Hershfield, D. G. Goldstein, W. F. Sharpe, J. Fox, L. Yeykelis, L. L. Carstensen, and J. N. Bailenson. Increasing saving behavior through age-progressed renderings of the future self. *Journal of marketing research*, 48(SPL):S23–S37, 2011. 2

[27] E. Hoddes, V. Zarcone, and W. Dement. Stanford sleepiness scale. *Enzyklopädie der Schlafmedizin*, 1184, 1972. 5

[28] M. Hoppe, A. Baumann, P. C. Tamunjoh, T.-K. Machulla, P. W. Woźniak, A. Schmidt, and R. Welsch. There is no first- or third-person view in virtual reality: Understanding the perspective continuum. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3491102.3517447 2

[29] T. Huisman, A. Ahrens, and E. MacDonald. Ambisonics sound source localization with varying amount of visual information in virtual reality. *Frontiers in Virtual Reality*, 2:722321, 2021. 3

[30] IBM. Ibm spss software. https://www.ibm.com/spss, 2024. Accessed: 2024-05-03. 6

[31] F. Immohr, G. Rendle, A. Lammert, A. Neidhardt, V. M. Z. Heyde, B. Froehlich, and A. Raake. Evaluating the effect of binaural auralization on audiovisual plausibility and communication behavior in virtual reality. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, pp. 849–858, 2024. doi: 10.1109/VR58804.2024.00104 2

[32] M. Kato, H. Uematsu, M. Kashino, and T. Hirahara. The effect of head motion on the accuracy of sound localization. *Acoustical science and technology*, 24(5):315–317, 2003. 3

[33] K. Kilteni, R. Groten, and M. Slater. The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments*, 21(4):373–387, 2012. 2

[34] K. Kilteni, A. Maselli, K. P. Kording, and M. Slater. Over my fake body: body ownership illusions for studying the multisensory basis of own-body perception. *Frontiers in human neuroscience*, 9:119452, 2015. 2

[35] K. Kilteni, A. Maselli, K. P. Kording, and M. Slater. Over my fake body: body ownership illusions for studying the multisensory basis of own-body perception. *Frontiers in Human Neuroscience*, 9, 2015. doi: 10.3389/fnhum.2015.00141 4

[36] M. Kim, J. Lee, C. Kim, and J. Kim. Tpvr: User interaction of third person virtual reality for new presence and experience. *Symmetry*, 10(4):109, 2018. 2

[37] R. L. Klatzky. Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In *Spatial cognition: An interdisciplinary approach to representing and processing spatial knowledge*, pp. 1–17. Springer, 1998. 2, 4, 5, 8

[38] S. Klockgether and S. van de Par. Just noticeable differences of spatial cues in echoic and anechoic acoustical environments. *The Journal of the Acoustical Society of America*, 140(4):EL352–EL357, 2016. 2

[39] L. Lin and S. Jörg. Need a hand? how appearance affects the vir-

tual hand illusion. In *Proceedings of the ACM symposium on applied perception*, pp. 69–76, 2016. 2

[40] T. Luong, Y. F. Cheng, M. Möbus, A. Fender, and C. Holz. Controllers or bare hands? a controlled evaluation of input techniques on interaction performance and exertion in virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 3

[41] M. Macchini, M. Lortkipanidze, F. Schiano, and D. Floreano. The impact of virtual reality and viewpoints in body motion based drone teleoperation. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 511–518, 2021. doi: 10.1109/VR50410.2021.00075 8

[42] P. Majdak, M. J. Goupell, and B. Laback. 3-d localization of virtual sound sources: Effects of visual environment, pointing method, and training. *Attention, perception, & psychophysics*, 72(2):454–469, 2010. 3

[43] D. A. Mauro, R. Mekuria, and M. Sanna. Binaural spatialization for 3d immersive audio communication in a virtual world. In *Proceedings of the 8th Audio Mostly Conference*, pp. 1–8, 2013. 3

[44] K. I. McAnally and R. L. Martin. Sound localization with head movement: implications for 3-d audio displays. *Frontiers in neuroscience*, 8:79254, 2014. 3

[45] A. Meshram, R. Mehra, H. Yang, E. Dunn, J.-M. Franm, and D. Manocha. P-hrtf: Efficient personalized hrtf computation for high-fidelity spatial sound. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pp. 53–61, 2014. doi: 10.1109/ISMAR.2014.6948409 3

[46] D. Monteiro, H.-N. Liang, W. Xu, M. Brucker, V. Nanjappan, and Y. Yue. Evaluating enjoyment, presence, and emulator sickness in vr games based on first-and third-person viewing perspectives. *Computer Animation and Virtual Worlds*, 29(3-4):e1830, 2018. 2

[47] B. C. Moore. *An introduction to the psychology of hearing*. Brill, 2012. 2

[48] J.-M. Normand, E. Giannopoulos, B. Spanlang, and M. Slater. Multisensory stimulation can induce an illusion of larger belly size in immersive virtual reality. *PloS one*, 6(1):e16128, 2011. 3

[49] A. Padmanabha, J. Gupta, C. Chen, J. Yang, V. Nguyen, D. J. Weber, C. Majidi, and Z. Erickson. Independence in the home: A wearable interface for a person with quadriplegia to teleoperate a mobile manipulator. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, p. 542–551. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3610977.3634964 2, 5

[50] M. Parger, J. H. Mueller, D. Schmalstieg, and M. Steinberger. Human upper-body inverse kinematics for increased embodiment in consumer-grade virtual reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, VRST '18. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3281505.3281529 2

[51] T. C. Peck, S. Seinfeld, S. M. Aglioti, and M. Slater. Putting yourself in the skin of a black avatar reduces implicit racial bias. *Consciousness and cognition*, 22(3):779–787, 2013. 2

[52] J.-M. Pernaux, M. Emerit, and R. Nicol. Perceptual evaluation of binaural sound synthesis: the problem of reporting localization judgments. In *Audio Engineering Society Convention 114*, Mar 2003. 3

[53] S. Perrett and W. Noble. The effect of head rotations on vertical plane sound localization. *The Journal of the Acoustical Society of America*, 102(4):2325–2332, 1997. 3

[54] D. Poirier-Quinot, M. S. Lawless, P. Stitt, and B. F. Katz. Hrtf performance evaluation: Methodology and metrics for localisation accuracy and learning assessment. *Advances in Fundamental and Applied Research on Spatial Audio*, 2022. 3

[55] T. Potter, Z. Cvetković, and E. De Sena. On the relative importance of visual and spatial audio rendering on vr immersion. *Frontiers in Signal Processing*, 2:904866, 2022. 3

[56] C. Rajguru, G. Brianza, and G. Memoli. Sound localization in web-based 3d environments. *Scientific Reports*, 12(1):12107, 2022. 3

[57] RootMotion. Final ik. `https://assetstore.unity.com/packages/tools/animation/final-ik-14290`, 2024. Accessed: 2024-04-30. 3

[58] S. Roßkopf, L. Kroczek, F. Stärz, M. Blau, S. Van de Par, and A. Mühlberger. Comparable sound source localization of plausible auralizations and real sound sources evaluated in a naturalistic eye-tracking task in virtual reality. 2023. 3

[59] P. Salamin, D. Thalmann, and F. Vexo. The benefits of third-person perspective in virtual and augmented reality? In *Proceedings of the ACM Symposium on Virtual Reality Software and Technology*, VRST '06, p. 27–30. Association for Computing Machinery, New York, NY, USA, 2006. doi: 10.1145/1180495.1180502 2

[60] N. Sawhney and C. Schmandt. Nomadic radio: speech and audio interaction for contextual messaging in nomadic environments. *ACM Trans. Comput.-Hum. Interact.*, 7(3):353–383, sep 2000. doi: 10.1145/355324.355327 3

[61] N. S. Schutte and E. J. Stilinović. Facilitating empathy through virtual reality. *Motivation and emotion*, 41:708–712, 2017. 1

[62] M. Slater, B. Spanlang, and D. Corominas. Simulating virtual environments within virtual environments as the basis for a psychophysics of presence. *ACM transactions on graphics (TOG)*, 29(4):1–9, 2010. 2

[63] M. Slater, B. Spanlang, M. V. Sanchez-Vives, and O. Blanke. First person experience of body transfer in virtual reality. *PloS one*, 5(5):e10564, 2010. 2, 5, 8

[64] D. A. Slutsky and G. H. Recanzone. Temporal and spatial dependency of the ventriloquism effect. *Neuroreport*, 12(1):7–10, 2001. 4, 5, 8

[65] M. A. Steadman, C. Kim, J.-H. Lestang, D. F. Goodman, and L. Picinali. Short-term effects of sound localization training in virtual reality. *Scientific Reports*, 9(1):18284, 2019. 3

[66] I. E. Sutherland et al. The ultimate display. In *Proceedings of the IFIP Congress*, vol. 2, pp. 506–508. New York, 1965. 1

[67] C. Valzolgher, G. Verdelet, R. Salemme, L. Lombardi, V. Gaveau, A. Farné, and F. Pavani. Reaching to sounds in virtual reality: A multisensory-motor approach to promote adaptation to altered auditory cues. *Neuropsychologia*, 149:107665, 2020. doi: 10.1016/j.neuropsychologia.2020.107665 3

[68] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, 94(1):111–123, 07 1993. doi: 10.1121/1.407089 2, 3

[69] F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening. II: Psychophysical validation. *The Journal of the Acoustical Society of America*, 85(2):868–878, 02 1989. doi: 10.1121/1.397558 3

[70] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, p. 143–146. Association for Computing Machinery, New York, NY, USA, 2011. doi: 10.1145/1978942.1978963 5

[71] N. Yee and J. Bailenson. The Proteus Effect: The Effect of Transformed Self-Representation on Behavior. *Human Communication Research*, 33(3):271–290, 07 2007. doi: 10.1111/j.1468-2958.2007.00299.x 3

[72] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst. Auditory distance perception in humans: A summary of past and present research. *ACTA Acustica united with Acustica*, 91(3):409–420, 2005. 4

[73] D. Zotkin, R. Duraiswami, and L. Davis. Rendering localized spatial audio in a virtual auditory space. *IEEE Transactions on Multimedia*, 6(4):553–564, 2004. doi: 10.1109/TMM.2004.827516 3