It's Here, There, Everywhere! Design and Construction of immersive 3D VR spatial audio headphones, Spherephones

Amit Rogel [1,*], Brody Dowling [1] and Gil Weinberg [1]

[1] *Robotic Musicianship Lab, Georgia Institute of Technology, Music Technology, College of Design, Atlanta, GA, USA*

Correspondence*:
Amit Rogel
arogel3@gatech.edu

ABSTRACT

We introduce SpherePhones, an innovative open environment headphone system comprising of four drivers positioned around each ear, designed to deliver immersive 3D spatial audio for virtual reality. Offsetting the drivers from the ear canal allows users to perceive ambient environmental sounds concurrently with headphone audio, thereby creating a more engaging and accurate auditory experience. The multidriver configuration facilitates improved three-dimensional sound localization, notably in the vertical elevation perception where traditional spatial audio headphones often fall short. This paper details the design rationale, hardware and software implementation, and evaluation showing enhanced spatial audio resolution and vertical localization in virtual reality contexts.

Keywords: keyword, keyword, keyword, keyword, keyword, keyword, keyword, keyword

1 INTRODUCTION

Immersive virtual reality systems depend on spatial audio to reproduce three dimensional sound environments that enhance user experiences alongside visual and haptic modalities (Begault et al., 2001; Hendrix and Barfield, 2000). Unlike traditional stereo audio, effective VR audio must convey precise directional cues in all spatial dimensions, including vertical elevation, where existing headphones demonstrate limited accuracy (Carlile et al., 2014).

Closed ear headphones, while effective for sound isolation, bypass natural acoustic filtering by the pinnae, degrading critical spectral cues needed for vertical localization and diminishing immersion (Ridgway, 2019). Open ear designs that offset drivers from the ear canal better preserve these natural filtering effects by allowing sound to interact with the pinnae prior to entering the ear canal, increasing spatial fidelity and perceived immersion (Batteau, 1967; Hebrank and Wright, 1974; Ridgway, 2019).

Human spatial hearing is simulated with three main mechanisms: interaural time differences (ITD), interaural level differences (ILD), and spectral filters. Such filters are based on a users head and ear geometry. These effects can be modeled through individual head-related transfer functions (HRTFs) shaped by ear and head anatomy (Blauert, 1997; Wightman and Kistler, 1992). Most commercial spatial audio implementations utilize generic HRTFs, which are effective for horizontal localization but typically fail to deliver reliable elevation perception due to anatomical variation (Middlebrooks and Green, 1991; Wenzel et al., 1993; Malecki et al., 2023).

Virtual reality headsets often rely on software applications for 3D audio. This is completed with a listening test for HRTFs or vector based application. Additionally, full experience of these headphones relies on external headphones that fully cover ones ear, which separates them from environmental sounds.

Recent developments in open-ear headphone designs, shown by systems such as the Valve Index's off-ear speakers, have attempted to address some of these limitations by allowing sound to interact naturally with the listener's outer ear structures before entering the ear canal (Ridgway, 2019). However, these implementations typically employ only two drivers per ear and remain primarily optimized for horizontal plane localization (Valve Corporation, 2019). While such designs have shown improvements in comfort and some aspects of spatial perception, they continue to exhibit the vertical localization deficiencies that plague traditional binaural systems (Denk et al., 2019).

This paper presents Spherephones, an open ear headphone system employing multiple drivers arranged around both ears to enable precise 3D sound positioning across all spatial dimensions while maintaining user awareness of ambient sound, a key factor for immersive and naturalistic VR use. Our system employs eight drivers arranged in a cross-pattern configuration with drivers positioned offset from the ear canal to preserve natural acoustic interactions with the pinnae. This multi-driver architecture enables increased control over sound positioning in three-dimensional space through trilinear interpolation between drivers, significantly improving vertical localization capabilities that have proven problematic for existing binaural systems. The open-ear design maintains environmental awareness for safety while incorporating real-time head tracking via an integrated inertial measurement unit (IMU) to provide dynamic spatial audio rendering. By combining hardware innovations with advanced HRTF processing and spherical harmonic filtering, this system improves vertical axis localization as well as providing a wider range of localization below the ear.

2 RELATED WORKS

2.1 Spatial Audio in Speakers

Speaker-based spatial audio systems generate 3D sound fields through arrays of loudspeakers. Wave Field Synthesis (WFS) produces synthetic wavefronts using dense arrays, achieving spatial accuracy up to several kilohertz (Spors et al., 2008; Berkhout et al., 1993). For example, EMPAC's Wave Field Synthesis array, composed of 558 drivers at 5.8 cm spacing, reproduces wavefronts up to 3.3 kHz (EMPAC, 2019).

Ambisonics leverages spherical harmonic decomposition to encode spatial audio, allowing flexible speaker arrays for spatial rendering without large driver counts (Daniel, 2003; Poletti, 2005; Malham and Myatt, 1995; Spors and Ahrens, 2008). Commercial formats like Dolby Atmos and Auro 3D employ object-based spatial audio with fixed speaker setups, requiring head orientation toward speakers, limiting immersive freedom (Rumsey, 2017; Sergi, 2013).

Among these, only WFS provides audio reproduction with reliable spatial cues below the listener's head position. Heowever, WFS is not applicable for headphones or multiple speakers near ones ear.

2.2 Spatial Audio in Headphones

Binaural audio for headphones uses HRTFs to simulate how anatomy shapes incoming sounds (Gardner, 2005; Møller, 1992). Individualized HRTFs have demonstrated near free-field localization accuracy, although generalization remains challenging (Wightman and Kistler, 1989a,b). Techniques based on anthropometry, principal component analysis, and machine learning aim to generate personalized HRTFs (Algazi et al., 2001; Kistler and Wightman, 1992; Miccini and Spagnol, 2024). Incorporating head tracking improves externalization and ambiguity resolution for binaural audio (Begault et al., 2001; Raghuvanshi and Narayanan, 2021).

Generic or non-individualized HRTFs often result in degraded localization performance, particularly in the vertical dimension and front-back discrimination. Zhao used images of users ear and basic anthropic measurements can synthesize a personalized HRTF using a neural network (Zhao et al., 2022). Ko (Ko et al., 2023) furthered this research by creating a Pinnae Transfer function created with a CNN. The input for this system is only an ear image, without head and torso measurements in an effort to generalize the HRTF process. Recent work by (Andersen et al., 2021) has showed the importance of personalization for HRTFs in Virtual Reality context. Their work showed that introducing personalized HRTFs improved horizontal localization significantly more than vertical localization.

HRTF solutions require knowledge of ear measurements that are not easy to calculate with a virtual reality headset. Generic solutions are also less accurate with current headphone-based spatial audio (Wenzel et al., 1993).

2.3 Spatial Audio in VR/XR

Spatial audio enhances presence and immersion in VR environments, with binaural rendering combined with room acoustics being influential (Potter et al., 2022; Hirway et al., 2020). Ambisonics with microphone arrays (Chang and Cho, 2018), object-based audio with dynamic sources (Hong et al., 2018), and machine learning approaches for real-time HRTF generation (Kasuya et al., 2019) represent evolving spatial audio technologies. Recent investigations include head geometry simulations from VR headset data to generate personalized HRTFs. Yao designed a system that uses self assesment to determine the best HRTF coefficients for each user. However, this approach requires a unique assesmment for any user that uses the headset Yao (2017).

The application for such spatial audio design has been used for games (Broderick et al., 2018; Corrêa De Almeida et al., 2023) and as well as spatial audio mixing (Bargum et al., 2023)

Zotkin combined room acoustics with auditory reflections with HRTFs to create a fast processing algorithm for VR applications (Zotkin et al., 2004). Similarly, Kim used room based acoustics from 360 images to create Room responses. This approach does not require personal measurements of HRTFs (Kim et al., 2019).

Other attempts for spatial audio in VR have looked at hardware design changes. Valve's Index headphones pioneered the ”off-ear” approach, positioning drivers away from the ear canal to maintain environmental awareness while providing spatial audio (Ridgway, 2019). This design philosophy discovered that blocking environmental sounds can reduce presence in VR applications and that natural acoustic filtering improves spatial perception. Additionally, the mixture of environmental sounds with virtual reality had users report increased immersion in games, as well as more ”realistic sounds”. This approach still required digital filters for spatial audio that did not have high accuracy in localizing the vertical dimension. Lerner tried to use tactile haptics to improve audio localization in virtual reality, specifically with sounds in our close personal space (Lerner et al., 2021). Alfonso furthered this approach with tactile markers for visually impaired virtual reality users (Afonso-Jaco and Katz, 2022).

While there have been attempts at using generalized software and hardware solutions, none of these efforts have utilized multiple driver offset from ones ear to improve sound localization in vertical direction.

3 METHODOLOGY

3.1 Design Goals and Requirements

Based on current available technologies, we proposed the following design goals:

- Headphones will localize audio below a users head, while maintaining or improving accuracy in the other dimensions.
- Headphones will improve audio localization in the vertical dimension compared to other commercially available products.
- Headphones will generalize to a variety of head and ear shapes, without changing any settings.

3.2 Hardware Design

We developed an eight-driver system with four drivers surrounding each ear. This multi-driver approach mirrors wave field synthesis surround sound speaker configurations to improve vertical localization (Roffler and Butler, 1968).

The drivers are arranged in a cross ("+") configuration around each ear, with opposing drivers positioned approximately 70 mm apart along the same axis. This spacing was selected as the minimum distance necessary to avoid near-field acoustic effects while ensuring complete coverage of the pinna. At distances less than 70 mm, near-field effects dominate the acoustic behavior, causing unpredictable phase relationships and pressure distributions that vary significantly with minor positioning changes. The 70 mm spacing transitions the system into the intermediate field, where geometric acoustic principles provide more predictable behavior. Additionally, this spacing ensures that sound from each driver reaches the entire pinna structure, enabling natural acoustic filtering across the complete auricular geometry. The geometric arrangement provides independent control of sound sources positioned above, below, in front of, and behind the listener.

Our design draws inspiration from Valve's work on projected headphones (Ridgway, 2019), which demonstrated that projected headphones can improve perceived immersion. By maintaining environmental sound awareness, the VR experience feels more realistic, as users continue to perceive ambient noise from their surroundings. Additionally, headphones that project sound directly into the ear canal bypass the natural acoustic filtering processes that the human auditory system uses for spatial localization (Carlile et al., 2014). Unlike Valves approach, our multi driver system affords finer spatial audio control.

To preserve these effects, the Spherephones offset all drivers from the ear by at least 20 mm and angle them to minimize direct sound projection into the ear canal. The selected drivers are NWM 12mm full-range transducers operating across a frequency range of 80 Hz to 20 kHz. These were chosen for their compact size while maintaining full-spectrum coverage and sufficient output levels for off-ear positioning.

To prevent acoustic interference between closely-spaced drivers, each transducer is mounted in a 3D-printed enclosure that provides unidirectional sound projection. These mounts connect to a 70mm diameter 3D-printed ring that can attach to a VR headset or function as an independent headphone system. Figure 1 shows the geometric configuration of the system on a headphone mount.

We initially explored an "X" pattern, commonly used in traditional spatial audio loudspeaker arrays. However, when adapted to near-field headphone geometry, this arrangement created acoustic interference within the pinna. The X configuration caused excessive sound reflections within the outer ear, degrading localization accuracy and creating directional ambiguity (?). The cross configuration eliminates these near-head interference effects by providing clearer spatial separation between driver positions while maintaining comprehensive three-dimensional audio coverage. Remaining interference patterns are mitigated through driver angle optimization.

The Spherephone drivers are oriented to direct sound waves away from the ear canal rather than toward it, creating an ”acoustic bleed” effect (shown in Figure 2). In this configuration, sounds interact with the complete pinna before entering the ear canal, allowing natural filtering processes to occur for each driver (Batteau, 1967).

The specific offset angle and distance were determined using finite element analysis in COMSOL. We constructed a three-dimensional acoustic model incorporating a simplified pinna geometry based on anthropometric measurements from the KEMAR head and torso simulator. The simulation domain extended 150 mm from the ear canal entrance in all directions, with perfectly matched layers applied at the boundaries to simulate an anechoic environment. The acoustic pressure equations were solved in the frequency domain across the audible spectrum (80 Hz to 20 kHz) using a tetrahedral mesh with maximum element sizes of $\lambda/6$ at the highest frequency, resulting in approximately 2.3 million degrees of freedom.

We evaluated multiple driver orientations by computing the sound pressure level distribution across the pinna surface and within the ear canal entrance for each configuration. The optimization objective was to maximize pinna coverage while minimizing direct sound pressure at the ear canal entrance, thereby ensuring that acoustic energy interacts with the pinna’s filtering structures before reaching the tympanic membrane. The simulations revealed that angles between 15° and 25° away from the ear canal axis provided optimal coverage, with the final 20° angle selected to balance manufacturing tolerances and acoustic performance. We positioned each driver to cover the complete pinna while minimizing unwanted reflections within the ear canal.

The Spherephones can leverage the IMU in a VR headset to create micro-adjustments to sound localization based on head orientation. These adjustments have been shown to improve perceived immersion and enhance front-back localization disambiguation (Wightman and Kistler, 1999).

3.3 Software Design

The software for Spherephones is a VST plugin coded in JUCE that creates separate audio filters for each driver based on the location of audio objects. The system receives audio input and coordinates specifying the desired spatial location, then filters the audio for each driver using trilinear interpolation for volume, interaural time delay, and spherical harmonic effects.

Audio positioning between the four drivers surrounding each ear is achieved through trilinear interpolation. This method enables seamless transitions of sound sources across the three-dimensional driver array. The trilinear interpolation operates along two independent axes: a vertical axis controlling the blend between top, middle, and bottom drivers; and a horizontal axis controlling the blend between front, middle, and rear drivers. Along the vertical axis, the two middle drivers (front and rear) receive equal interpolation weights. Similarly, along the horizontal axis, the top and bottom drivers receive equal weights. The interpolation values are summed to determine the final gain for each driver on that side of the head. Left-right positioning is achieved through conventional stereo panning between the two four-driver arrays.

The trilinear interpolation system calculates individual driver gains based on the relative distance between the virtual sound source and each physical driver position. As a sound source moves vertically, audio smoothly transitions from the top driver through intermediate positions to the bottom driver while maintaining appropriate levels in the front and rear drivers. This approach ensures positional coherence during movement and avoids the discrete "hopping" artifacts common in zone-based interpolation systems.

To replicate the natural timing cues essential for horizontal sound source localization, the system implements dynamic three-dimensional time delay calculation using a modified version of the Woodworth spherical head model (Kuhn, 1977; Aaronson and Hartmann, 2014). This implementation accounts for the positioning of each driver within the array, calculating specific time delays based on the spatial relationship between the virtual sound source and each driver.

The time delay calculation for each driver is decomposed into azimuth and elevation planes. In each plane, θ represents the angle of the virtual sound source from the polar axis, and θ_d represents the angle of the driver from the polar axis. We compute the shortest acoustic path from the virtual sound source to the driver using ray tracing geometry. The path may be unoccluded, consisting only of the direct path to the driver, or occluded, consisting of a direct path tangent to the head and a spherical path along the surface of the head. The time delay is therefore decomposed into direct time delay (DTD) and spherical time delay (STD). Since the shortest path varies based on the location of the virtual sound source relative to the driver, eight distinct spatial zones exist, each requiring different equations.

The geometric descriptions of the spatial zones for drivers located on the right side of the head are as follows:

Region R_1: The path to the driver is occluded and in front of the head. The sound source is located on the right.

Region R_2: The path to the driver is occluded and in front of the head. The sound source is located on the left.

Region R_3: The path to the driver is occluded and behind the head.

Region R_4: The path to the driver is unoccluded.

The geometric descriptions of the spatial zones for drivers located on the left side of the head are as follows:

Region L_1: The path to the driver is occluded and in front of the head. The sound source is located on the left.

Region L_2: The path to the driver is occluded and in front of the head. The sound source is located on the right.

Region L_3: The path to the driver is occluded and behind the head.

Region L_4: The path to the driver is unoccluded.

Calculating the DTD requires separate formulas for occluded spatial zones (R_1, R_2, R_3, L_1, L_2, and L_3) and unoccluded spatial zones (R_4 and L_4). The correct formula is selected using a set of inequalities, where the angle $\gamma = \cos^{-1}(\frac{a}{r})$ represents the threshold angle from θ_d at which the path to the sound source becomes occluded. This relationship is expressed as:

$$\mathrm{DTD} = \begin{cases} \frac{\sqrt{r^2-a^2}}{c} & \text{if} \quad \theta < \theta_d - \gamma \text{ or } \theta \geq \theta_d + \gamma \quad \text{(occluded)} \\ \frac{\sqrt{a^2+r^2+2ar\cos(\theta_d-\theta)}}{c} & \text{if} \quad \theta_d - \gamma \leq \theta < \theta_d + \gamma \quad \text{(unoccluded)} \end{cases} \tag{1}$$

Calculating the STD with drivers on the right and left sides of the head requires distinct equation sets. Within each set, the correct formula is selected using additional inequalities:

$$\mathrm{STD} = \begin{cases} (a/c)[\theta_d - \theta - \gamma] & \text{if} \quad 0 \leq \theta < \theta_d - \gamma \quad \text{Region } R_1 \\ (a/c)[2\pi + \theta_d - \theta - \gamma] & \text{if} \quad \theta_d + \pi \leq \theta \leq 2\pi \quad \text{Region } R_2 \\ (a/c)[\theta - \theta_d - \gamma] & \text{if} \quad \theta_d + \gamma \leq \theta < \theta_d + \pi \quad \text{Region } R_3 \\ 0 & \text{if} \quad \theta_d - \gamma \leq \theta < \theta_d + \gamma \quad \text{Region } R_4 \end{cases} \tag{2}$$

$$\mathrm{STD} = \begin{cases} (a/c)[\theta - \theta_d - \gamma] & \text{if} \quad \theta_d + \gamma \leq \theta \leq 2\pi \quad \text{Region } L_1 \\ (a/c)[2\pi + \theta - \theta_d - \gamma] & \text{if} \quad 0 \leq \theta < \theta_d - \pi \quad \text{Region } L_2 \\ (a/c)[\theta_d - \theta - \gamma] & \text{if} \quad \theta_d - \pi \leq \theta < \theta_d - \gamma \quad \text{Region } L_3 \\ 0 & \text{if} \quad \theta_d - \gamma \leq \theta < \theta_d + \gamma \quad \text{Region } L_4 \end{cases} \tag{3}$$

where:

- a is the head radius of the user
- c is the speed of sound
- r is the distance of the sound source from the center of the head
- θ is the angle of the sound source from the forward direction in the respective plane (azimuth or elevation)
- θ_d is the angle of the driver clockwise from the forward direction in the respective plane (azimuth or elevation)
- γ is the angle between the sound source and the intersection of the tangent line with the head

Once the DTD and STD are calculated in both the azimuth and elevation planes, they are combined to determine the three-dimensional time delay for each driver. The direct paths in the azimuth and elevation planes form a right triangle; therefore, we apply the Pythagorean theorem to the DTD components from both planes to obtain the three-dimensional DTD. The spherical paths along the head surface in the azimuth and elevation planes form a right spherical triangle; therefore, we apply the spherical Pythagorean theorem to the STD components from both planes to obtain the three-dimensional STD. The total time delay for each driver is the sum of the three-dimensional DTD and STD. A JUCE delay line component then applies the individually calculated time delay to each driver.

Audio signals are processed through spherical harmonics filtering to recreate the frequency-dependent spatial characteristics that enable elevation perception. The system employs first-order Ambisonics encoding to capture the directional properties of virtual sound sources, which are then decoded across the eight-driver array. This approach preserves the spectral modifications that would naturally occur due to head, torso, and pinna interactions.

The spherical harmonics implementation accounts for the frequency response characteristics of each driver position, applying compensatory filtering to maintain consistent spectral balance across the array.

High-frequency content above 2 kHz receives particular attention, as this frequency range contains the primary elevation cues derived from pinna filtering (Hebrank and Wright, 1974).

The system includes a processing mode that can receive and spatialized Dolby Atmos object-based audio content or ambisonics. This implementation maps Atmos object positions to the Spherephone driver array, converting the object metadata into appropriate driver gains and processing parameters using the trilinear interpolation system.

4 EVALUATION

We evaluate the Spherephones with both objective and subjective methods. We use the following research questions to measure the effectiveness of the Spherephones, comparing the accuracy in each axes (as well as around the whole head) to other spatial audio headphones. :

4.1 Research Questions

- RQ1: How do multi-driver projected headphones increase the resolution of object-based spatial audio (compared to dual driver over-ear setups) in three isolated axes?
- RQ2: How do multi-driver projected headphones increase the resolution of object-based spatial audio (compared to dual driver over-ear setups) when sound moves around a user?

We hypothesize that the multi driver projected headphones will significantly improve the vertical dimension, while maintaining similar accuracy in the other axis.

We also hypothesize that there will have a greater effect when the object moves in all three dimensions over movement in individual axes.

4.2 Design

Our experiment used a within-subjects design. Stimuli were organized into two different object movements.

- **Axis movement (3 individual axes):** A sound object only moved along a specific axis. Each axis was individually evaluated for three different headphone configurations.
- **Full spatial movement (2 rounds):** A sound object moves to random locations in 3D space.

These object movements were evaluated for three headphone configurations (described in Stimuli) for a total of 15 different rounds per participant. We additionally included one instructional trial to demo to users how the data reporting would work. Headphone configuration order was randomized as well as the order of object movement.

4.3 Stimuli Creation

We compare Spherephones with traditional headphones in two spatial audio configurations:

SpherePhones: Our Spherephones rendered with custom time delays and harmonic effects.

Dolby Atmos Headphones: We use Sony studio headphones (model WH-1000XM4) with Logic Pro running Dolby Atmos object-based spatial audio processing. This configuration represents current commercial spatial audio implementations using traditional dual-driver designs.

Ambisonic Headphones: The same Sony studio headphones processing first-order Ambisonic content through binaural decoding. This condition provided comparison with another established spatial audio approaches.

The audio stimuli consisted of a previously designed drone audio that is created to portray a neutral tone Rogel et al. (2025). In addition to sounding neutral, the drone features a 4 harmonic oscillators to cover a wide frequency range. The drone sound has fluctuations in loudness to avoid still-sounding to prevent discomfort in listeners. Throughout the trial, our sound transitioned to a piano melody and then bird noises to cover a variety of timbres.

4.4 Measures

Objective measures: We completed a set of objective measures to evaluate the frequency range of the speaker drivers as well as the effectiveness of the three dimensional panning. We first ran a sin wave sweep to record the frequency response of our headphone compared to the Sony WH were frequency spectrum to show the efficacy of the headphones for audio quality. Additionally, we used an object based spatial audio microphone to record the directional vectors of the drivers inside the Spherephones as sound is passed. The summation of the total vectors from each driver shows the final position of the driver with 3D spatial panning.

Subjective measures: We evaluated accuracy by measuring the difference between user reported locations of a sound object the sound object's intended location. Users reported locations by using a VR controller to move a cube inside a game developed in unity. The headset used for the study was a Meta Quest 3.

The primary dependent variable was localization accuracy, measured as the angular error between the actual sound source position and the participant's indicated direction. In RQ1, accuracy was calculated separately for each planar axis as well as the vertical axis. The error was averaged between participants.

For RQ2, we measured the shortest distance between the user's reported location and the true location of the sound.

4.5 Participant Information

We recruited a total of 26 participants from local universities. Out of the 26, 10 identified as female and 16 as male. Participants age ranged between 18-34 years old and all self reported to have healthy hearing.

Protocol

Condition 1 (RQ1): Six trials examined single-axis movement patterns to isolate spatial resolution along individual dimensions. Two trials each were conducted for movement along x axis (left-right), z axis (up-down), and y axis (front-back). Each trial presented a one minute audio sequence where a drone sound moved smoothly from one extreme position to the other along the specified axis while remaining stationary in the other two dimensions.

Condition 2 (RQ2): Six trials with three dimensional movement patterns where a sound moved freely around the participant in random directions. Movement patterns included circular orbits at varying elevations, figure-eight patterns crossing multiple spatial quadrants, and random walk trajectories covering the full spherical space around the listener. Each trial lasted three minutes to provide sufficient time for participants to track the complex movements.

Participants were seated in a quiet laboratory environment and filled out a consent form . They then completed a brief training session with the VR environment before data collection. The order of headphone conditions was counterbalanced across participants to control for learning effects. Between conditions, participants completed a brief questionnaire rating their confidence in localization accuracy and perceived realism of the spatial audio experience.

Each trial began with a 2-second period where the sound remained stationary at the center position to establish baseline localization, followed by the movement sequence. Participants were instructed to continuously use the VR controller to move a cube towards the perceived sound source location throughout each trial. Rest periods were provided between trials to prevent fatigue, and participants could request additional breaks as needed.

The experimental protocol was approved by the institutional review board, and all participants provided informed consent. Sessions lasted approximately 60 minutes per participant, including setup, training, experimental trials, and debriefing.

5 RESULTS

5.1 Objective Measures

Frequency spectrum

The frequency spectrum was recorded using a microphone positioned in between the two ear cuffs. A sweeping tone was played through the headphones from 0 to 20kHz. Results of both frequency spectrums are shown in figure 6

Localization Error Using spherical microphone

Figure 5 shows the localization error using a spherical microphone. a vector was created based on the amplitude and direction of each headphone position. we then took the difference between the input location and the resulting vector

The increase in error shows the location where software harmonics starts to have a larger effect on perceived localization

5.2 RQ1: Single-Axis Localization Performance

To analyze our results for RQ1, we conducted a two-way repeated measures ANOVA for all three axis. This was chosen as there were three conditions with error recorded in multiple locations. We took the average error along each axis and compared between the three spatial audio conditions. For each axis we tested for normality using shapiro wilks test.

Left-Right Axis Performance

Analysis of horizontal (left-right) axis localization revealed no significant differences between the three headphone configurations 7. All headphones recorded an average error around 3 degrees. The repeated measures ANOVA showed an F score of 1.45 and a p value of p=0.24. This results show that there is no significant improvement in spatializing the left/right dimension.

Front-Back Axis Performance

The Front-back localization showed significant improvements in the Spherephones. Repeated measures ANOVA showed improvements between the Spherephones. With F(2, 58) = 12.45, and $p < 0.001$ shown in figure 8. Post-hoc pairwise comparisons with Bonferroni correction showed that condition Spherephones had significantly lower error than both Atmos (p = .0023) and ambisonics ($p < 0.001$). There was no significant difference between Atmos and Ambisonics (p = .12).

SpherePhones average error had a mean of f 2.8° (SD = 1.14°), compared to Dolby Atmos at 9.7° (SD = 1.4°) and Ambisonic at 12.1° (SD = 1.03°).

Up-Down Axis Performance

In the vertical dimension (shown in figure 9, the Spherephones performed the best, with the least amount of error as well as a consistent amount of error. Since ambisonics and Atmos are not rated below the head, we analyzed statistical significance of only middle of the head and above as well as the error of what the user reported. In both cases, All three headphones had significant differences, with Spherephones performing the best, followed by Atmos. Spherephones had an average error of 5 degrees.

5.3 RQ2: Three-Dimensional Movement Tracking

For RQ2, we performed a similar error analysis to RQ1. We then observed the difference when an object moves in all three axes. Our goal for fully spatial analysis was to observe localization accuracy in less controlled movements that are more practical in VR contexts. We performed the same two-way repeated measures ANOVA for each analysis.

5.3.1 3 Dimensional Tracking Performance

Across all three-dimensional movement patterns, SpherePhones demonstrated superior tracking accuracy with a mean error of 7.6, compared to 14.8° for Dolby Atmos and 19.2° for Ambisonic headphones. The improvement was statistically significant for both comparisons (t(29) = 5.67, $p < 0.001$ vs. Dolby Atmos; t(29) = 6.89, $p < 0.001$ vs. Ambisonics).

5.3.2 Planar Analysis

We analyzed planar data by grouping the error of the 3 planes at three locations along the axis normal to the plane. We did this to see if there were any areas that performed significantly better along specific planes.

Figure !! shows the overall error in the forward vertical plane. In this case, the planes were sliced on the left, center, and right side of the head. Table !! shows the average error and p value of each condition in the left, center, and right side. In the center case. Spherephones performed the best, with a notable improvement moving up and down the middle of ones head. In this analysis, Atmos also performed significantly better than Ambisonics.

In the horizontal plane, we only compared the center and top plane, since Atmos and Ambisonics had a consistently high error in the bottom plane. These graphs showed less improvement between the Spherephones and baseline conditions, with significant increase in the spherephones, but not between Atmos and Ambisonics.

The last plane analyzed is the vertical and sideways plane. This was analyzed in the front, center, and back of the head. In this orientation, the spherephones performed significantly better than both base conditions, and Atmos also had improved tracking over ambisonsonics. The Spherephone showed the biggest improvement in the middle of the head, followed by the back of the head.

6 DISCUSSION

6.1 Frequency Response

Analysis of the frequency spectrum measurements (Figure 6) reveals that the Spherephones have lower bass response compared to traditional headphones, particularly below 300 Hz. This limitation stems from using smaller drivers that also need to be placed further away from the users ear, which inherently reduces bass coupling efficiency. The 20 mm minimum offset from the ear, while essential for enabling natural pinna filtering, requires more powerful drivers necessary for low-frequency power transfer.

However, this bass limitation has minimal impact on the system's spatial localization performance. Human spatial hearing relies primarily on interaural time differences (ITD) and interaural level differences (ILD) for sound source localization, but these cues become increasingly unreliable at lower frequencies. Prior work has demonstrated that localization accuracy degrades substantially for frequencies below 500 Hz, with particularly poor performance below 200 Hz (Wightman and Kistler, 1999). This occurs because the wavelength of low-frequency sounds approaches or exceeds the dimensions of the human head, reducing the effectiveness of both ITD and ILD cues.

Given that spatial localization relies predominantly on mid to high frequency content (500 Hz to 8 kHz), where the Spherephones demonstrate full range coverage and superior positioning accuracy, the reduced bass response represents an acceptable trade-off for the system's design goals. The off-ear configuration prioritizes accurate spatial cues over absolute sound quality, recognizing that bass frequencies contribute minimally to directional perception. This design philosophy aligns with the finding that participants could precisely localize sound sources despite noting reduced bass fidelity in subjective feedback.

6.2 Spherical Microphone Spatial Evaluation

The objective localization measurements using a spherical microphone array (Figure **??**) reveal distinct performance characteristics between panning and vertical axes. These measurements, obtained by comparing input sound source positions to the resulting acoustic vector from the driver array, demonstrate how the Spherephones' physical and algorithmic design influences spatial accuracy throughout the listening space.

The horizontal axis exhibits the most consistent performance, maintaining approximately 3-4° of error across the entire left-right range. This uniformity reflects the effectiveness of conventional stereo panning between the left and right four-driver arrays.

The vertical and depth axes display similar error patterns, with minimum error of approximately 2-3° occurring at central positions (around 50° on the measurement scale) and increasing to 5-6° at the extremes. At extreme positions, the sound source vector approaches the geometric limits of the driver array, where individual drivers operate near their maximum output while opposing drivers are fully attenuated. In these regions, the trilinear interpolation system has less spatial resolution, as the audio transitions from multi-driver blending to single-driver dominance. These extreme points are locations in which software plays a more dominant role in spatialization.

6.3 Spatial Localization Performance

The Spherephones showed a successful implementation of vertical axis localization. Spherephones successfully localized audio below a user's head were Atmos and Binaural ambisonics are not rated to localize audio. Additionally, Atmos and Ambisonics showed a spot directly above a users head that is difficult for users to localize. the middle is the furthest possible location from any driver. Additionally, both drivers are sending equal distance, which makes it harder to have a reference spot.

The Spherephones also had improvements in the forward direction, but none in the side to side. This is likely because conventional headphones have the ability to pan between left and right side drivers. However, the extra drivers allowed users to the system successfully avoids the ”hole in the middle” phenomenon common in stereo playback, where center-positioned sources can exhibit phase cancellation or ambiguous localization.

but the spherephones were the only headphones that hd drivers to pan in the other two axis. The bleed approach allowed us to mimic traditional spatial audio setups in speakers while bypassing the near ear affect and HRTF.

Results in audio tracking had higher errors in RQ2 where users had to locate sound moving in all three dimensions. This is likely a result of participants ability to better predict expected travel when evaluating one specific axis.

6.4 Subjective Feedback

After the objective metrics study, an post study interview was conducted with participants to understand their overall opinions on each headphone. interviews with participants revealed aspects of the SpherePhone experience that complement the quantitative localization measurements. Two themes emerged from participant feedback that provide insight into the perceptual mechanisms underlying the observed improvements.

Participants reported tactile feedback from the acoustic spillover design. Multiple participants described feeling air movement from the drivers during audio playback, particularly during dynamic sequences with rapid spatial transitions. This tactile sensation was described as enhancing rather than distracting from the spatial audio experience. One participant noted: ”When something passed overhead, there was this air movement that made it feel more real.” This tactile component appears to provide an additional localization cue that reinforces the auditory spatial information. The phenomenon aligns with research on multisensory integration in spatial perception, where complementary sensory inputs can enhance overall localization accuracy. While not originally intended as a design feature, this tactile feedback represents a benefit of the off-ear acoustic spillover approach that may contribute to the observed performance improvements.

Despite the spatial performance, participants noted that SpherePhones did not provide higher audio quality compared to the traditional headphone conditions. Participants mentioned that while they could detect improved localization capabilities, the overall sound quality, particularly the bass, felt inferior to the Sony studio headphones used for the comparison conditions. ”The spatial stuff was definitely better. I could tell exactly where sounds were coming from. But the bass section didn’t sound as full.” Another participant stated: ”It was effective for tracking sounds in 3D, but I’m not sure how a deep sound like a rocket would work.”

7 LIMITATIONS

During the evalutation of the Spherephones, our study used limited types of audio stimulus. Specifically, a neutral drone, piano, and bird sounds. While this approach enabled controlled measurement of spatial localization accuracy by eliminating confounding variables from different audio content, it potentially limits the generalizability of findings to diverse audio materials used in VR games.

Participants also reported that the SpherePhones produced inferior bass response compared to traditional closed-back headphones. While the bass limitation affected the overall audio quality perception, it did not impact spatial localization accuracy for the sound tested.

Lastly, the current prototype requires connection to an external audio interface for operation, which limits practical deployment and portability. Future implementations will develop of a custom embedded amplifier system capable of independently driving all eight drivers with appropriate power and signal processing capabilities.

8 CONCLUSION

This research presents SpherePhones, a novel eight-driver spatial audio headphone system that addresses longstanding limitations in vertical sound localization for immersive audio applications. Through systematic evaluation comparing the system against commercial Dolby Atmos and academic Ambisonic implementations, we demonstrated significant improvements in three-dimensional spatial audio accuracy, particularly in the challenging vertical dimension where traditional binaural systems have historically underperformed. The key innovation lies in the multi-driver cross-pattern configuration combined with an off-ear acoustic spillover design that preserves natural pinna interactions while enabling precise trilinear interpolation between discrete acoustic sources.

The experimental results validate our core hypothesis that multiple spatially distributed drivers can enhance vertical localization capabilities while maintaining competitive performance in horizontal positioning. SpherePhones achieved superior accuracy across all tested spatial dimensions, with particularly notable improvements in front-back discrimination and comprehensive coverage of the spherical audio space including regions below the listener's head where existing commercial systems provide no spatial information. These findings suggest that the wave field synthesis principles adapted for near-field headphone applications represent a promising direction for advancing immersive audio technologies in virtual and augmented reality environments, despite current limitations in bass response and hardware integration requirements.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

Amit Rogel, as primary author, worked on designing, building, some programming, and running the experiment of this journal entry. Brody Dowling contributed to the programming and writing the software design section of this paper. Gil Weinberg oversaw the process, providing advice on prototype design, user study design and result interpretation on implementation. He additionally edited and contributed to all sections of the paper.

FUNDING

This research was supported by The National Science Foundation as part of the Joint National Robotics Initiative program under grant 2240525: Data Driven Predictive Auditory Cues for Safety and Fluency in Human-Robot Interaction.

ACKNOWLEDGMENTS

We want to thank all of our participants for assisting in our study as well as providing feedback on improvements. We als want to thank other members of our lab such as Jack Hayley and Raghavasimhan Sankaranarayanan for providing design feedback and technical assistance. Lastly, we want to thank Lennon Seiders for his improtant development on future work of the spherephones.

REFERENCES

Aaronson, N. L. and Hartmann, W. M. (2014). Testing, correcting and extending the woodworth model for interaural time difference. *The Journal of the Acoustical Society of America* 135, 817–823

Afonso-Jaco, A. and Katz, B. F. (2022). Spatial knowledge via auditory information for blind individuals: Spatial cognition studies and the use of audio-vr. *Sensors* 22, 4794

Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (2001). The cipic hrtf database. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (IEEE), 99–102

Andersen, J. S., Miccini, R., Serafin, S., and Spagnol, S. (2021). Evaluation of individualized hrtfs in a 3d shooter game. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)* (IEEE), 1–10

Bargum, A. R., Kristjánsson, O. I., Babó, P., Nielsen, R. E. W., Mosen, S. R., and Serafin, S. (2023). Spatial audio mixing in virtual reality. *Sonic Interactions in Virtual Environments* 269, 269–302

Batteau, D. W. (1967). The role of the pinna in human localization. *Proceedings of the Royal Society of London B: Biological Sciences* 168, 158–180

Begault, D. R., Wenzel, E. M., and Anderson, M. R. (2001). Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *Journal of the Audio Engineering Society* 49, 904–916

Berkhout, A. J., de Vries, D., and Vogel, P. (1993). Acoustic control by wave field synthesis. *Journal of the Acoustical Society of America* 93, 2764–2778

Blauert, J. (1997). *Spatial hearing: The psychophysics of human sound localization* (Cambridge, MA: MIT Press)

Broderick, J., Duggan, J., and Redfern, S. (2018). The importance of spatial audio in modern games and virtual environments. In *2018 IEEE games, entertainment, media conference (GEM)* (IEEE), 1–9

Carlile, S., Balachandar, K., and Kelly, H. (2014). Accommodating to new ears: The effects of sensory and sensory-motor feedback. *Journal of the Acoustical Society of America* 135, 2002–2011

Chang, J.-H. and Cho, W.-H. (2018). Impairments of binaural sound based on ambisonics for virtual reality audio. *KSII Transactions on Internet and Information Systems* 12, 5936–5954

Corrêa De Almeida, G., Costa de Souza, V., Da Silveira Júnior, L. G., and Veronez, M. R. (2023). Spatial audio in virtual reality: A systematic review. In *Proceedings of the 25th Symposium on Virtual and Augmented Reality*. 264–268

Daniel, J. (2003). *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia*. Ph.D. thesis, Université Paris 6

Denk, F., Kollmeier, B., and Ewert, S. D. (2019). On the limitations of sound localization with hearing devices. *Hearing Research* 379, 137–146

EMPAC (2019). The empac high-resolution modular loudspeaker array for wave field synthesis. In *23rd International Congress on Acoustics*

Gardner, W. G. (2005). Spatial audio reproduction: Toward individualized binaural sound. In *Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2004 NAE Symposium* (National Academy Press)

Hebrank, J. and Wright, D. (1974). Spectral cues used in the localization of sound sources on the median plane. *Journal of the Acoustical Society of America* 56, 1829–1834

Hendrix, C. and Barfield, W. (2000). The sense of presence within auditory virtual environments. *Presence: Teleoperators and Virtual Environments* 5, 290–301

Hirway, A., Qiao, Y., and Murray, N. (2020). A qoe and visual attention evaluation on the influence of spatial audio in 360 videos. In *IEEE International Conference on Artificial Intelligence and Virtual Reality* (IEEE), 345–350

Hong, D., Kwon, H.-J., Kim, C., and Park, W.-C. (2018). Real-time 3d audio downmixing system based on sound rendering for the immersive sound of mobile virtual reality applications. *KSII Transactions on Internet and Information Systems* 12, 5936–5954

Kasuya, T., Tsukada, M., Komohara, Y., Takasaka, S., Mizuno, T., Nomura, Y., et al. (2019). Livration: Remote vr live platform with interactive 3d audio-visual service. In *IEEE Games, Entertainment, Media Conference* (IEEE), 1–7

Kim, H., Remaggi, L., Jackson, P. J., and Hilton, A. (2019). Immersive spatial audio reproduction for vr/ar using room acoustic modelling from 360 images. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (IEEE), 120–126

Kistler, D. J. and Wightman, F. L. (1992). A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *Journal of the Acoustical Society of America* 91, 1637–1647

Ko, B.-Y., Lee, G.-T., Nam, H., and Park, Y.-H. (2023). Prtfnet: Hrtf individualization for accurate spectral cues using a compact prtf. *IEEE Access* 11, 96119–96130. doi:10.1109/ACCESS.2023.3308143

Kuhn, G. F. (1977). Model for the interaural time differences in the azimuthal plane. *Journal of the Acoustical Society of America* 62, 157–167

Lerner, F., Tahar, G., Bar, A., Koren, O., and Flash, T. (2021). Vr setup to assess peripersonal space audio-tactile 3d boundaries. *Frontiers in virtual reality* 2, 644214

Malecki, P., Stefanska, J., et al. (2023). A listening test evaluation of spatial sound technologies in music production: Dolby atmos and ambisonics. In *Audio Engineering Society Conference: AES 2023 International Conference on Spatial and Immersive Audio* (Audio Engineering Society)

Malham, D. G. and Myatt, A. (1995). 3-d sound spatialization using ambisonic techniques. *Computer music journal* 19, 58–70

Miccini, R. and Spagnol, S. (2024). A review on head-related transfer function generation for spatial audio. *Applied Sciences* 14, 11242

Middlebrooks, J. C. and Green, D. M. (1991). Sound localization by human listeners. *Annual Review of Psychology* 42, 135–159

Møller, H. (1992). Fundamentals of binaural technology. *Applied Acoustics* 36, 171–218

Poletti, M. A. (2005). Three-dimensional surround sound systems based on spherical harmonics. *Journal of the Audio Engineering Society* 53, 1004–1025

Potter, T., Cvetkovi'c, Z., and De Sena, E. (2022). On the relative importance of visual and spatial audio rendering on vr immersion. *Frontiers in Signal Processing* 2, 904866

Raghuvanshi, N. and Narayanan, A. (2021). Personalizing head related transfer functions for earables. In *ACM SIGCOMM Conference* (ACM), 147–161

Ridgway, E. (2019). Deep dive series: Valve index ear speakers Accessed: 2024-12-18

Roffler, S. K. and Butler, R. A. (1968). Factors that influence the localization of sound in the vertical plane. *Journal of the Acoustical Society of America* 43, 1255–1259

Rogel, A., Hayley, J., Savery, R. J., and Weinberg, G. (2025). What sounds dangerous? establishing correlations of musical features and perceived safety in hri. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (IEEE), 1578–1582

Rumsey, F. (2017). Surround by sound: A review of spatial audio recording and reproduction. *Applied Sciences* 7, 532

Sergi, G. (2013). Knocking at the door of cinematic artifice: Dolby atmos, challenges and opportunities. *The New Soundtrack* 3, 107–121

Spors, S. and Ahrens, J. (2008). A comparison of wave field synthesis and higher-order ambisonics with respect to physical properties and spatial sampling. In *AES Convention 125* (Audio Engineering Society)

Spors, S., Rabenstein, R., and Ahrens, J. (2008). The theory of wave field synthesis revisited. In *AES Convention 124* (Audio Engineering Society)

Valve Corporation (2019). Valve index vr kit technical specifications Accessed: 2024-12-18

Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *Journal of the Acoustical Society of America* 94, 111–123

Wightman, F. L. and Kistler, D. J. (1989a). Headphone simulation of free-field listening i: Stimulus synthesis. *Journal of the Acoustical Society of America* 85, 858–867

Wightman, F. L. and Kistler, D. J. (1989b). Headphone simulation of free-field listening ii: Psychophysical validation. *Journal of the Acoustical Society of America* 85, 868–878

Wightman, F. L. and Kistler, D. J. (1992). The dominant role of low-frequency interaural time differences in sound localization. *Journal of the Acoustical Society of America* 91, 1648–1661

Wightman, F. L. and Kistler, D. J. (1999). Resolution of front–back ambiguity in spatial hearing by listener and source movement. *Journal of the Acoustical Society of America* 105, 2841–2853

Yao, S.-N. (2017). Headphone-based immersive audio for virtual reality headsets. *IEEE Transactions on Consumer Electronics* 63, 300–308

Zhao, M., Sheng, Z., and Fang, Y. (2022). Magnitude modeling of personalized hrtf based on ear images and anthropometric measurements. *Applied Sciences* 12, 8155

Zotkin, D. N., Duraiswami, R., and Davis, L. S. (2004). Rendering localized spatial audio in a virtual auditory space. *IEEE Transactions on multimedia* 6, 553–564

FIGURE CAPTIONS

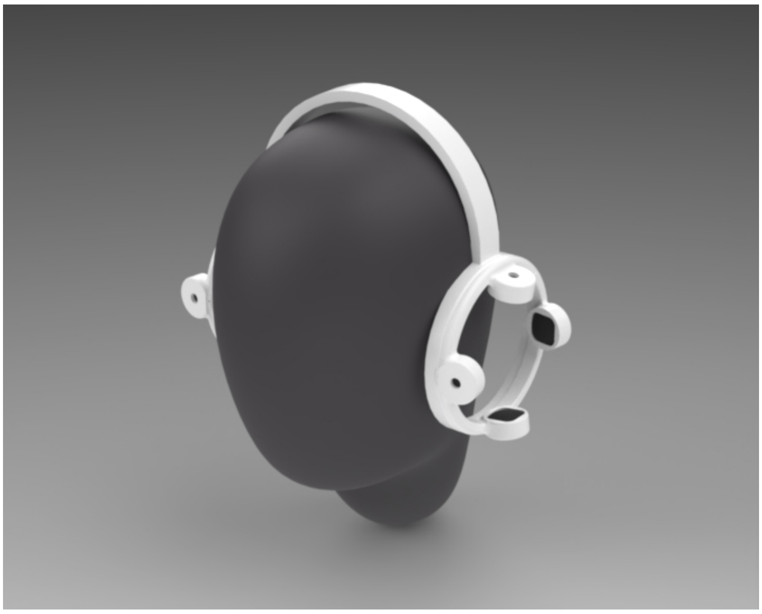

Figure 1. Model of Spherephones

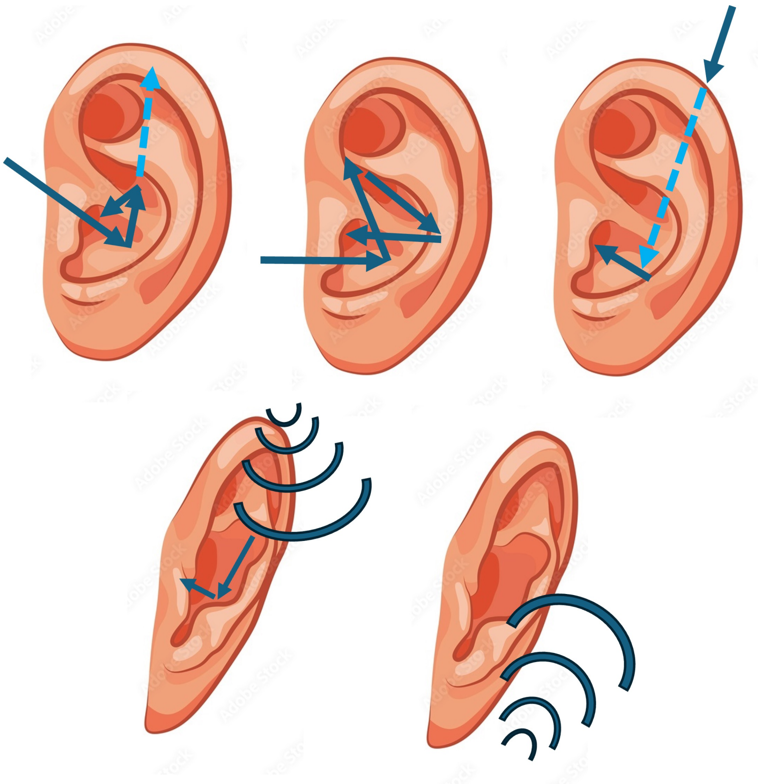

Figure 2. Comparison of ”X” pattern direct geometry versus ”bleed” approach for driver angles

Figure 3. Picture of Spherephones

Figure 4. Picture of Spherephones with Meta Quest

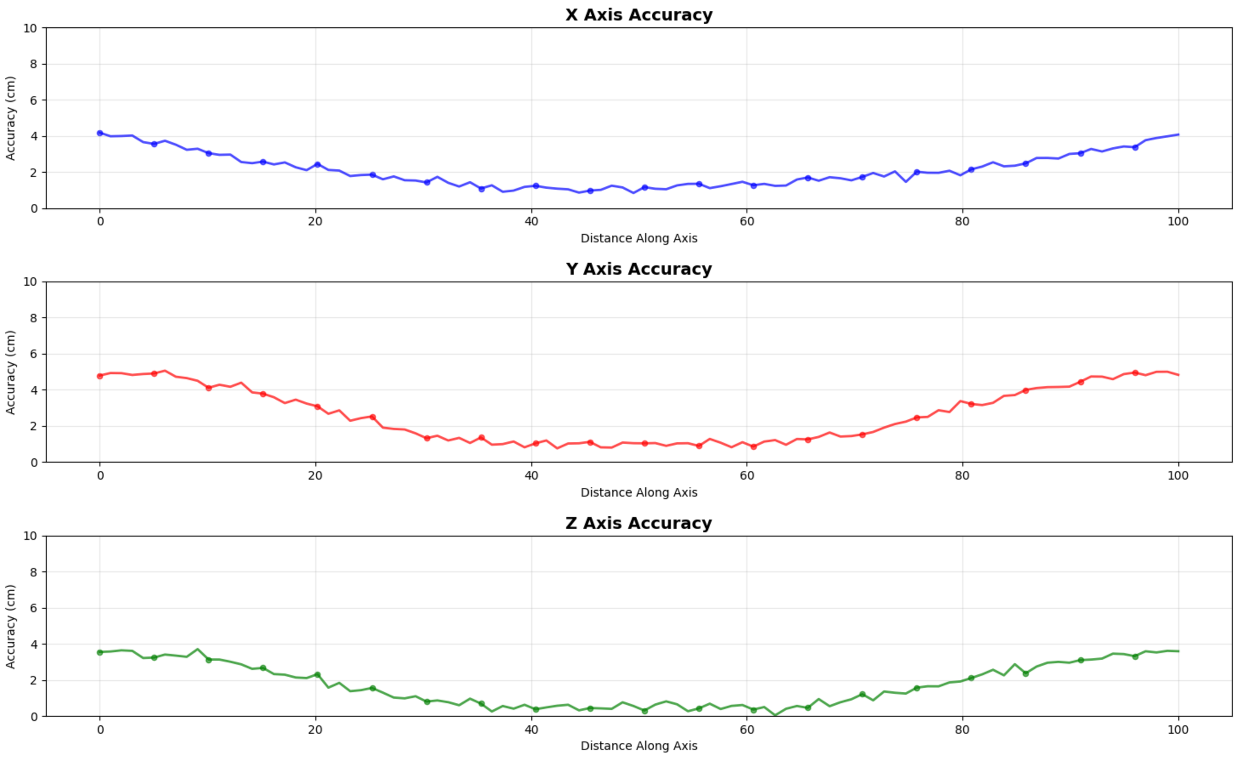

Figure 5. Accuracy of Spherephone panning in three axis based on ambisonic microphone

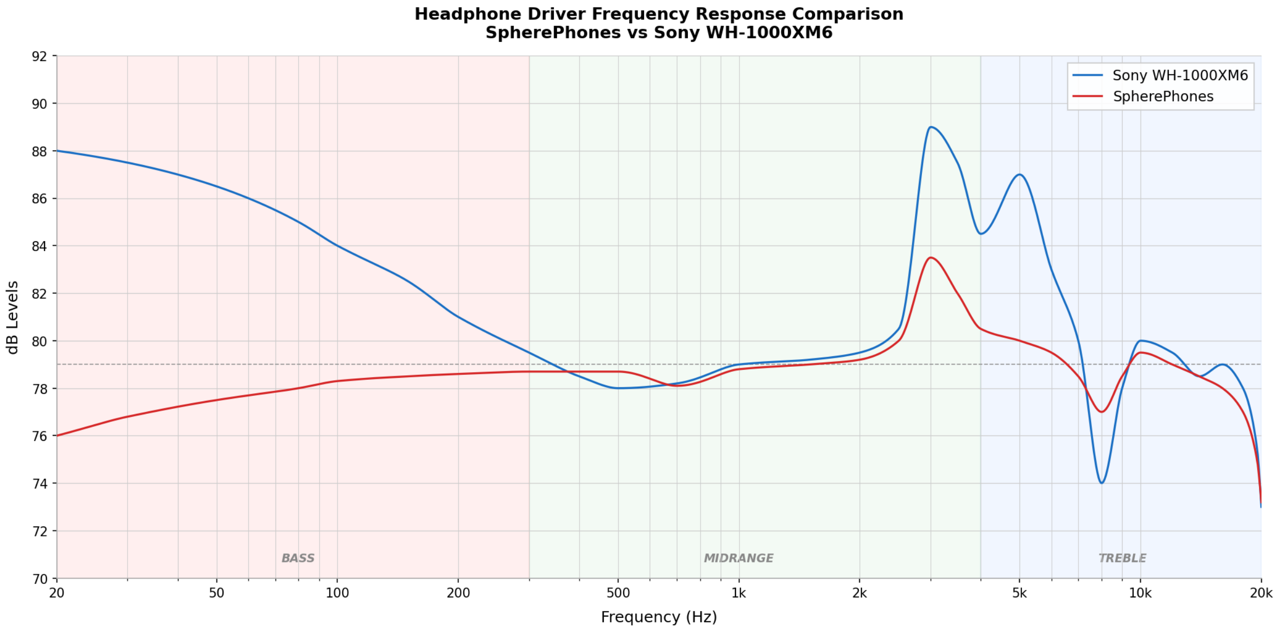

Figure 6. Frequency Response of Spherephones compared to Sony Headphones

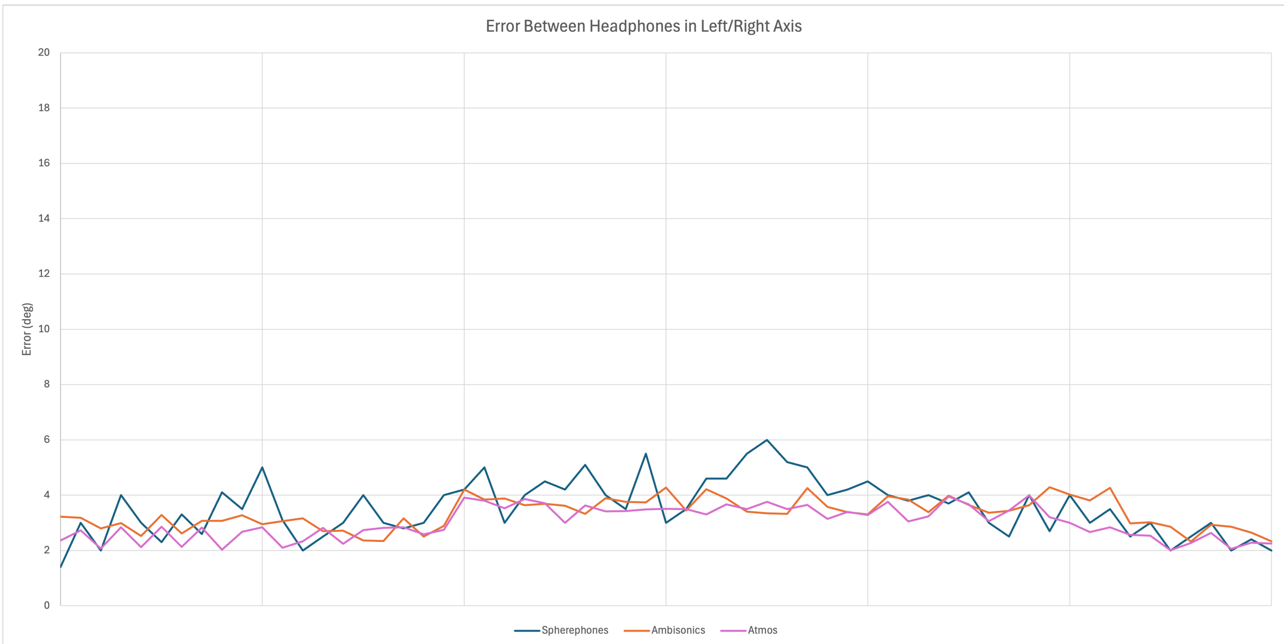

Figure 7. Accuracy of Spherephones along left/right axis

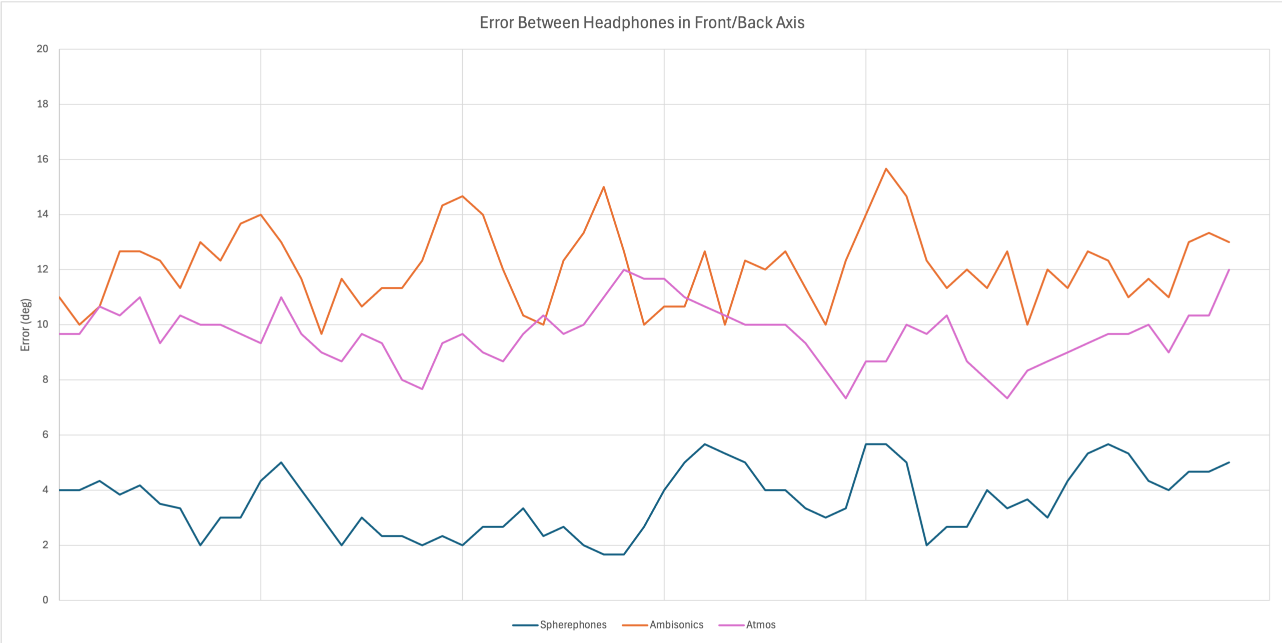

Figure 8. Accuracy of Spherephones along front/back axis

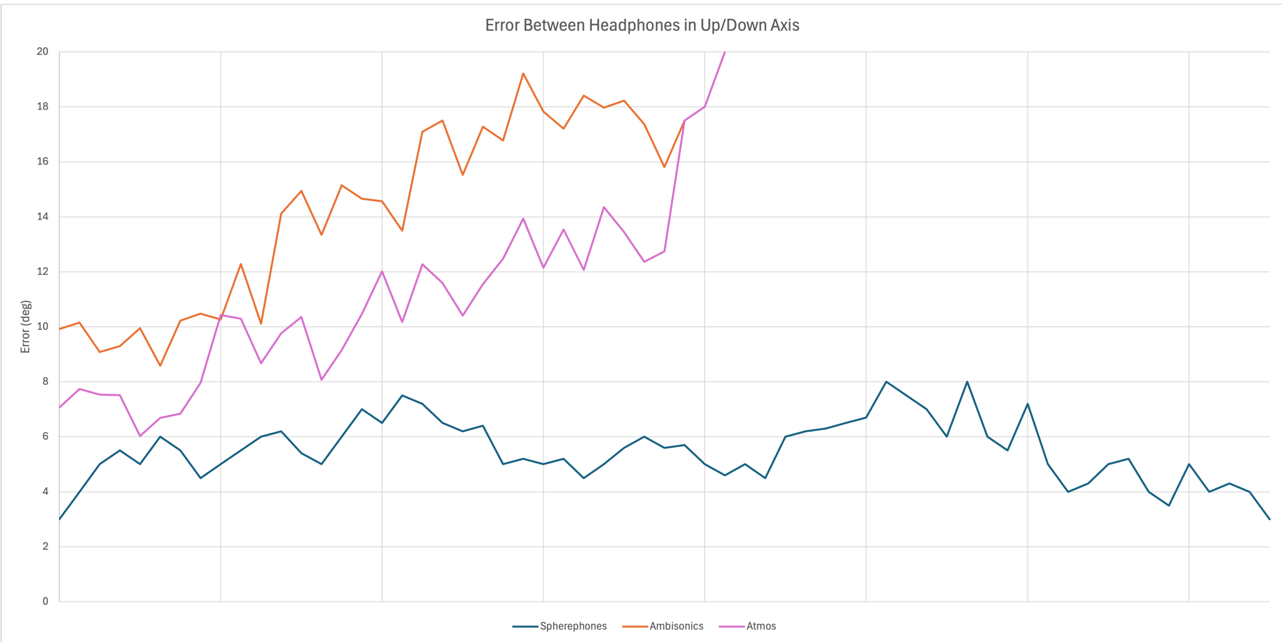

Figure 9. Accuracy of Spherephones along vertical axis