Beyond Vision: The Sound of Motion in Virtual Reality

HIEU CHU, University of Twente, The Netherlands

Sound plays a vital role in shaping perceptual experience in Virtual Reality (VR), particularly in environments that require accurate responses to dynamic stimuli. While prior research has demonstrated the value of spatial and immersive audio for enhancing presence, the impact of auditory motion cues on the perception of movement remains underexplored. Therefore, in this work, we attempt to examine how audio cues affect users' accuracy in intercepting moving object, and assess the influence of cue congruency on users' ability to discriminate speed. While qualitative feedback indicated that participants found the inclusion of sound helpful, quantitative results did not strongly support this observation, suggesting a supplementary interaction between the two modals. Overall, our findings highlight the potential of well-designed auditory motion cues to reduce visual cognitive load and enhance perceptual accuracy in virtual environments.

Additional Key Words and Phrases: Virtual Reality, Spatial Audio, Auditory Cues, Motion Perception

1 INTRODUCTION

Virtual Reality (VR) has become a prominent medium for immersive interaction, simulation, and research. Its adoption spans a wide range of domains, including gaming, training [1], healthcare [2], education [3], and scientific experimentation [4]. As the technology develops, the demand for increasingly realistic and immersive VR experiences has risen, not only among researchers, but also from consumers and developers. Game designers, for instance, rely on nuanced interactions and feedback mechanisms to create a compelling virtual environment. In VR, one of the most crucial factors in enhancing realism and engagement is the accurate object perception within these environments. Accurate spatial awareness underpins navigation, object interaction, and responsiveness, especially in dynamic tasks such as object avoidance, timing, or trajectory estimation, highlighting the importance of accurately perceiving motion and location in a virtual space.

While the visual modality has historically dominated VR design, auditory input also plays an essential role in shaping users' perception. Visual modalities in VR have advanced significantly, specifically improvements in render resolution, frame rate, and visual depth cues. However, vision remain inherently limited by occlusion, field-of-view constraints [5] [6], and cognitive overload [7]. These constraints highlight the need for other complementary sensory modalities that can augment perception beyond what vision alone can provide. Sound itself is omnidirectional and is capable of conveying information even outside the user's field of view [5]. For instance, auditory signals can helps localization [8], guide attention [5], or communicate motion [9]. Therefore, designing VR systems that account for the full range of human sensory capabilities is essential.

TScIT 43, July 4, 2025, Enschede, The Netherlands

A key component of VR audio design is spatialized audio. Techniques such as interaural level differences (ILDs), interaural time differences (ITDs), and head-related transfer functions (HRTFs) allow for precise spatial positioning of sound sources, enhancing realism and user presence [5]. Binaural rendering and real-time spatialization through headphone-based systems further enable immersive soundscapes aligned with head movement [10] and environment geometry [11]. Reviews have shown that spatial audio increases task performance [12], realism, and engagement in VR [5]. However, spatial audio has primarily been used to anchor users within static or ambient environments, rather than conveying dynamic information such as motion or trajectory.

Beyond static spatial localization, audio can also convey motion cues, which are sound features that communicate the movement of objects. Sound can dynamically convey information about the movement of objects through space, including their speed, direction, or time to arrival by modulating features like pitch [13] [14], amplitude [15], or intensity [16] over time. These dynamic sound features, when perceptually mapped to motion characteristics, serve as intuitive indicators that complement or even enhance visual information. For instance, looming sounds, where the audio intensity increases as an object approaches, have been shown to strongly influence users' time-to-contact (TTC) judgments, triggering faster and more accurate responses [16]. Similarly, rising pitch can signal acceleration or proximity [14], while rhythmic amplitude modulation serve as a temporal cue indicating rate of movement [15]. In multisensory scenarios, congruent audio-visual stimuli can facilitate perception and action, while incongruent cues may produce conflict or increase uncertainty [17] [18]. Yet, despite their proven perceptual salience, such cues are rarely studied in VR to inform user behavior or support interactive tasks.

Although VR audio research has made significant progress in improving presence and immersion, most studies have concentrated on passive listening or ambient effects within static environments. There is a lack of systematic investigation into how discrete auditory cues influence the perception of dynamic events, especially in relation to user performance on temporally sensitive tasks like object interception or motion tracking. This research gap highlights the need to bridge established findings in auditory motion perception with VR contexts, where such cues could meaningfully enhance task effectiveness and realism. To address this gap, the present study investigates how sounds can be used to enhance motion perception in VR. Specifically, we examine (1) which auditory cues are essential for perceiving object movement, (2) how these cues affect users' ability to accurately time or judge moving virtual objects, and (3) how congruent and incongruent sound–motion pairings influence user accuracy and confidence. By evaluating how individual auditory cues affect perception and user response to moving virtual objects, the research informs the design of perceptually effective VR interactions. This contribution is not only scientifically relevant in

understanding cross-modal perception, but also practically important for developers and designers seeking to implement intuitive and responsive audio feedback in VR systems.

2 PROBLEM STATEMENT

To address this gap, our study investigates how specific spatial and temporal auditory cues affect users' ability to perceive and respond to moving virtual objects in VR. The goal is to evaluate the perceptual and behavioral outcomes—such as accuracy and reaction time—associated with each auditory cue type, thereby informing future VR design principles and enhancing user experience in dynamic virtual settings.

Therefore, this research aims to address the following overarching question:

RQ: How can sounds be used to enhance motion perception in Virtual Reality?

To address this, the research will investigate these specific subquestions:

- **SQ1:** Which auditory cues are involved in motion perception?
- **SQ2:** How do different auditory cues affect users' in accurately intercepting a moving object in VR?
- **SQ3:** To what extent does auditory cues bias users' judgement of object speed?

3 RELATED WORKS

3.1 Immersive audio technologies

Sound has become an essential modality in Virtual Reality (VR), complementing visuals to increase immersion, realism, and spatial awareness. Serafin et al. (2018) provided a comprehensive review of auditory technologies in VR, highlighting their contributions to presence and interaction [5]. Their work emphasizes spatial audio as a key enabler of immersive experiences, a theme echoed by Naef et al. (2002), who demonstrated that 3D localization is a crucial part of an immersive audio rendering pipeline can significantly enhance realism [19]. Similarly, a study conducted by Fialho et al. (2021) also made use of 3D sound for spatial navigation [8]. These developments emphasize the value of sound not only as a complementary channel but as a core component in shaping VR interactions. Building on this foundation, recent research has turned to spatial audio as a means to support precise object localization, a critical factor in enabling accurate motion perception and timely user responses.

In computer graphics, accurate auditory localization not only enhances immersion but also enables users to track moving sources, anticipate trajectories, and time their responses accordingly. This makes spatial audio particularly important in experiments involving motion perception, where directionality and spatial orientation are fundamental. A work by McKenzie et al. (2019) has highlighted that the human auditory system can precisely localize a sound source based on the difference of sound signals received by the left and right ears: that is interaural time difference (ITD), interaural level difference (ILD) [20]. Another study done by Bertonati et al. (2021), which focused on auditory processing in blind individuals, also highlighted the importance of ITD and ILD in speed perception [21]. These audio technique was implemented and extended upon in a study by Peterson et al. (2010) through a location aware game. In addition to ITD and ILD, the game utilized Head Related Transfer Function (HRTFs), which provide a realistic spatialization as it also takes the sound wave's interaction with the user's head into account [22]. Binaural rendering and real-time spatialization through headphone-based systems further enable immersive soundscapes aligned with head movement and environment geometry. Reviews have shown that spatial audio increases task performance, realism, and engagement in VR [23]. These findings directly inform our experimental design by supporting the integration of ITD/ILD and HRTFs through real-time spatial audio rendering. By spatializing all sound cues with directional audio, we ensure that participants can accurately localize the object's position. This is particularly critical in VR, where motion perception must occur in three-dimensional space with accurate timing and localization.

However, the dominant focus in VR audio research remains on enhancing realism or immersion within static or ambient environments, rather than conveying dynamic information such as motion or trajectory. The effect of sound on perceptual accuracy, particularly in dynamic interactions such as intercepting or evaluating moving objects, remains underexplored. Research on VR audio has shown that spatial and immersive sound is crucial to user experience. However, these works largely overlook the role of sound in enhancing perceptual judgments of motion or timing. Our study aims to build upon these foundations by examining how specific auditory cues—beyond spatial location—can shape users' perception of motion events in VR. We adopt and extend prior findings on spatialization and realism by applying them to time-sensitive, motion-dependent tasks that better reflect interactive and perceptually demanding VR use cases.

3.2 Motion auditory cues

Auditory cues offer an effective means for conveying motion-related information in virtual environments. A particularly important aspect of motion is object velocity, and several studies have shown that sound characteristics can reliably inform users about the speed of moving stimuli. A study by Zhang et al. (2021) explored the relationship between perceived motion speed and auditory pitch. In their study, under the influence of high tone, participants perceived the object as faster and therefore reacted earlier compared to a low tone [13]. Another study by Senna et al. 2017 has shown the correlation between amplitude modulation (AM) frequency and perceived speed, whereas higher AM-frequency are preceived as moving faster [15]. Additionally, according to a finding by Lutfi et al. (1999), Doppler shifts provide the most salient cues for velocity discrimination [24]. These findings establish that auditory features, such as pitch, modulation rate, and spectral shifts, can serve as effective proxies for speed, suggesting that speed-based sound cues are viable tools for enhancing motion perception in virtual environments.

Building on this, sound is not only capable of representing speed but also inherently conveys urgency and time to arrival. This is achieved by mapping temporal changes in auditory features, such as pitch, intensity, and modulation rate, to an approaching trajectory. These dynamic sound features, when perceptually mapped to motion characteristics, serve as intuitive indicators that complement

or even enhance visual information. One foundational contribution in this area is a work by Neuhoff et al. (2016) on the looming sounds which demonstrated that rising-intensity sounds are perceived as approaching objects and tend to trigger earlier responses than physically equivalent receding or constant sounds [16]. These findings have directly inspired our experimental design to include of time-to-contact (TTC) estimation as a key measure in our first task. Another study made by Neuhoff et al. (1998) also confirmed that the pitch of a moving sound source rises as the source approaches [14]. These findings collectively suggested that temporal characteristics of sound play a critical role in motion prediction—an insight we utilized through our use of distance-based auditory cues in our first experiment task. By examining how users respond to dynamic auditory cues that vary over time, such as pitch elevation, amplitude modulation, and looming intensity, we assess how well these features support anticipation and reaction to approaching objects in a VR environment.

In the domain of cross-modal integration, Sekuler et al. (1997) and Hülsdünker et al. (2021) showed that audiovisual stimuli are more accurately judged when congruent cues are provided across modalities [17] [18]. For instance, matching visual motion with sound that increases in pitch results in better temporal predictions and spatial tracking. These findings emphasized that sound cues are not merely supplementary but can actively bias perceptual outcomes, particularly in tasks involving motion. Drawing on this insight, our second experiment was designed to examine whether auditory cues can bias perceived object speed in VR. By including both congruent and incongruent audio-visual conditions, we tested whether conflicting audio cues influence participant judgments, and to what extent they override or reinforce visual information. This design directly builds on earlier findings by exploring cross-modal effects in a dynamic VR context, where object speed discrimination is more complex and ecologically valid.

Yet despite these robust perceptual findings, few studies have tested these auditory motion cues in immersive or interactive contexts. Most research relied on simplified auditory stimuli presented in controlled laboratory environments, with limited spatial realism or user movement. The implications of these cues in realistic, three-dimensional spaces have not been systematically evaluated. Furthermore, the application of these cues to real-time interactive tasks remained an open research area. The literature provided strong evidence that auditory cues such as pitch, looming, amplitude modulation, and Doppler shift influence perception of motion, speed, and contact timing. However, these findings have rarely been applied or validated within VR, where audio is typically used to enhance atmosphere rather than perception. Our research bridges this gap by translating these perceptual insights into dynamic VR scenarios, evaluating how these cues affect user judgments in interactive tasks that reflect real-world spatial and temporal demands.

4 METHODOLOGY

To investigate how auditory cues support motion perception in Virtual Reality (VR), we designed two experimental tasks that target distinct but complementary aspects of dynamic perception: temporal anticipation and relative speed estimation. These tasks were motivated by prior research discussed in the Related Work section, which highlighted the importance of sound in shaping user responses to motion. The first task centers on TTC estimation, where participants respond to an approaching object by predicting the moment of contact with a virtual plane. This task draws directly on the concept of auditory looming and temporal modulation cues, which have been shown to influence anticipatory behavior. The second task focuses on speed discrimination, requiring participants to compare the speed of two consecutively presented objects and judge which one is faster. Together, these experiments allow us to systematically evaluate the contribution of specific auditory cues on motion-related judgments.

4.1 Experiment 1: Ball-Plane

This experiment examined whether temporal auditory cues improve participants' ability to estimate TTC. For this goal, we have setup an environment where a ball and a plane is situated in front of the viewer. Figure 1 visualizes the setup of the experiment. Once starts, the ball moves linearly at a constant speed (5-7u/s)[1] and approaches the plane to the right. Participants were instructed to press a predefined button when they believed the ball hits the plane. The auditory cues implemented were designed to provide increasing urgency as the ball approached its destination:

- Doppler: Pitch shifting due to relative motion was simulated for added realism.
- Pitch: The audio pitch gradually increase as the ball gets closer the plane.
- Intensity (volume) looming: The loudness increase to simulate looming intensity.
- Amplitude modulation: The pulse rate of a tone increased with proximity.

Each cue was presented in separate conditions. A control condition with no sound served as a baseline. Spatialized sound based on ITDs/ILDs and HRTFs was applied in all sound conditions. During the experiment, the system logs different information on the users' response, including the actual TTC and user response time (the time between the ball starts moving and the user reacts). From these, we calculated the time difference between actual and perceived TTC, categorized responses as early or late, and computed descriptive statistics per condition. This experiment directly addresses Subquestion 2: "How do different auditory cues affect users' ability to accurately intercept a moving object in VR?". To address the question, the task aims to examine the extent to which temporal audio gradients assist or bias motion judgments.

4.2 Experiment 2: Red or Blue

The second experiment explored participants' ability to distinguish which of two sequentially presented balls moved faster. This task focused on speed-based auditory cues, including Doppler effect and pitch mapping to speed. It tested whether users could leverage these cues to make accurate speed comparisons. Additionally, in conditions where sound and visual are in conflict, we wanted to see how accurate and certain the participants are, and whether sound

[1] "u" denotes "unit", which is a distance unit in Unity

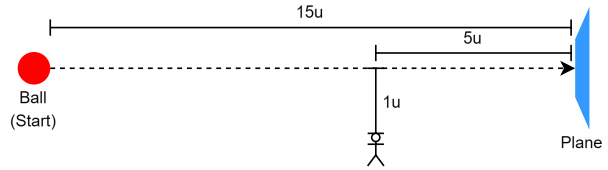

Fig. 1. The diagram visualizes the environment for the first task. In this task, a ball spawns to the left of the user and moves linearly toward a fixed plane on the right at a fixed, randomized speed between 5-7u/s. A drone sound is played from the ball. Auditory cues change gradually based on the ball's distance to the plane to help estimate time-to-contact. This setup is designed to evaluate how temporal auditory cues affect interception timing accuracy.

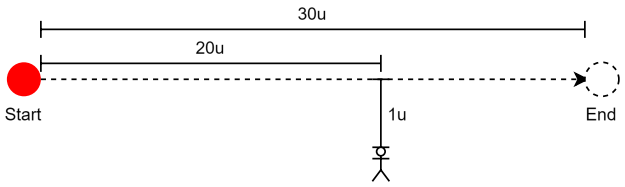

Fig. 2. The diagram visualizes the setup for the second task. This task presents two balls sequentially, one red and one blue (only red is shown), each traveling in the same direction and along the same path but at different speeds. Auditory cues are tied to each ball to reflect the ball's speed. The experiment investigates how auditory speed-based cues influence motion discrimination.

information places a perceptual bias. Figure 2 visualizes the setup of the task. Each trial involved two balls of differing speeds, with one being slower (randomized between 5-7 u/s), and the other 1.5 times faster. We wanted to make sure that it is still possible to judge the faster ball without an obvious speed difference. The balls moved along identical linear trajectories and were colored red and blue. Participants were asked to select which ball appeared faster, with an additional "Not sure" option to eliminate forced guessing and assess confidence. The auditory cues implemented were:

- Doppler: Pitch shifting due to relative motion was simulated for added realism, similarly to the first experiment.
- Pitch: Audio pitch was mapped to the ball's speed. Unlike the first experiment, the sound pitch for each ball stays constant.

As in Experiment 1, all sound conditions were spatialized. Trials were conducted across multiple conditions: Doppler effect, congruent pitch (higher pitch for faster ball), and incongruent pitch (lower pitch is faster). Each condition was tested across five trials. Each trial logged the timestamp, audio cue condition, ground truth speed of both balls, participant's response, response time, sound cue parameters, and whether the participant was uncertain. We calculated participants' accuracy based on the proportion of correct answers. Additionally, confidence (based on "Not sure" selections) and susceptibility to misleading cues in incongruent pitch conditions were measured. This experiment addresses Subquestion 3: "How do users respond to congruent versus incongruent auditory cues when judging motion in VR?". To answer this question, the task was designed to measure accuracy, reaction time, and the effect of misleading cues the experiment.

4.3 Evaluation

4.3.1 ***Statistical analysis.*** For the first task, the primary dependent variable was the time difference between the user's response and the actual collision time. Responses were further categorized into early or late decisions to assess perceptual timing biases. Descriptive statistics were calculated for each auditory condition, including mean time difference, standard deviation, absolute timing error, and the proportion of early versus late responses. To determine whether the sound conditions had a statistically significant impact on performance, a one-way ANOVA was performed on the time difference data. This method was chosen as it allows for comparison across more than two independent conditions. Given the total number of observations (over 50 trials per condition across 12 participants) and the balanced design of the task, ANOVA is appropriate and offers robust inference for between-group comparisons. In cases where a significant main effect was found ($p < 0.05$), post-hoc analysis was performed using Tukey's HSD test to identify pairwise differences between cue types.

For the Ball Speed task, we analyzed three primary outcome measures: (1) accuracy (percentage of correct responses), (2) response time (RT), and (3) uncertainty rate (frequency of "not sure" selections). Descriptive statistics were computed for each metric by condition. Similar to the first task, one-way ANOVAs were conducted for each dependent variable to evaluate whether different auditory cue mappings significantly affected user performance. Additionally, we compared the congruent and incongruent pitch conditions directly using independent t-tests to determine whether conflicting audio-visual information induced perceptual bias. The effect size was calculated using Cohen's d for further interpretation of practical significance.

4.3.2 ***Interview questions.*** Following the completion of both experimental tasks, a short semi-structured interview was conducted to gather qualitative feedback on participants' subjective experience, perceived effectiveness of the auditory cues, and decision-making strategies. These interviews were conducted verbally and responses were noted manually.

For the first task, participants were asked:

- Do you think the inclusion of sound was any helpful?
- What do you think about each specific cue?
- Which sound cue seems to be more helpful to you?

For the second task, the questions included:

- Which condition was the easiest to you?
- Do you notice what was going on with the sound? (referring to incongruent conditions)
- Do you feel like you relied more on visual or on sound?

Depending on participants' answers, additional follow-up questions were asked to explore their thought process, perception of realism, and general feedback on the auditory design and experimental procedure. These responses were used to contextualize the

quantitative findings and to uncover subjective patterns that may not be captured by behavioral metrics alone.

5 EXPERIMENT

5.1 Setup

A total of 12 participants were recruited for the study. All recruited participants report normal or corrected-to-normal vision and hearing and no prior diagnosed neurological or perceptual disorders. No restrictions were placed on age, gender, or prior VR experience, though participants must be at least 18 years of age. Prior to participation, each individual is briefed on the experimental procedures and asked to provide informed consent, in accordance with ethical research practices. All experimental tasks were developed in Unity and deployed on a Meta Quest 3 headset. The built-in stereo speakers on the headset are used to deliver audio, taking advantage of spatial audio capabilities. Sound spatialization is handled by Meta XR Audio SDK, while sound cues such as pitch, intensity, and modulation are configured programmatically. Participants interact with the system using the standard handheld Quest controllers. The study is conducted in a spacious and minimally lit room to ensure stable ambient lighting and to reduce distractions that may interfere with the passthrough view. Participants remain seated on an adjustable office chair throughout the experiment to maintain a consistent head position and minimize body movement, ensuring perceptual consistency across trials.

5.2 Procedure

Prior to beginning the session, participants were presented with an informed consent form detailing the purpose of the study, procedures, potential risks, and their rights as participants. Only those who gave written consent were allowed to proceed. Upon consent, participants were shown a brief presentation introducing the study's goals and a concise overview of the two tasks they would be performing. Before each experimental task, participants completed as many practice trials as they needed to familiarize themselves with the setup and to eliminate as much learning effect as possible.

Figure 3 visualizes the experiment procedure. The first experiment is structured into auditory conditions, including No sound, Pitch, Volume, and Amplitude. Each condition was tested in a block of 5 trials. The order of blocks was counterbalanced across participants to mitigate order effects and learning bias. Similarly, the second experiment is also divided into three conditions: Doppler, Pitch (congruent), and Pitch (incongruent). While the Doppler condition is tested in a single block of 5 trials, both Pitch congruent and incongruent conditions are tested in a block of 10 trials in total, in which the order is shuffled. This is to prevent the player from noticing the consistent pattern and relying completely on this pattern. A long break was provided between the two tasks to avoid motion sickness and fatigue. Throughout the session, the scripts automatically recorded detailed information about each trial, including timestamps, auditory cue parameters, participant responses, and performance metrics. Logs were backed up during the inter-task break to prevent data loss.

After completing both tasks, we gave the participants an interview to reflect on their experience. This included their confidence, perceived usefulness, and intuitiveness of the audio cues, and qualitative feedback on their strategies or difficulties encountered during the tasks. The session concluded with a debriefing, during which the experimenter explained the purpose of the study in more detail, clarified the nature of the sound cues used, and answered any questions participants had.

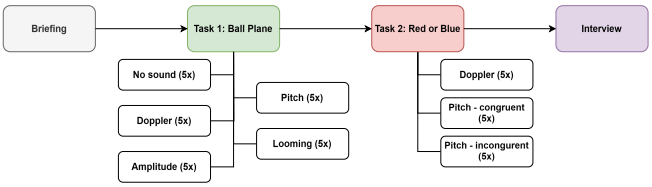

Fig. 3. The flow diagram describes the experiment procedure. In Task 1, the conditions are divided into blocks, each with 5 trials. The order of these blocks is randomized to counterbalance. For task 2, Pitch-congruent and incongruent conditions were performed in between in a shuffled order.

5.3 Experiment results

Table 1. Descriptive statistics of timing error by auditory condition in the first task.

Condition	N	Mean diff. (s)	SD	Early (%)
None	51	−0.075	0.056	88.2
Doppler	51	−0.074	0.047	94.1
Pitch	52	−0.062	0.060	78.8
Volume	51	−0.048	0.049	80.4
Amplitude	53	−0.062	0.051	90.6

Table 2. One-way ANOVA results for effect of auditory condition on time difference.

Source	SS	df	MS	F	p-value
Condition	0.0257	4	0.0064	2.29	0.0602
Residual (Error)	0.7084	253	0.0028		
Total η^2	0.035 (small effect size)				

5.3.1 ***First experiment.*** Descriptive statistics (Table 1) showed that all conditions resulted in anticipatory responses (mean time difference < 0). The Doppler and Amplitude cues yielded the highest rates of early responses (94.1% and 90.6%, respectively), while the control condition had the largest variability in timing ($SD = 0.056\ s$). These results suggest that dynamic sound cues may stabilize user response behavior. A one-way ANOVA revealed no statistically significant differences between conditions, $F(4, 253) = 2.29, p = .060$, although the effect size was small ($\eta^2 = 0.035$), indicating a trend toward condition-based variation (Table 2). Post-hoc analysis was not conducted due to the non-significant result.

Table 3. Accuracy and confidence statistics per condition in the second task

Condition	N	Acc. (%)	Acc.% SD	Unsure (%)
Doppler	53	83.0	37.9	17.0
Pitch–Congruent	55	74.5	44.0	21.8
Pitch–Incongruent	54	74.1	44.2	16.7

Table 4. Response time statistics per condition

Condition	N	Mean (s)	SD (s)	Median (s)
Doppler	53	1.846	0.850	1.722
Pitch–Congruent	55	1.541	0.781	1.291
Pitch–Incongruent	54	1.774	0.952	1.458

5.3.2 ***Experiment 2.*** Accuracy was highest in the Doppler condition (83.0%), followed by Pitch–Congruent (74.5%) and Pitch–Incongruent (74.1%), as shown in Table 3. However, one-way ANOVA revealed no statistically significant difference in accuracy between conditions, $F(2, 159) = 0.76, p = .469$. Similarly, response times did not differ significantly, $F(2, 159) = 1.85, p = .161$, though the Pitch–Congruent condition showed the fastest mean response time (1.54 s). Participants were also more likely to respond "Not sure" in the Pitch–Congruent condition (21.8%) than in Pitch–Incongruent (16.7%) or Doppler (17.0%), but this difference was not significant, $F(2, 159) = 0.30, p = .744$. Finally, the Pitch–Incongruent condition was used to assess susceptibility to misleading cues. Of all trials, 9.3% were classified as "misled" (i.e., the participant chose the ball that matched the misleading pitch cue). However, error rates were nearly identical between Congruent and Incongruent conditions (25.5% vs. 25.9%), with no significant difference found ($t = 0.056, p = .956$).

5.4 Interview results

In the first task, the majority of participants reported that pitch-based cues were the most helpful. The rising pitch appeared to assist them in anticipating the point of contact, allowing better synchronization with the ball's movement. Some participants also found the looming intensity cue helpful, but a few noted that it interfered with the spatialized sound effects, especially in peripheral positions. Amplitude modulation was often described as overly noisy or jittery, which made it distracting and difficult to interpret. Several participants indicated that, even if the cues were subtle, the inclusion of sound generally helped reduce reliance on visual tracking and lowered visual cognitive load. Notably, definitions of "contact" varied between participants, with some perceiving it as the first point of overlap, while others interpreted it as full alignment or center-overlap of the ball with the plane.

For the second task, responses were more mixed. Some participants easily perceived speed differences, especially in the congruent condition where the faster object had a higher-pitched sound. Others struggled, particularly when the speed difference was subtle or when incongruent cues were present. Many participants reported initially relying on pitch cues, but later switched to visual judgments when they noticed inconsistencies. In incongruent conditions, several participants expressed difficulty trusting auditory cues, stating that rapid switching between congruent and incongruent blocks made them skeptical of using pitch as a reliable indicator. For participants who were less confident in judging speed visually, pitch remained a fallback strategy. Across participants, a higher pitch was generally perceived as naturally corresponding to faster motion.

6 DISCUSSIONS

6.1 Statistics

The quantitative results revealed mixed outcomes across the two experimental tasks. In the first task, we hypothesized that temporally modulated auditory cues would improve users' ability to accurately judge the TTC of an approaching object. While the inclusion of sound cues did result in lower average timing error and higher proportions of early responses compared to the no-sound condition, the differences across cue types were not statistically significant. This outcome partially supports SQ2, as auditory cues appeared to influence perception but not strongly enough to yield significant differences in performance.

In the second task, we expected that congruent auditory cues would enhance discrimination accuracy, while incongruent cues might lead to perceptual bias or errors, providing insight into SQ3. However, no statistically significant differences were found in accuracy, response time, or uncertainty rates across conditions. Although the incongruent pitch condition resulted in slightly more errors and uncertainty, the differences were minimal and not statistically meaningful. These results suggest that while auditory cues may influence initial perception, they are often overridden by dominant visual input, particularly in tasks where visual speed estimation is relatively easy.

Collectively, these findings suggest that auditory cues alone may not reliably improve task performance in controlled VR tasks with clear visual information. However, the subtle patterns observed such as earlier responses and lower error variability in sound conditions indicate that auditory cues may support perception in ways that are not fully captured by traditional significance testing. These findings support the premise of SQ1 and SQ2, albeit with the caveat that their effects may be context-dependent or subject to individual variation.

6.2 Interview

The qualitative findings from participant interviews offer additional context for interpreting the statistical results. Participants noted that with the absence of sound information, they have to divert a lot more attention to visual. This may explain why differences between conditions were not more pronounced: when sound is unavailable or ineffective, users compensate for the lack of information with heightened visual focus. These observations suggest that auditory cues can serve as a supportive modality that reduces visual load, even if their effects on accuracy are not always statistically significant.

For the second task, the interviews highlighted perceptual biases in congruent and incongruent conditions. During the experiment, the congruent and incongruent condition was shuffled within a block. Therefore, even though the participants noticed a difference in the sound the balls make, they soon disregard it as the difference is not consistent. This led several of them to rely increasingly on visual cues instead, consciously ignoring the pitch when it appeared

unreliable. These responses indicate that while auditory cues can provide initial guidance, their influence is easily undermined when cue reliability is inconsistent, further reinforcing the notion that visual information dominates in multimodal motion perception unless it becomes unreliable or ambiguous.

These qualitative findings strengthen the interpretation of the quantitative data, particularly for SQ3. Although the results did not show a significant difference in accuracy, the interview responses suggest that participants were perceptually influenced by sound cues, especially in uncertain conditions. In such cases, sound served as a helpful secondary reference that could guide decision-making. However, when visual information was clear and unambiguous, auditory input was largely discounted, indicating that sound cues alone are not strong enough to override visual perception. This underscores an important insight: while auditory motion cues have the potential to influence perceptual judgments, their impact is highly context-dependent and largely limited to situations where visual information is degraded, ambiguous, or insufficient.

7 CONCLUSION

The present study explored how different auditory cues influence users' perception of motion in VR, with a specific focus on two fundamental tasks: time-to-contact estimation and speed discrimination. Across both experiments, we aimed to investigate whether specific sound cues can enhance motion perception, how these cues influence users' accuracy and response timing, and whether they introduce perceptual biases when conflicting with visual information. While the qualitative feedback from participants suggested that sound was helpful, the quantitative analysis did not yield a statistically significant improvement in performance across conditions. These findings suggest that although auditory motion cues are perceived as supportive, their measurable influence on task accuracy under the tested parameters remains limited.

Despite the lack of significant quantitative effects, this research provides valuable insights into the perceptual role of auditory cues in VR. The modest impact of audio may stem from several factors, including individual variability in auditory and visual processing abilities, and the difficulty of designing universally interpretable auditory cues. These limitations highlight that the design and tuning of auditory motion cues must be carefully considered in future studies. Moving forward, future work should refine the cue parameters and investigate adaptive audio systems that account for user-specific perceptual thresholds. Additionally, more valid VR task designs that effectively isolate audio influence, and a broader participant pool may reveal stronger effects.

In conclusion, while the present study does not definitively demonstrate performance gains from auditory motion cues in VR, it underscores their perceived utility, highlights key design considerations, and lays the groundwork for future work exploring how carefully crafted sound design can complement and enhance motion perception in immersive environments.

REFERENCES

[1] Jessica Sharon Putranto, Jonathan Heriyanto, Kenny, Said Achmad, and Aditya Kurniawan. Implementation of virtual reality technology for sports education and training: Systematic literature review. *Procedia Computer Science*, 216:293–300, 2023.

[2] Marileen M. T. E. Kouijzer, Hanneke Kip, Yvonne H. A. Bouman, and Saskia M. Kelders. Implementation of virtual reality in healthcare: a scoping review on the implementation process of virtual reality in various healthcare settings. *Implementation Science Communications*, 4(1), June 2023.

[3] Maged Soliman, Apostolos Pesyridis, Damon Dalaymani-Zad, Mohammed Gronfula, and Miltiadis Kourmpetis. The application of virtual reality in engineering education. *Applied Sciences*, 11(6):2879, March 2021.

[4] Andries van Dam, David H Laidlaw, and Rosemary Michelle Simpson. Experiments in immersive virtual reality for scientific visualization. *Computers amp; Graphics*, 26(4):535–555, August 2002.

[5] Stefania Serafin, Michele Geronazzo, Cumhur Erkut, Niels C. Nilsson, and Rolf Nordahl. Sonic interactions in virtual reality: State of the art, current challenges, and future directions. *IEEE Computer Graphics and Applications*, 38(2):31–43, March 2018.

[6] Kai-uwe Doerr, Holger Rademacher, Silke Huesgen, and Wolfgang Kubbat. Evaluation of a low-cost 3d sound system for immersive virtual reality training systems. *IEEE Transactions on Visualization and Computer Graphics*, 13(2):204–212, March 2007.

[7] Cesare V. Parise and Marc O. Ernst. Noise, multisensory integration, and previous response in perceptual disambiguation. *PLOS Computational Biology*, 13(7):e1005546, July 2017.

[8] L. Fialho, J. Oliveira, A. Filipe, and F. Luz. Soundspace vr: spatial navigation using sound in virtual reality. *Virtual Reality*, 27(1):397–405, November 2021.

[9] Simon Carlile and Johahn Leung. The perception of auditory motion. *Trends in Hearing*, 20, January 2016.

[10] Stefan Riedel, Matthias Frank, and Franz Zotter. Effect of hrtfs and head motion on auditory-visual localization in real and virtual studio environments. *Acta Acustica*, 9:21, 2025.

[11] Finnur Pind, Cheol-Ho Jeong, Hermes Sampedro Llopis, Kacper Kosikowski, and Jakob Strømann-Andersen. Acoustic virtual reality–methods and challenges. In *Proceedings of Baltic-Nordic Acoustic Meeting (BNAM), Reykjavík, Iceland*, 2018.

[12] Emil R. Hoeg, Lynda J. Gerry, Lui Thomsen, Niels C. Nilsson, and Stefania Serafin. Binaural sound reduces reaction time in a virtual reality search task. In *2017 IEEE 3rd VR Workshop on Sonic Interactions for Virtual Environments (SIVE)*, page 1–4. IEEE, March 2017.

[13] Gangsheng Zhang, Wei Wang, Jue Qu, Hengwei Li, Xincheng Song, and Qingli Wang. Perceptual influence of auditory pitch on motion speed. *Journal of Vision*, 21(10):11, September 2021.

[14] John G. Neuhoff. Perceptual bias for rising tones. *Nature*, 395(6698):123–124, September 1998.

[15] Irene Senna, Cesare V. Parise, and Marc O. Ernst. Modulation frequency as a cue for auditory speed perception. *Proceedings of the Royal Society B: Biological Sciences*, 284(1858):20170673, July 2017.

[16] John G. Neuhoff. Looming sounds are perceived as faster than receding sounds. *Cognitive Research: Principles and Implications*, 1(1), November 2016.

[17] Robert Sekuler, Allison B. Sekuler, and Renee Lau. Sound alters visual motion perception. *Nature*, 385(6614):308–308, January 1997.

[18] Thorben Hülsdünker, David Riedel, Hannes Käsbauer, Diemo Ruhnow, and Andreas Mierau. Auditory information accelerates the visuomotor reaction speed of elite badminton players in multisensory environments. *Frontiers in Human Neuroscience*, 15, November 2021.

[19] Martin Naef, Oliver Staadt, and Markus Gross. Spatialized audio rendering for immersive virtual environments. In *Proceedings of the ACM symposium on Virtual reality software and technology*, VRST02, page 65–72. ACM, November 2002.

[20] Thomas McKenzie, Damian T. Murphy, and Gavin Kearney. Interaural level difference optimization of binaural ambisonic rendering. *Applied Sciences*, 9(6):1226, March 2019.

[21] Giorgia Bertonati, Maria Bianca Amadeo, Claudio Campus, and Monica Gori. Auditory speed processing in sighted and blind individuals. *PLOS ONE*, 16(9):e0257676, September 2021.

[22] Natasa Paterson, Katsiaryna Naliuka, Soren Kristian Jensen, Tara Carrigy, Mads Haahr, and Fionnuala Conway. Design, implementation and evaluation of audio for a location aware augmented reality game. In *Proceedings of the 3rd International Conference on Fun and Games*, Fun and Games '10, page 149–156. ACM, September 2010.

[23] Gustavo Corrêa De Almeida, Vinicius Costa de Souza, Luiz Gonzaga Da Silveira Júnior, and Maurício Roberto Veronez. Spatial audio in virtual reality: A systematic review. In *Symposium on Virtual and Augmented Reality*, SVR '23, page 264–268. ACM, November 2023.

[24] Robert A. Lutfi and Wen Wang. Correlational analysis of acoustic cues for the discrimination of auditory motion. *The Journal of the Acoustical Society of America*, 106(2):919–928, August 1999.

A DIAGRAMS

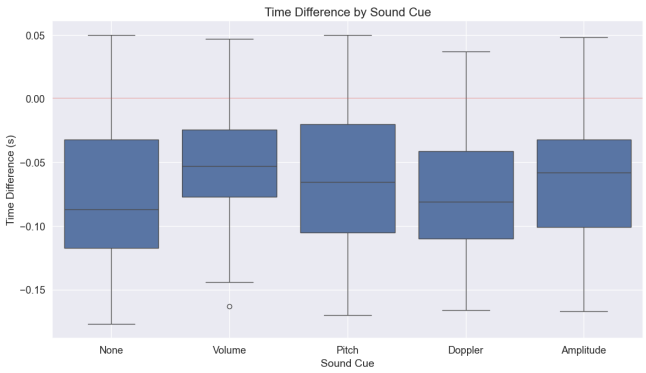

Fig. 4. From the first experiment: The box plot visualizes the distribution of time difference by different sound cues. Time difference are defined by the difference of actual TTC and the user response time.

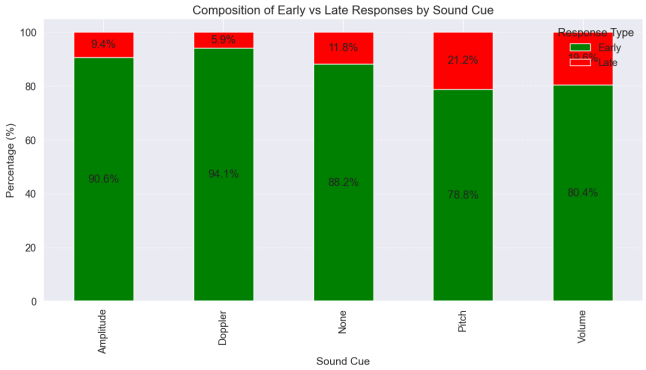

Fig. 5. From the first experiment: The bar chart visualizes the proportion between early and late time difference by different sound cues. Responses are considered early when the recorded time difference are negative, and late if it is positive.

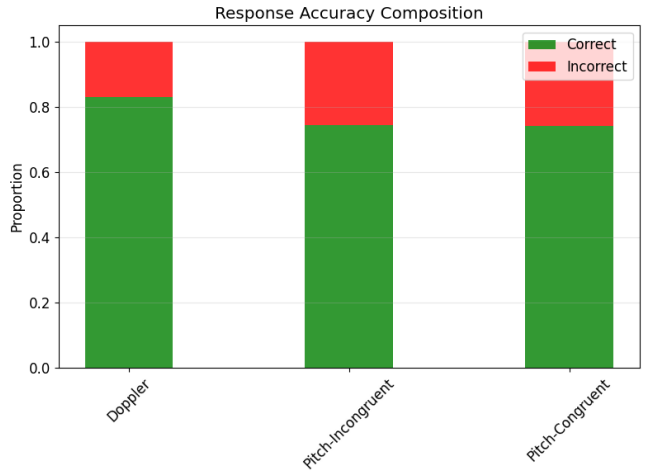

Fig. 6. From the second experiment: The bar chart visualizes the proportion between correct and incorrect responses. Not sure responses are not included in the proportion.

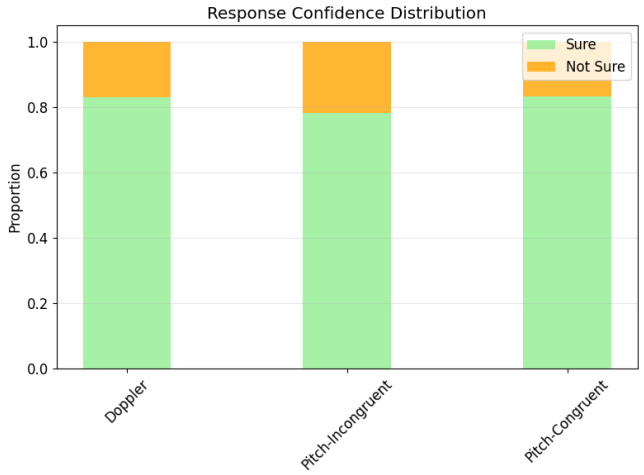

Fig. 7. From the second experiment: The bar chart visualizes the proportion between certain and uncertain responses. Uncertain responses are "Not sure" instances.

B AI STATEMENT

Throughout the research, I have used Generative AI (ChatGPT) in the process of writing to improve my language, tone and clarity. Additionally, it assisted me in LaTeX, Unity (game scripts), Python (data analysis/visualization). Any written information including paper structure, literature citations, methodology concepts, both quantitative and qualitative data mentioned in this paper originates from the author. The AI model did not, and is not allowed to gather and include additional information outside of the provided information.