RESEARCH-ARTICLE

Repurposing Audio Playback Tools to Test Human Localization with 6DoF Sound

DANIEL REHBERG, Colorado State University System, Denver, CO, United States

ADAM SINCLAIR WILLIAMS, Colorado State University, Fort Collins, CO, United States

ANIL UFUK BATMAZ, Concordia University, Montreal, QC, Canada

FRANCISCO RAUL ORTEGA, Colorado State University, Fort Collins, CO, United States

Open Access Support provided by:

Colorado State University

Colorado State University System

Concordia University

Published: 28 September 2025

Citation in BibTeX format

UIST '25: The 38th Annual ACM Symposium on User Interface Software and Technology
September 28 - October 1, 2025
Busan, Republic of Korea

Conference Sponsors:
SIGCHI
SIGGRAPH

Repurposing Audio Playback Tools to Test Human Localization with 6DoF Sound

Daniel Rehberg
Computer Science
Colorado State University
Fort Collins, Colorado, USA
Computer Science and Engineering
University of Notre Dame
Notre Dame, Indiana, USA
drehberg@nd.edu

Adam S. Williams
Computer Science
Colorado State University
Fort Collins, Colorado, USA
adamwil@colostate.edu

Anil Ufuk Batmaz
Department of Computer Science & Software Engineering
Concordia University
Montreal, Quebec, Canada
ufuk.batmaz@concordia.ca

Francisco R. Ortega
Computer Science
Colorado State University
Fort Collins, Colorado, USA
fortega@colostate.edu

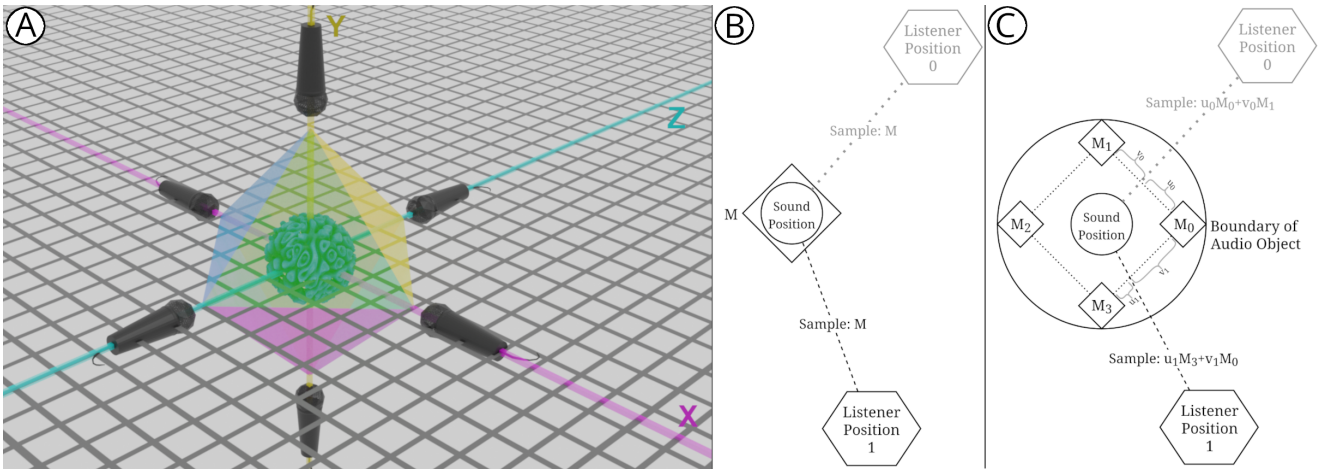

Figure 1: A) Visualization of recording a sound with a first-order spherical microphone arrangement to capture source interference by radial position. The recordings are stored as a triangulated octahedral mesh. B) Conventional object-based 3D sound uses a single recording, limiting playback to 3DoF (translational) as rotation about the sound does not change observed output. C) Sampling from a spherical recording mesh means translation and rotation alters which wavefront segment will reach a listener (6DoF). Interpolation of a boundary's recordings are assessed with linear weights at the intersection of the boundary and the listener's direction to the sound – corresponding to spherical weights on an idealized hull of the audio object.

Abstract

Six-degree-of-freedom audio is a growing interest in interactive software, but it does not easily conform to object-based rendering when achieved with arrays of ambisonics microphones. Prior studies rely on subjective metrics also, which do not clearly indicate how this additional audio interaction might aid a human in a localization task – an indication of enhanced spatial awareness of a sound event. In this paper, we propose an alternative recording and playback technique to achieve six-degree-of-freedom audio to minimize recording overhead, yield object-based rendering, and verify enhanced spatialization through objective testing. The approach taken in this paper utilizes existing audio playback tools in the Unity game engine, and can be redeployed quickly to allow researchers outside of audio engineering exploration in six-degree-of-freedom audio applications. Two studies were conducted within a group of participants using a Microsoft Hololens 2 – testing for interpretation of directional sound cues in a stationary position, and testing the proposed technique in a mobile task. Participants were

UIST '25, Busan, Republic of Korea

ACM ISBN 979-8-4007-2037-6/25/09
https://doi.org/10.1145/3746059.3747682

able to discern additional information within the front-facing "blind spots" and were effectively perfect in a localization task with the proposed audio technique. Participants did not achieve the same performance level with a head-related transfer function alone – indicating meaningful cueing with six-degree-of-freedom sound.

CCS Concepts

• **Human-centered computing** → **Auditory feedback**; **Systems and tools for interaction design**; *Empirical studies in HCI*; *Mixed / augmented reality*.

Keywords

Interactive Audio, Psychoacoustics, Human-computer Interaction

ACM Reference Format:
Daniel Rehberg, Adam S. Williams, Anil Ufuk Batmaz, and Francisco R. Ortega. 2025. Repurposing Audio Playback Tools to Test Human Localization with 6DoF Sound. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST '25), September 28–October 01, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3746059.3747682

1 Introduction

Interactive audio utilizes spatialization techniques to inform listeners of sounds in 3D spaces. Two predominant choices are used in interactive software to account for individual body characteristics of a listener or the characteristics of a soundscape. One of them is the use of a head-related transfer function (HRTF) in which a monaural sound clip is transformed into a stereo output to provide position cues [6, 15, 16, 38] and ambisonics, putting a listener at a fixed location within a sound field where facing different directions changes what scene sounds are heard. Ambisonics has resurged for its use in virtual reality (VR) enhancing experiences such as 360-degree videos [2, 5], but prior to this was used to record sounds for periphony spaces [11]. Studies have moved beyond ambisonics for stationary VR to full six-degrees-of freedom (6DoF) audio in extended reality (XR) applications [17, 26, 27].

Both HRTF and ambisonics have a place in XR applications, but HRTF has limited degrees-of-freedom with point-source sounds and 6DoF ambisonics has inhibitive technical concerns including cost and a misalignment with object-based rendering. Currently, XR studies using 6DoF ambisonics-based audio do not necessarily illustrate why this additional audio information might be pertinent to a listener. 6DoF could provide supplemental localization cues about an object's orientation relative to a listener. This is best explained by first understanding an HRTF limitation.

HRTF is well studied as its constituent parts define the primary cues for auditory spatial perception [6, 15, 16, 38]. An HRTF is based on three physical interactions of an audio event as heard by a listener with two ears [6, 17]. Two of these are interaural differences as attenuation and timing of a sound will vary between the ears — referred to as interaural level (ILD) and time (ILT), respectively. The third attribute is the most complex and fairly unique to an individual — the head-related impulse response (HRIR). The HRIR describes how a person's body will reflect, occlude, or diffract sound, altering its spectral signature before it is heard. When simulated in software, a single monaural clip (a 1-channel recorded sound) is passed through an HRTF (as a pair of filters for each ear) to form a stereo output presenting this audio as a point-source sound in 3D space.

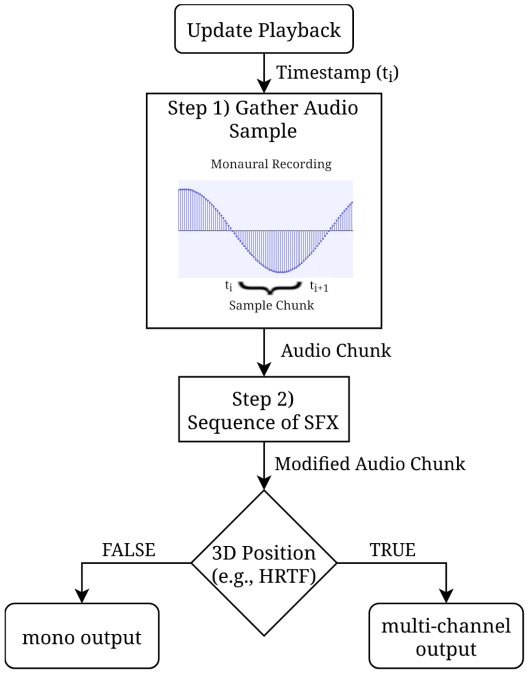

Figure 2: Interactive audio presentations begin by sampling chunks of a sound file over timesteps before being processed by SFX for environmental or positional simulations of spatialization.

HRTFs work in a paradigm of object-based rendering as a monaural sound can be arbitrarily instantiated and run through an HRTF in interactive software. This allows sound recordings to be repurposed and reused for different scenes, videogames, or XR and psychoacoustic experiments [6, 26] – being instantiated at different positions and times during the lifespan of the program. Many audio libraries support HRTF [4, 8, 9, 24, 33, 34] among other filtering techniques (e.g., doppler, reverberation, etc..) through this object-based rendering approach of reusing monaural recordings. These interactions work by sampling chunks of monaural recordings over time, and then passing them through sound effects (SFX) as shown in Figure 2.

What conventional HRTF playback lacks is directional audio information. Specifically, by always presenting a single sound clip as a point in space, it means nothing to rotate this object – it has no differentiable surfaces to enable a listener to interpret an orientation state of the audio. Of course, an HRTF is not seeking to resolve this issue – it focuses on presenting a sound at a location in 3D space about a listener (i.e., an egocentric presentation). This is shown in the segment B of Figure 1, where a listener's position does not change which audio file to sample for an instance of an sound object in a 3D environment.

Ambisonics differs from the playback of monaural audio because an entire soundscape is captured from a coincident microphone

3D Sound	Translational	Rotational	Playback
HRTF	3DoF	0DoF	Object-based
FOA/HOA	0DoF	3DoF	Soundscape
FOA/HOA Array	3DoF	3DoF	Soundscape

Table 1: Technique Features, Sound Relative to Listener

array [11]. This means multiple channels are recorded, e.g., one per microphone, from a common location. The arrangement of the microphones are inside-out, such that the microphones perceive the world around their common center, i.e., sound is converging towards the device. Each audio recording can be mapped directly to a single physical sound channel in a periphony or general surround sound system [11, 17, 30]. The multiple directional recordings of a soundscape allow rotational degrees-of-freedom [2, 5] but do not provide changes in audio positions like HRTF — which is useful in XR given a potential for limited space and mobility of an end user. Audio engineering studies have extended ambisonic capabilities by filling rooms with several ambisonics devices [21, 25, 27] — enabling near 6DoF reproduction experiences in XR by dynamically switching which microphone recordings are header based on position and orientation of the end user. Table 1 summarizes the capabilities of these existing practices.

Given 6DoF audio objects, we might be able to ask new spatialization questions to end users. Instead of only asking a listener "where a sound is located," there exist the potential to also ask a listener what direction a sound is facing. For example, given a 6DoF recording of a tea kettle boiling, a listener might be able to not only localize the position of the kettle as it whistles, but whether the steam is blowing towards or away from them – and potentially directions in between.

6DoF audio research in XR is still fairly new and has only evaluated subjective qualities with human subjects, meaning limited objective answers and metrics exist to evaluate the full effects of interactive 6DoF playback. To build a controlled – repeatable – experiment to ask position and orientation localization questions, it is beneficial to simplify 6DoF audio reproduction. We prescribe a methodology for recording and playback to yield object-based 6DoF capabilities in commodity software – like game engines. We isolate directions from individual sound events by purposefully arranging microphones around a single sound source – following an outside-in recording approach (sphericial arrangement) instead of inside-out from conventional ambisonics. This technique can be combined with HRTF playback (among other SFX), which we tested with, and without, in human-subject studies. Our results indicate that directional recordings can influence perception of a sound object's position and direction and that, for human speech, 6DoF playback is meaningful to orientation localization over HRTF alone.

2 Related Work

2.1 Variance of Audio Observed by Direction

6DoF sound, pertaining to rotational and translational freedom, requires capturing how an object emitting a sound is altered through self-imposed physical interference(s). Traditional HRTF with a monaural sound source, throughout we will refer to as *monaural HRTF*, only has one directional perspective of a sound event, and the HRIR is concerned with how one's body alters the spectral form of a sound [6, 15, 26]. That is, HRTF with its HRIR represent an egocentric perspective of an audio event and is not directly concerned with the orientation of the object making noise relative to the listener.

Theoretical models demonstrate how the shape of an object will alter the waveforms emitted from that object [14]. This is further illustrated in sound source directivity, where amplitude of frequencies change when heard from different orientations about a sound source. Additional studies on sound source directivity with human speech indicate spectral modifications are perceived when heard from different directions [18]. This makes sense as sound waves are mechanical, and can be interfered with as they travel about the surfaces adjacent to their emission source.

Leishman et al. [18] used a spherical microphone arrangement to map the variance in waves that observers would hear if located at different orientations about a person talking. We show this quality visually with a set of spectrograms as seen in Figure 3, where we recorded six different locations around a human reading a prompt. These microphones were in a first-order arrangement, capturing along the positive and negative directions of a 3D coordinate basis.

In studying near-field HRTF, directivity considerations are important, but concern capturing a monaural recording which excludes variance based on directionality [19] – a description of an omnidirectional sound that is consistent regardless of where it is heard. This is important as monaural recordings used with HRTF generate a point-source sound, and environmental influences like a room-impulse response (RIR) – reverberations of a sound in an environment – can cue a listener into information about the space they occupy [3, 12, 22, 28, 31, 32, 39] rather than the location of a sound source.

This *"ideal"* recorded sample for a point-source sound does not necessarily account for the precedent effect [28, 35], where humans tune in to the origin of a sound source. A monaural HRTF playback only provides the locale of the original sound source, but it should be possible that a 6DoF recording might provide a listener the additional audio information to hone in to directionality given the precedent effect.

These are the types of questions that could be explored with 6DoF object-based sound. The closest objective research regarding enhanced perception of sound locality using directional information extends from periphony studies using ambisonics [11]. We discuss some of these studies briefly as they demonstrate an advantage to increased directional recording resolution in regards to localization and have led into research where XR HMDs are utilized.

2.2 Ambisonics, Periphony, and Sound Localization

Periphony studies use a room with a circular or spherical arrangement of speakers. If a listener stands in the center of the room, the resolution of the audio presented to them from all directions can be dictated directly by the order of an ambisonics recording. As order increases, a listener is provided with higher density information

from a recorded soundscape. The intriguing results of higher-order ambisonics (HOA) is the increase in resolution can overcome deficiencies from ILD and ITD confusion with point-source sounds. Accuracy in localization is determined by having a participant (sometimes blindfolded) provide a response (e.g., point in the direction) about where they perceive a sound source [7, 13, 29].

Power et al. [29] is one such example wherein 3rd-order ambisonics was found to be an ideal resolution for human performance. Their study utilized a spherical arrangement of speakers to assess accuracy in a vertical range, *-35 to +35 degrees*, and in a 360-degree horizontal range. Their ambisonics recordings were of real-world sound events, and their results indicated that accuracy increased in vertical and horizontal localization tasks with HOA compared to first-order ambisonics (FOA). In direct relation to an HRTF, theoretical modeling has agreed with studies like Power et al. [29] indicating that ILT and ILD interpretation of position should increase with HOA over FOA by minimizing gaps (errors) in timing and amplitude of sounds [7].

HMDs have been involved in validating 3rd-order HOA as an ideal resolution of spherical surround sound. Huisman et al. [13] performed a periphony localization study with FOA and HOA in 3rd, 5th, and 11th-order arrangements. Their study included conditions of participants being blindfolded – with and without an HMD – or not blindfolded and seeing the real environment – no HMD – or a VR representation of the periphony room – with an HMD. Regardless of the conditions, the gap between FOA and HOA was notable, but the margins between 3rd and higher-orders reveal diminishing returns beyond 3rd order.

Huisman et al. [13] study indicates a transition from real-world periphony studies to XR HMDs — a desire existing as far back as the 1980's in NASA research [36, 37]. While their results indicate that the HMD did not effect localization, they were actively looking to see how much an HMD would disrupt a listener's natural HRTF when using physical speakers – a motivation to fully simulate the audio space in, for instance, VR. They point to research where this has been found to be problematic [10].

While HOA studies reveal human capabilities to locate a sound from a soundscape, the method does not allow listeners to walk around sounds within that soundscape. Recent studies delve into 6DoF reproduction using arrays of ambisonics devices with HMDs for user tracking, but do not ask localization questions like those from traditional periphony research, but in fully traversable 6DoF 3D space.

2.3 6DoF Developments with Human Subjects

A motion towards exploratory play of *"6DoF like"* (not 6DoF but more than 3DoF) interactions in VR through ambisonics can be found in Plinge et al. [27]'s paper. Their work used a single ambisonics recording in which they altered the direction-of-arrival (DoA) to a listener as they walked around a virtual environment. They compared a baseline using a MUSHRA test to there DoA translation interactions with both an FOA and HOA recording. The "6DoF like" playback with HOA was marked as closer to the baseline compared to the the FOA condition.

Truly representing 6DoF audio with ambisonics is more involved than this for both recording and reproduction in software, relying on a room filled with ambisonics devices and some type of interpolation mechanism to blend between device recordings in software [17, 23, 26]. Effectively, this amounts to recording a soundscape from multiple locations and playing back the recordings that are closest in proximity to an end user in a virtual scene.

Patricio et al. [25] advances upon prior work by providing a closer to 6DoF experience by filling a room with ambisonics devices. Both FOA and 3rd-order HOA recordings were captured as conditions to evaluate against a baseline in another MUSHRA study. Participants ranked the 6DoF HOA case close to 80 on average (out of 100) whereas 6DoF FOA was ranked closer to 70 on average. This form of subjective testing and result in VR is consistent in new research expanding upon ambisonics-based 6DoF [20, 21].

There appears to be a gap in measuring any performance advantages provided by 6DoF audio cues in XR. While qualitative testing is important, spatialized sound is a technique which exploits human perception to deliver information about an environment and objects therein [17, 26]. It is possible that objective localization tests could be performed in current ambisonics 6DoF practices, but the cost is potentially prohibitive to fields outside of audio engineering given the number of ambisonics devices and recording space required. Because these recordings are of a fixed soundscape, the recorded audio cannot be repurposed to build arbitrary virtual environments – because it does not fit an object-based rendering paradigm. McCormack et al. [21] outline an approach using beamformers to help isolate individual sound events in an ambisonics 6DoF set of recordings, but this does not address the initial investment required to produce a 6Dof ambisonics (preferrably HOA) recording.

To begin exploring 6DoF questions of localization, we outline a 6DoF model which is more cost effective, inherently object-based, and is easily implemented using existing interactive audio software tools. We use the knowledge from sound source directivity and periphony studies to infer that directional recordings may enhance user perception of spatial audio – information which leads our hypotheses in two studies.

3 6DoF with an Octahedron Mesh

To produce a 6DoF object-based sound, we utilized a spherical microphone array in a novel way. For ambisonics, a spherical microphone arrangement records in an inside-out method, capturing the world around the centroid of the microphones. As seen in sound source directivity studies [18], a spherical arrangement of microphones could be made to record in an outside-in manner.

This style of microphone arrangement allow us to form 6DoF audio objects. Prior 6DoF studies achieve multidirectional sound samples from their array of ampbisonics devices, but this fails to isolate a single audio event. The recording practice described here illustrates how to prioritize recording samples to a single event – enabling object-based audio that can be rotated and translated. Moreover, by keeping to existing object-based playback tools any standard SFX can be applied to these 6DoF audio objects, including ILT, ILD, or full HRTF simulation. The method described here

guarantees 6DoF interactions, which are not the case with one floor-plane of ambisonics devices – as that does not allow for rotations or translations into further vertical planes.

The use of six axis-aligned microphones ensures an $O(1)$ evaluation of picking which microphones to sample when delivering moment-to-moment audio information to the listener. This is achieved by forming a triangulated mesh of the microphone recordings; each triangle therefore exists in one of the 8 octants of 3D space. Because sound propagates outward in all directions from a source, we assume a uniform expansion which means only one of these triangles would reach a listener (based explicitly on the octant the listener is observing the sound event from).

Our method breaks down into two parts offline recording and online audio sampling.

3.1 Recording Method

A visualization of the outside-in spherical arrangement is shown in Figure 1. For this research, six AT2020 microphones were placed 0.25 meters from the recording subject along a 3D standard basis. The radial distance separating the microphones is arbitrary, but sound will attenuate meaning placing microphones too far apart may lower recording quality (loss of resolution in pressure wave differentials). This 0.25 meter placement was to allow enough space for a human to fit within the spherical arrangement of microphones.

Each microphone corresponded to a positive or negative location along these X, Y, and Z basis axes. The microphones faced each other acting as *"observers"* hearing a common sound from different locations. By being positioned a common radius from the recording subject (i.e., spherical arrangement), their observations converge to a common center.

This common center acts as the local origin to the triangulated mesh formed from each of the six recording samples. Recording simultaneously from each microphone at a consistent radial placement, ensure a common parametric representation of time can be used during playback in software. Saying at some time **t** what did each *"observer"* (i.e., microphone) hear. Physically, this time parametric is a plausible heuristic given the spherical arrangement of microphones – as a sphere represents the uniform outward expansion of sound over time.

3.2 Playback Method

We utilize the existing tools of the Unity game engine to illustrate how these six separate recordings can be used to produce a 6DoF audio experience given basic audio playback functions. Game engines allow an arbitrary number of sounds to be played, which means all six recordings can be played back simultaneously. A mix of these six recordings would yield a potentially incomprehensible sound, but the volumes of each playback can be altered in real-time. For instance, if a listener is co-located at the position of one of the recordings in the triangulated mesh, then its playback volume could be set to 1.0 (from a normalized range) while the others are set to 0.0. This means the mesh boundary most relevant to the listener's observation point is presented.

Playing all six recordings at once ensure parametric alignment over time, and by modifying the volumes of each recording, the final sound samples heard can relate to a listener's orientation about the sound mesh. In addition, each sound played can be spatialized (i.e., have a position in 3D space). This means the engine's capabilities to produce positional sounds – ILT, ILD, or even full HRTF – are still available. Technically, sampling from boundaries of the audio mesh allow for both translational and rotational degrees of freedom, but allowing use with an HRTF incorporates well established egocentric perspective of positional sound.

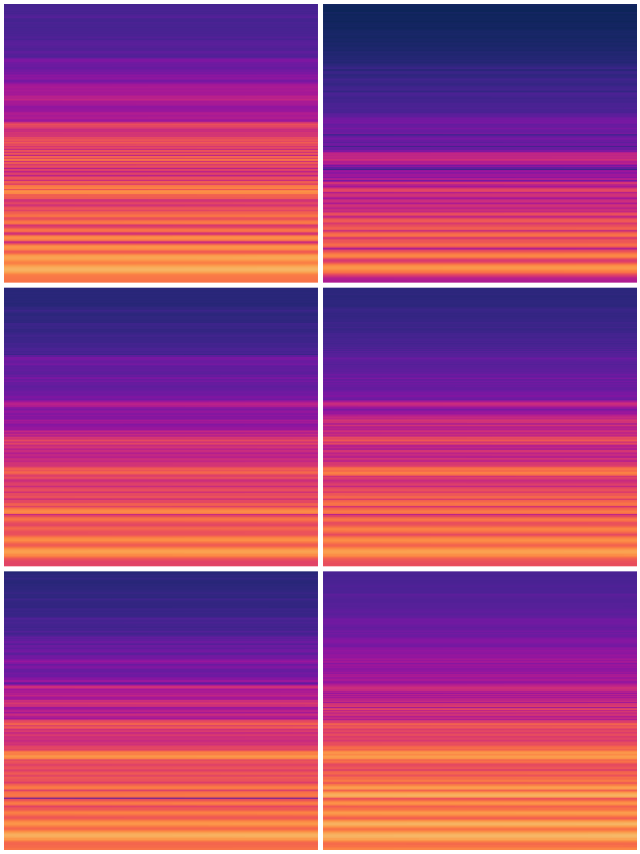

Figure 3: Six small snippets of spectrograms from human speech. Each spectrogram is from a different recorded direction from a first-order arranged spherical microphone array. Top row: front and back microphones. Middle row: left and right microphones. Bottom row: top and bottom microphones.

The hull of the triangulated mesh of recordings forms an octahedron. The goal is to only play a mix of sounds closest to the listener, which means finding the closest triangle. We exploit that the octahedron is formed from recordings along a standard basis (a first-order arrangement) to define an O(1) algorithm. The object-based instantiation of a rotatable and translatable object means the object-instance has a position and orientation state. By applying an inverse transformation of these states to a listener's position, we can check a few conditions to determine which octant the listener is observing the sound from. Because each octant in the proposed arrangement (see the octahedron in the recording arrangement

of Figure 1 A) only has one triangle, the result of finding an octant reveals which triangle (therefore recordings) to sample.

Once the nearest triangle is found, the listener can be projected to the closest point on the sphere hull that the recording vertices exist. Once there, barycentric projection is performed to return weights which indicate how much volume should be applied for each of the three nearest sounds. All sound volumes are first set to zero, then the barycentric weights are applied as volumes to the three nearest sound objects. The barycentric weights ensure the sum of volumes never exceeds 1.0 – and enable gaps in the audio mesh to be filled via interpolation of boundary samples. The pseudocode describing the algorithm used in Unity is shown in Algorithm 1.

Algorithm 1 Update Sound Object

Require: $volumes = (front, back, left, right, top, bottom)$ AND
$volumes_i = [0, 1]$ AND
$volume_i$ is a $radius$ away from recorded subject
Ensure: $\sum_{i=0}^{5} volumes_i = 1$
procedure UpdateVolumes($listener$)
$relative \leftarrow objectRotation^{-1} \times (listener - objectPosition)$
$octant \leftarrow octantOfPosition(relative)$ ▷ Integer of octant
$u \leftarrow (0, radius, 0)$ ▷ Default vertical sound position
$r \leftarrow (radius, 0, 0)$ ▷ Default horizontal sound position
$f \leftarrow (0, 0, radius)$ ▷ Default depth sound position
$indices = (4, 3, 0)$
if $octant$ is 3 **or** $octant$ is 4 **or** $octant$ is 7 **or** $octant$ is 8 **then**
$u.y \leftarrow -radius$ ▷ Negative vertical position
$active_0 \leftarrow 5$
end if
if $octant$ is 2 **or** $octant$ is 3 **or** $octant$ is 6 **or** $octant$ is 7 **then**
$r.x \leftarrow -radius$ ▷ Negative horizontal position
$active_1 \leftarrow 2$
end if
if $octant$ is 5 **or** $octant$ is 6 **or** $octant$ is 7 **or** $octant$ is 8 **then**
$f.x \leftarrow -radius$ ▷ Negative depth position
$active_2 \leftarrow 1$
end if
$weights \leftarrow barycentricWeights(u, r, f, relative)$
$i \leftarrow 0$
while $i \neq 6$ **do**
$sounds_i$ set volume to 0
$i \leftarrow i + 1$
end while
$i \leftarrow 0$
while $i \neq 3$ **do** ▷ Apply volumes to closest triangle only
$soundVolume_{active_i} \leftarrow weights_{active_i}$
$i \leftarrow i + 1$
end while
end procedure

4 Methods

4.1 Equipment

The experiments for each study were developed in the Unity game engine using the Mixed-Reality Toolkit (MRTK) API to build and deploy the application on a Microsoft HoloLens 2 HMD. MRTK with Unity automates the use of hardware accelerated HRTF on HoloLens 2 when sound objects are set to both "spatialized" and "3D." These settings were activated as independent variables when needed in Study 1 and 2. The microphones were Audio-Technica AT2020 connected to a Focusrite 18i20 audio interface. The HRTF on the HoloLens 2 is generalized, and while not tailored to each participant, does determine approximate ear separation from initial interpupillary calibration.

An open-back pair of Audio-Technica 900X headphones was used. The headphones were connected to the HoloLens 2 through a Google 3.5mm to USB-C adapter. A standard decibel meter was sealed around the headphones while connected to the HoloLens 2 to calibrate all sound playback to fall between 55-65 dBA – the standard amplitude of conversation. The calibration of sounds occurred on a per-recording basis such that this target decibel range was reached when setting the HoloLens 2 system volume to a constant level.

The HoloLens 2 and open-back headphones were chosen to ensure participants did not feel isolated from reality, using this equipment to yield an augmented experience to minimize distractions. Participants were asked for verbal responses that a researcher would document using a Bluetooth keyboard. The keyboard was paired with the HoloLens 2 to record responses. This input also progressed the experiment to subsequent trials, keeping the previous trial locked until the participant's response was entered.

4.2 Recordings

Six directional recordings were taken of human speech. The microphone arrangement was approximately 0.254 meters in front, behind, above, below, left, and right of the speaker with an attempt to have the speaker's mouth centered about the microphones. Study 1 used a recording of the statement "hello" for short-form dialogue. Study 2 had a section of the ACM Code of Ethics [1] recited for a long-form dialogue.

The calibration of the directional audio occurred in two steps. In the first step, the audio interface had the individual gains for each XLR input modified to match a signal response repeated at the same distance and front-facing orientation of each microphone. The second step was reducing the amplitude of a recording to match a target 55-65 dBA output in headphones. Given the front recording has the largest amplitude signature, and the first calibration ensured all other microphones had matching gains during recording, only the front-directional recording was used to determine the required amplitude change to reach the dBA target. This modification was applied to all directional recordings to provide consistency for the amplitudes of each recording.

4.3 Participants and Experiment Structure

27 college students, mixed between graduate and undergraduate populations, participated in this study – 10 female, 17 male. The participants were run through two studies in one 30-minute session. Participants were compensated with extra credit in a computer science course or with a $10 gift card.

The two studies were preceded by a hearing test. A series of sinusoidal frequencies were played between the left and right ear.

The frequencies were in the lowest to highest ranges of average human voices. Each frequency had its volume manually set to match the target 55-60 dBA of the experimental trials' audio. Participants responded by saying the sound was on the left or right after it had played. Participants moved on to experimental trials upon completion of the hearing test.

Both studies had a total of 4 blocks of trials. The order of blocks were pseudorandomized in a factorial manner, giving 24 permutations of blocks per experiment. The order of experiments provided was alternating, meaning a total of 48 permutations existed within the study. Aside from counterbalancing the order of experiments, each experiment was performed twice to help acclimate participants to trials – e.g., Study 1, then 2, then 1, and 2 again.

4.4 Study 1

Study 1 was a stationary sound localization task in which participants sat in front of a virtual (augmented) cube. This cube was a fixation point representing the approximate distance to sound events. The cube was 1.2 meters away and had dimensions of 0.25 meters. The 0.25 meters was chosen to match the distance of the spherical microphone arrangement used to record human speech.

This study had four randomized blocks of trials. Each block had different independent variables applied, described momentarily. Each block involved arranging six sounds at the centers of each face of the fixation cube. A single trial consisted of a 3-second countdown (shown visually on the HMD) before playing one of these sounds 3 times – with a 1-second pause between each playback. The sound heard was from one of the six spherical recordings of a human speaker saying "hello." The participant was then presented with text in the HMD asking if they heard the sound from the top, bottom, left, right, front, or back of the cube. The participant's response was entered by the researcher on a keyboard to record the result and initiate the next trial. The subsequent trial played one of the next sounds located on the cube in the same manner, until all six trials were completed (i.e., all six sounds about the cube had been played) after which the next block of trials would occur.

In each block, the six sounds located on the cube's faces were pre-randomized such that no repeat order of playback would occur – e.g., two blocks could not play sounds in the same order, such as left, top, right, back, front, bottom, etc. The primary independent variable per block was whether each sound was the same monaural recording using HRTF, separate directional recordings, or separate directional recordings while also using HRTF. A naming convention is presented to distinguish the independent variables and aid in disseminating the results.

Direction Only (Dir Only) refers to the trials where no HRTF was used, and each sound sample was one of the six directional recordings. *Misaligned* refers to trials where HRTF was used with each of the six directional recordings, but none of the directional recordings aligned with their position on the fixation cube (e.g., front-directional sound would not be located at the front of the cube). This meant the Misaligned block could be analyzed in two ways for human localization: by directional sound, *Misaligned (Dir)*, and by sound position, *Misaligned (HRTF)*. *Aligned* refers to trials that used HRTF and the six directional recordings, where each of the directional recordings aligned with their position on the cube (front-direction sound at front of cube, left-direction sound at left, etc..) *HRTF Only* refers to trials that used the contemporary monaural HRTF approach of utilizing a single "ideal" recording, placed in each of the six different positions about the fixation cube. The "ideal" recording in the HRTF Only block was the same front-directional recording used in the other blocks, as convention for audio recording is to record a speaker talking directly into a microphone. Once all blocks were concluded, the program would move onto the next experiment.

4.5 Study 2

The second study utilized the 6DoF playback technique, utilizing existing interactive audio controls in Unity. A fixation cube was again used as a focal point to help a participant visualize the approximate location of sounds. In each trial, a longer dialogue was played as participants walked around the cube. Their objective was to stop in the location where they thought they would be facing the person talking.

To help distinguish orientation around the cube, each face was a different color. Before each trial, a 3-second countdown was displayed to the participants to signify when the trial would begin. While dialogue was playing, the participant walked around the cube to listen for changes based on their movement – determining where to stop based on these interactive cues. After the dialogue concluded, text was presented to the participants in the HMD asking them how confident they were – on a scale from 1 to 10 – that they were standing in front of the person speaking. Their response was recorded by the researcher, at which point the confidence value and the relative position of the person about the cube – a vector – was recorded for accuracy and analysis. At this point, the program would move on to the next trial.

The experiment in Study 2 had four blocks with one trial per block. The conditions of the block were whether the 6DoF playback method was used, or if monaural sound was used with HRTF. Three blocks used the specialized rendering method with varying conditions. By playing all six recordings as separate 3D-positioned objects, each individual audio track could be spatialized (via HRTF) independently. This enabled additional conditions where the set of sounds used for 6DoF playback could be located at different or shared positions – offering more variance while using an HRTF.

6DoF Only was a condition using only the playback method – no HRTF. This meant that 3D positions of the sounds would not alter the output – only the modulation of volumes through the 6DoF playback. If a listener can pick up on directional audio cues, then this should provide localization by orientation about a sound event. In contrast, HRTF with a single sound could potentially provide orientation cueing – say positioning a sound on one of the faces of the cube (see Monaural HRTF conditions) – because an observer could listen for amplitude changes (i.e., maximum amplitude means closest to sound object). However, HRTF strictly provides egocentric perspective and finding an orientation through audio cues may not always relate to focusing on a loudest audio sensation.

6DoF w/Surfaces was the second block condition which used the 6DoF playback method but with HRTF. Each sound sample from the octahedron mesh was positioned on distinct faces of the cube (left recording to left side, right recording to right side, etc..).

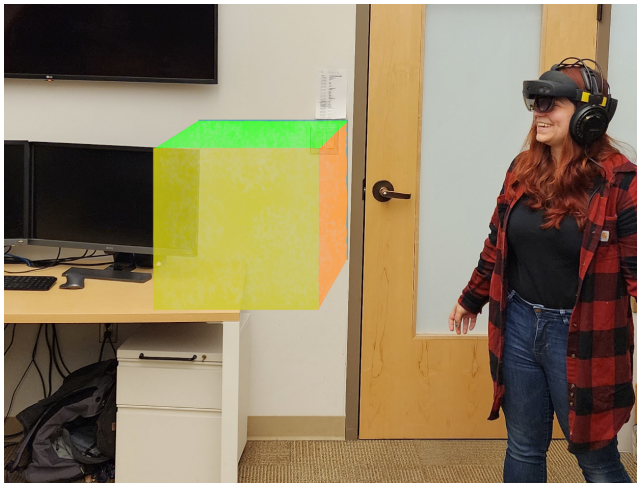

Figure 4: Depiction of Study 2's experiment.

This meant volume modulation provided 6DoF playback, but each audible sound (non-zero volume) was additionally altered through the HoloLens 2's HRTF – providing distinct ILD, ILT, and HRIR effects to each sound. Depending on the results of the other three blocks, this would help determine if amplitude changes from an HRTF outweighed 6DoF orientation cues.

6DoF Centered was the condition in the third block, which used the 6DoF playback and HRTF. This condition had all six sounds collocated at the center of the cube so that they were spatialized with the HoloLens 2's HRTF in a uniform manner. These conditions most accurately represent how the original recorded sound would be heard.

Monaural HRTF was the condition in the fourth block, which did not use the 6DoF playback method. The sound was located on one face of the cube to ensure that a listener could still be directed to a *"front-facing-direction"* of the cube given the observation of amplitude changes by proximity to the sound source. This condition should indicate if HRTF is useful enough to provide cues for orientation localization about an object.

Once all four blocks were concluded, if the participant had not completed both experiments twice, then we continued to Study 1's experiment.

4.6 Hypotheses

Four overarching hypotheses are presented in these studies, three within Study 1 and one in Study 2. The names of hypotheses are defined here in a monotonically increasing manner. The Misaligned block provides conditions that fit either the directional audio cues or monaural HRTF cues, and so this block is reflected twice as appropriate for Study 1 hypotheses below.

H_1 is from Study 1 and is that the accuracy is greater for front and back sound locations when using directional recordings compared to using a single sound with HRTF. This is indicated within a Wilcoxon analysis by each directional audio block – Dir Only, Misaligned (Dir), and Aligned – compared to blocks relying on spatialization with a HRTF – Misaligned (HRTF) and HRTF Only blocks. Because independent variables are changed between each block, this hypothesis is conducted between all pairs of directional audio blocks and HRTF spatialization blocks. Given a block with directional audio condition D and a block with monaural HRTF spatialization M, each hypothesis is $H_1 : D \neq M$.

H_2 is from Study 1 and is that directional sounds have participants answering with more *"non-front"* responses compared to only having HRTF provide spatialization with a constant monaural sound. This hypothesis involves having three separate block comparisons given the independent variable changes in the directional audio blocks: Dir Only, Misaligned, and Aligned. These three blocks are compared separately to the HRTF Only block. Given a block with directional audio condition D and the HRTF Only block R, each hypothesis is $H_2 : D \neq R$.

H_3 is the last hypothesis in Study 1 and is that directional audio alters the overall accuracy of a listener localizing a sound. It is tested separately with all the Dir Only, Misaligned (Dir, HRTF), and HRTF Only blocks against the Aligned block. Given one of the listed blocks X tested against the Aligned block A, each hypothesis is $H_3 : X \neq A$. This hypothesis is based on having more coherent information regarding spatialized sound compared to direction only audio samples or a monaural HRTF presentation.

Study 2 has one hypothesis, H_4, which is that the localization of a sound about a position and direction in 3D space is more accurate provided one of the 6DoF conditions – 6DoF Only, 6DoF w/Surfaces, and 6DoF Centered blocks – compared to a traditional monaural HRTF presentation – the Monaural HRTF block. Given a 6DoF block condition D and a traditional HRTF presentation R, each hypothesis is $H_4 : R \neq D$.

The Wilcoxon tests are performed in a two-sided manner, and so each hypothesis is described with *not equal* operators. However, the positive and negative weights of the Wilcoxon test statistic indicate conditions have greater influence in side-by-side comparisons.

5 Results

The results were collected at the end of each participant's run through Study 1 and 2. One participant was removed from the analysis due to malfunctions occurring with the HMD during the experiment trials resulting in a total of 26 participants analyzed.

All hypotheses involved a comparison of monaural HRTF to a treatment condition involving additional directional sounds or a 6DoF reproduction. Data analysis in Study 1 required integers for counting the frequency of a response for a given block whereas Study 2 involved floating point values derived from vector comparisons. The non-continuous values to analyze in Study 1 prompted use of the Wilcoxon signed-rank test. Given fewer than 30 participants, the data in Study 2 was considered to not fall into a normal distribution which was validated through Shapiro-Wilk tests. For these reasons, the Wilcoxon signed-rank test was used to observe comparisons and test hypotheses in both Study 1 and 2.

Average rank was used in each Wilcoxon analysis as Study 1 had duplicate integer differences and because some repeated differences could occur in Study 2 from finite representation with floating point values. All zero difference comparisons were removed before computing the final W statistic. Wilcoxon analysis was performed manually and then verified in R. Approximate p-values were computed in R and are provided. All Wilcoxon results tables show the

Trial	Frequency	Ear
1	80 Hz	Left
2	260 Hz	Right
3	170 Hz	Right
4	260 Hz	Left
5	80 Hz	Right
6	100 Hz	Left
7	170 Hz	Left
8	100 Hz	Right

Table 2: Hearing Test Trial Order

rejection or non-rejection for its, respective, null hypothesis given a significance level of $\alpha = 0.05$ for a paired, two-tailed Wilcoxon test. In the results table, an observed W statistic indicates if the smaller ranked sum came from the positive (+) or negative (-) values – indicating which block condition had greater effect.

5.1 Frequency Hearing Test

All participants passed the initial left-right ear hearing test. The tests played each of the following frequency in either the left or right speaker of headphones: 80 Hz, 100 Hz, 170 Hz, and 260 Hz. While the sounds were only played in one ear at a time, a total of eight trials existed to ensure that each frequency was played between the right and left ears once. The order of frequencies and ear speaker playback was staggered but constant for all participants and can be seen in Table 2. The participant then reported which ear, if any, they heard the sound. These sounds cover the general frequency range of adult speech and were a means to gauge if problems might occur in data collection for subjects with self-reported hearing loss given Study 1 and 2 used human speech in all trials. Three participants had reported minor hearing loss in one or both ears, but passed the hearing test and were included in the final data analysis.

5.2 Study 1 Results

Study 1 analysis began by preprocessing the response data collected from each block. The preprocessing was a conversion of each participant's two trials per block into an integer count. The integer count was formalized based on the three hypotheses to test. Supplemental information on overall direction guesses and accuracy can be seen in Table 3 and Figure 5.

Block	Back	Left	Right	Top	Bottom	Front
Dir Only	96.2%	9.6%	5.8%	30.8%	3.9%	90.4%
Misaligned	80.8%	0%	5.8%	23.1%	1.9%	88.5%
Aligned	84.6%	92.3%	94.2%	55.8%	7.7%	73.1%
HRTF Only	1.9%	96.2%	88.5%	28.8%	1.9%	69.2%

Table 3: Study 1 Accuracy

Figure 6 shows a boxplot for front/back response accuracy between each block before the Wilcoxon comparisons. Two sets of

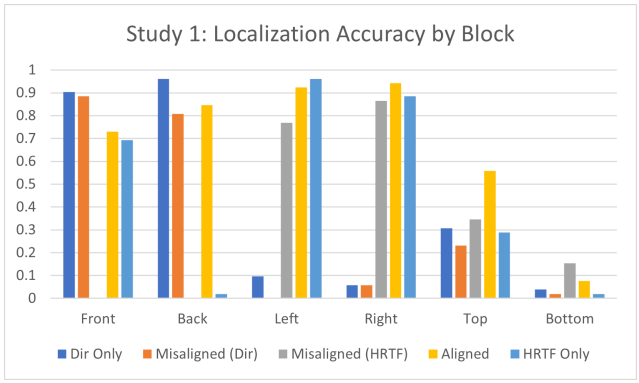

Figure 5: Total accuracy from Study 1. Correctness dictating accuracy is either based on the localization of a sound position on the fixation cube with HRTF – in the case of the Misaligned (HRTF) and HRTF Only blocks – or by the directional recording played for a participant – in the case of Dir Only and Misaligned (Dir) blocks. The Aligned block had directional sounds collocated at their corresponding locations on the fixation cube such that sound position and direction were matched.

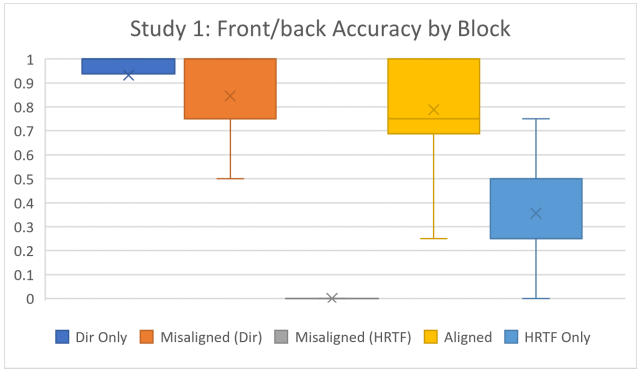

Figure 6: Front and Back localization accuracy during Study 1. Each block presented two front and two back sound conditions either based on a sounds position represented with HRTF – for blocks Misaligned (HRTF) and HRTF Only – or the directional recording played – in blocks Dir Only and Misaligned (Dir). The Aligned block used both directional recordings and spatialization with HRTF, but had position collocated with their corresponding directional recordings.

comparisons were performed to assess HRTF localization performance against the directional recordings. These comparisons defined the tests for H_1. Table 4 show the results for H_1 testing where a (-) W observed value indicates a trend in favor of monaural HRTF for front/back perception while a (+) W observed value indicates a trend in favor of directional sounds for this perception.

Figure 7 is a box plot showing counts of participant responses in the vertical plane (top, bottom, and back) that were not answered with a *front* response. Misaligned and Aligned conditions excluded trials where the directional sound was in the front direction and

H_1 Tests	W	p-value
$Misaligned(HRTF) \neq Dir\ Only$	0(+)	$p < 0.001$
$Misaligned(HRTF) \neq Misaligned(Dir)$	0(+)	$p < 0.001$
$Misaligned(HRTF) \neq Aligned$	0(+)	$p < 0.001$
$HRTF\ Only \neq Dir\ Only$	2.5(+)	$p < 0.001$
$HRTF\ Only \neq Misaligned(Dir)$	0(+)	$p < 0.001$
$HRTF\ Only \neq Aligned$	4(+)	$p < 0.001$

Table 4: Results for H_1 hypothesis tests in Study 1.

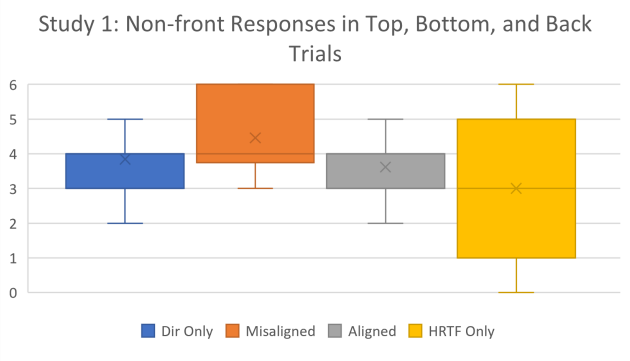

Figure 7: Non-front responses within the vertical plane of hearing. Two trials existed per block for top, bottom, and back conditions with either directional recordings or sound positions – totaling six trials in focus. Non-front responses in these conditions indicate potential localization in typically hard spatialization scenarios.

trials where the sound was located on the left or right position of the fixation box. The Misaligned (HRTF) condition is not tested as the focus is on comparing directional sound cues to a traditional monaural HRTF audio presentation. The HRTF Only block excluded trials where the monaural sound was located at the front, left, or right positions of the fixation box. These comparisons are the necessary tests to assess H_2. Table 5 show the results for H_2 testing where a (-) W observed value indicates a trend in favor of monaural HRTF while a (+) W observed value indicates a trend in favor of directional sounds.

H_2 Tests	W	p-value
$HRTF\ Only \neq DirOnly$	63(+)	≈ 0.038
$HRTF\ Only \neq Misaligned(Dir)$	24(+)	$p < 0.001$
$HRTF\ Only \neq Aligned$	64.5(+)	≈ 0.072

Table 5: Results for H_2 hypothesis tests in Study 1.

Figure 8 is a box plot showing the count of correct responses with regard to sound direction or sound location. These results are used in tests for H_3. This hypothesis is an extension, and therefore dependent, of H_1 for front and back accuracy being higher with directional sounds, but includes considerations of greater left/right localization given an HRTF. The final formulation of the hypothesis

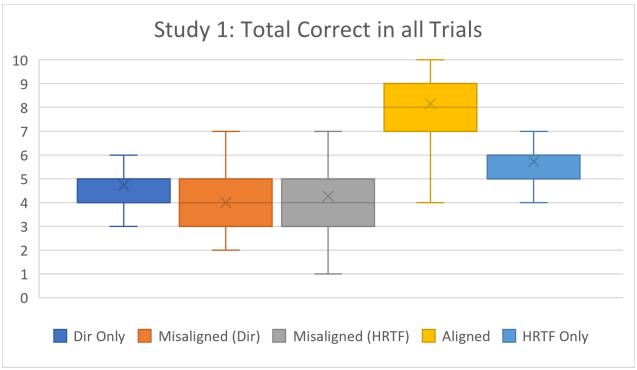

Figure 8: Total count of correct responses in Study 1. A perfect response of 12 trials was not achieved in any block, instead the maximum achieved was 10. Dir Only and Misaligned (Dir) are correct responses with regard to directional sounds, while Misaligned (HRTF) and HRTF Only are correct responses of sound locations through the use of a HRTF. The Aligned block paired directional recordings to their corresponding spatial location.

asks if coherency of vertical position and directional sounds are more meaningful for perception and sound localization. Table 6 show the results for H_3 testing where a (+) W observed value indicates a trend for greater accuracy in the Aligned block condition.

H_3 Tests	W	p-value
$DirOnly \neq Aligned$	0(+)	$p < 0.001$
$Misaligned(Dir) \neq Aligned$	1.5(+)	$p < 0.001$
$Misaligned(HRTF) \neq Aligned$	0(+)	$p < 0.001$
$HRTF\ Only \neq Aligned$	2.5(+)	$p < 0.001$

Table 6: Results for H_3 hypothesis tests in Study 1.

5.3 Study 2 Results

Study 2 analysis began by preprocessing the vector response data of a participant's position from the center of the fixation cube. This three-part vector was first normalized to represent a unit distance – therefore representing only direction (orientation) information of a participant about the fixation cube – and then had a dot product applied with the direction of the forward sound vector. The forward sound vector was a normalized vector representing the direction (orientation) for which the forward sound was located – relative to the center of the fixation cube. Figure 9 shows a box plot of the total accuracy from the two trials for each participant across each block in Study 2.

Study 2 had one hypothesis, H_4, which is the 6Dof conditions should provide a listener with greater localization ability compared to the traditional spatialization of Monaural HRTF. Three tests were conducted to compare the Monaural HRTF block against the three 6DoF blocks. Table 7 shows the results for H_4 testing where a (-) W observed value indicates a trend in favor of the Monaural HRTF block condition for localizing sound about an object in 3D space (+)

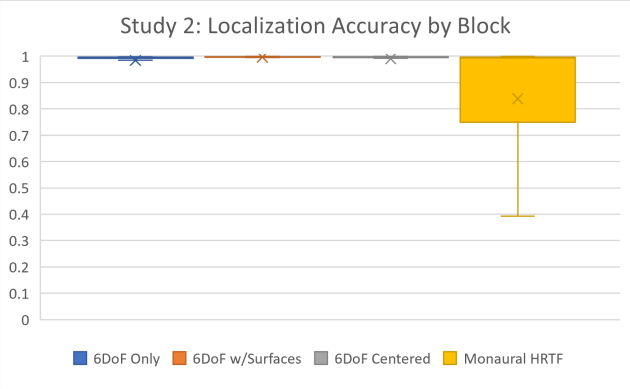

Figure 9: Localization accuracy in a mobile task between all blocks of Study 2. Participants walked around a virtual cube, augmented in their space, which was the source of our 6DoF audio method or monaural HRTF audio of human speech. Participants stopped in a position around the cube where they thought they would be facing the human speaker. IV changes were present in the 6DoF method between its three blocks, but consistency was found in localization accuracy.

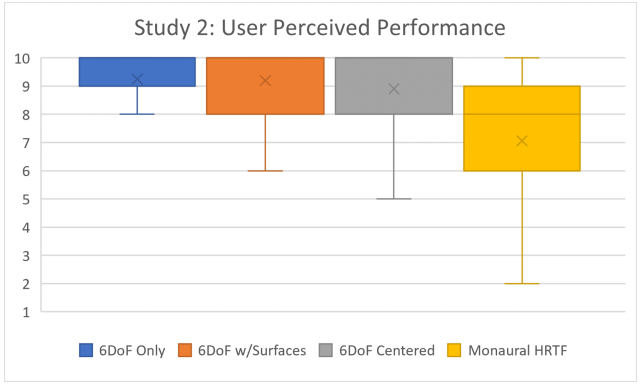

Figure 10: Perceived performance by block in Study 2. Participants reported a value on a scale of 1-10 for their level-of confidence to be standing in front of the human speaker based on the interactive audio located at the augmented fixation cube.

W observed value indicates a trend in favor of 6DoF information for this type of localization task.

H_4	W	p-value
Monaural HRTF ≠ 6DoF Only	90(+)	≈ 0.031
Monaural HRTF ≠ 6DoF w/Surfaces	34(+)	p < 0.001
MonauralHRTF ≠ 6DoF Center	56(+)	p < 0.001

Table 7: Results for H_4 hypothesis in Study 2.

Participants provided a verbal response after each trial in Study 2 to indicate how confident they were in finding the forward speaking position around the fixation cube – on a scale from 1 to 10. Figure 10 shows the average of the two trial responses from each participant. Comparing this graph to 9 helps show a correlation between tangible and subjective performance in Study 2.

6 Discussion

In this paper, we outlined an object-based rendering technique for 6DoF sound that was prototyped and tested within human subjects. The goal was to simplify the process of producing 6DoF interactive sound and to define objective metrics to test for meaningfulness of 6DoF sound for human listeners.

The results of Study 1's front/back performance indicate directional sound recordings favoring the hypothesis H_1. These results indicate directional audio is a significant treatment in aiding participants to near and far cues about the fixation cube. This is well summarized in the Misaligned (HRTF) comparisons to the other directional audio blocks. While there existed sounds positioned in the near and far locations of the fixation box, no participants responded correctly with respect to the simulated attenuation provided by the HRTF in the Misaligned (HRTF) block. Asserting that this minimal distance could be perceived with certainty appears untenable given the 1.9% accuracy of participants response to the far location (back response) in the HRTF Only block. However, overall accuracy in blocks Dir Only, Misaligned (Dir), and Aligned are at least 80.0% and validated by the Wilcoxon analysis suggesting that sounds diffracting around an object (in this case a human head) can provide audio cues that are more substantial than short-distance attenuation.

For hypothesis H_2, there is not a definite indication as strongly in favor of this hypothesis compared to the results of H_1. These hypotheses are related because of their involvement in the perception of sound within the vertical plane of a listener's cone-of-confusion. In Table 5, comparing HRTF Only to Dir Only and Misaligned blocks indicates results in favor of H_1 for these conditions, but not necessarily for HRTF Only compared to the Aligned block under the same $\alpha = 0.05$ significance level.

The analysis for H_3 was in favor of the hypothesis, suggesting coherent directional sound and HRTF spatialization is meaningful for a listener to perceive sound events about a 3D object. Table 3 is worth examining to see where some of the differences exist between each block. The accuracy for interaural cueing is clear for the left and right sounds striking around an 80-90% accuracy in the Aligned and HRTF Only blocks and can be seen in Figure 5 for the Misaligned (HRTF) block. Front/back accuracy was already shown to be favorable with the direction cues in H_1, so the information of concern is in the top/bottom directions. Bottom response accuracy was consistently the worst across all trials, but seems to have had greater accuracy in the Misaligned (HRTF) and Aligned blocks. The margins are not nearly as wide when compared to the left/right interaural performance and front/back directional sound performance. What might be of interest is the visual correlation in improved accuracy for both top and bottom directions in the Aligned block compared to the other blocks. It could be that the egocentric spectral modifications provided by the HRTF combined with directional sounds of top and bottom locations paint a familiar experience of talking with someone taller or shorter than the

participant. However, more samples would be required to see if the accuracy increase for top/bottom directions is consistent for listeners or if the results here were random.

Hypothesis H_4 uses a mathematical basis in geometry with vectors to assert a metric that expands how questions can be asked about audio objects in virtual environments as degrees of freedom increase. The results indicate favoring H_4 from an objective perspective, and that all uses of the 6DoF method were preferred to Monaural HRTF. Theoretically, participants should be able to orient themselves about the fixation box based on the attenuation of sound in the Monaural HRTF block, and while some participants did get high accuracy, the overall consistency was low. Further difficulty correlated positively with the performance in the blocks to the perceived performance response reported by participants. 6DoF only had near-perfect results across all pairs of trials for all participants, and the reported perception of performance was consistently above an 8 average. 6DoF w/Surfaces had near perfect response again combined with strong perceptions of performance sitting around an average minimum of 8. 6DoF Centered accuracy was also near perfect, but with a larger standard deviation in perceived performance – albeit the majority of responses were still high. However, Block 4 had the most varied accuracy and the widest spread of perceived performance with half the participants reporting a perceived performance less than 8. Overall, the 6DoF method(s) appear to provide participants with the most meaningful cues to find an orientation and position about a sound object – with and without the interaural cues of HRTF,

7 Limitations

As a preliminary investigation concerned with producing a simplified 6DoF method for human performance testing, some aspects of recording could be more controlled. Recordings for the experiments occurred in a room with a fixed noise floor but without substantial noise-canceling properties. A full anechoic chamber may not be necessary for recording audio with the outlined 6DoF approach, but a recording studio that eliminates reverberations would be beneficial for reproduction control of object-based audio.

While the use of a single audio interface for all six recordings minimizes the temporal differences between each recording, there is no guarantee that time differences are not slightly shifted. The recording approach outlined can be performed on a single audio interface, but if greater precision is required in analysis, then a thorough breakdown of the latency of recordings due to hardware constraints/tolerances would be worth investigating.

Similar to FOA, this 6DoF method is a first-order approach. This object-based audio mesh technique could instead have a higher number of audio recordings to form a higher-resolution triangulated mesh. In this regard, higher-order versions of this technique can be defined. Similar to studies showing greater sound locations with HOA (e.g., 3rd-order ambisonics), it could be that higher-order audio meshes could be more useful than the first-order approach prototyped in this study. This is crucial to ensure the most relevant directions of individual sound sources are accurately captured. The major benefit of the first-order approach is minimizing software latency as it fits a O(1) algorithm – but at the cost of potential audio resolution.

The range of frequency for the hearing test were chosen as human speech was the audio presented to participants. However, for interaural cues, larger frequency ranges are meaningful. The studies in this paper use both monaural and interaural playback – but future studies focused on consistent interaural cues would benefit from a standard frequency test. Additionally, calibration of the headphones with a device capable of accurate near-field testing would be beneficial. The use of a standard decibel meter should not interfere with the results as the study was within-subjects.

H_2 was not as conclusive in the anaylsis as other hypotheses. Changing to a one-tailed test or a higher significance level would change the strength of favoring H_2, but a larger sample size is ultimately necessary to see how strong directional audio cues are in assisting a listener within the vertical plane of the cone-of-confusion.

8 Future Work

Future work would greatly benefit from studies involving more complex audio meshes. It would be beneficial to analyze both the accuracy of reproduction with varying arrangements of a spherical microphone array as well as continuing to test human performance.

The 6DoF method proposed is not limited to recordings from real-world sounds. Directional audio information could be generated through complex simulation or with generative AI. The benefit of the proposed 6DoF technique is that all formats of offline produced audio can be applied to this technique for real-time interactive playback. Moreover, comparing generated samples with real-world examples could be a means to validate new techniques in computer generated sound.

The control of the studies in this paper benefit from starting with only presenting the 6DoF audio objects. However, RIR and general environmental interference is known to alter the perception of a listener. Expanding on how 6DoF objects interact in a fully simulated environment is an incremental space to further explore human performance in localization studies.

The object-based nature of the 6DoF playback could mean further subjective testing could occur between ambisonics studies and this technique. Specifically, soundscapes from 6DoF ambisonics recordings could be compared to the same environments reconstructed with 6DoF audio meshes.

9 Conclusion

In this paper, the two user studies conducted indicate that object-derived directional sounds can cue listeners into additional spatial information. Differences emerged in Study 1 indicating how a HRTF can be complimented with additional directional audio information for discerning near/far sound events in the centered-vertical plane of the cone-of-confusion. Aligned directional recordings relative to a listener also show that an upward sound direction combined with an HRTF (generalized) may be sufficient to simulate a virtual sound location vertically.

A geometry-based metric of performance was able to demonstrate an objective means to evaluate a 6DoF interactive sound for sound localization. This provides an example for tesing performance of human subjects in 6DoF audio environments. This metric, combined with a 6DoF interactive audio experience, open new

ways to ask questions about localization within psychoacoustics, accessibility, and HCI studies.

The proposed object-based 6DoF technique in this study was shown to cue listeners into spatial information with and without an HRTF. Having orientation information without interaural feedback could be a meaningful way to explore navigation by audio without the imperfection of left/right disparity within the cone-of-confusion. Additionally, outstanding questions in audio scene complexity could be explored, knowing that this technique inherently culls additional sound sources when not immediately facing a listener. Such explorations could further enhance how 6DoF audio might influence immersion in virtual worlds and XR settings.

A benefit of this technique is it minimizes the overhead of equipment and space required to produce a 6DoF interactive audio experience. This includes minimizing the expertise required with audio playback tools in order to present 6DoF sounds for interactive software. Further, the method is inherently object-based meaning these 6DoF sound objects can be reused through arbitrary instantiation in software to represent an infinite number of soundscapes – as opposed to having only one soundscape from HOA techniques. This should allow better scaling for researchers and audio engineers interested in exploring 6DoF interactive sound by reusing sound objects others have recorded, and by requiring a minimal set of audio equipment to record new 6DoF audio objects.

Acknowledgments

We would like to acknowledge the Office of Naval Research for their support with awards N00014-24-1-2214 and N00014-23-1-2298. Additionally, we would like to acknowledge the National Science Foundation for its award #2238313.

References

[1] ACM 2018. ACM Code of Ethics and Professional Conduct. https://www.acm.org/code-of-ethics.

[2] Robert Anderson, David Gallup, Jonathan T. Barron, Janne Kontkanen, Noah Snavely, Carlos Hernández, Sameer Agarwal, and Steven M. Seitz. 2016. Jump: virtual reality video. *ACM Trans. Graph.* 35, 6, Article 198 (dec 2016), 13 pages. doi:10.1145/2980179.2980257

[3] Johannes M. Arend, Sebastià V. Amengual Garí, Carl Schissler, Florian Klein, and Philip W. Robinson. 2021. Six-Degrees-of-Freedom Parametric Spatial Audio Based on One Monaural Room Impulse Response. *Journal of the Audio Engineering Society* 69, 7/8 (July 2021), 557–575. doi:10.17743/jaes.2021.0009

[4] Drew Batchelor, Kent Sharkey, David Coulter, Mike Jacobs, and Michael Satran. 2021. Render Spatial Sound Using Spatial Audio Objects. https://learn.microsoft.com/en-us/windows/win32/coreaudio/render-spatial-sound-using-spatial-audio-objects.

[5] Enda Bates and Francis Boland. 2016. Spatial Music, Virtual Reality, and 360 Media. In *2016 AES International Conference on Audio for Virtual and Augmented Reality*.

[6] Jens Blauert and Robert A. Butler. 1985. Spatial Hearing: The Psychophysics of Human Sound Localization by Jens Blauert. *The Journal of the Acoustical Society of America* 77, 1 (01 1985), 334–335. doi:10.1121/1.392109 arXiv:https://pubs.aip.org/asa/jasa/article-pdf/77/1/334/7372815/334_1_online.pdf

[7] Sam Clapp, Anne Guthrie, Jonas Braasch, and Ning Xiang. 2014. Evaluating the Accuracy of the Ambisonic Reproduction of Measured Soundfields. In *EAA Joint Symposium on Auralization and Ambisonics 2014*.

[8] Epic Games. [n. d.]. Spatialization Overview. https://docs.unrealengine.com/5.1/en-US/spatialization-overview-in-unreal-engine/.

[9] Facebook. [n. d.]. Audio v47 Reference Guide. https://developer.oculus.com/reference/audio/v47/o_v_r_audio_8h.

[10] Andrea Genovese, Gabriel Zalles, Gregory Reardon, and Agnieszka Roginska. 2018. acoustic perturbations in hrtfs measured on mixed reality headsets. *journal of the audio engineering society* P8-4 (august 2018).

[11] Michael A. Gerzon. 1973. Periphony: With-Height Sound Reproduction. *Journal of the Audio Engineering Society* 21 (feb 1973), 9 pages. Issue 1.

[12] Dukki Hong, Tae-Hyoung Lee, Yejong Joo, and Woo-Chan Park. 2017. Real-Time Sound Propagation Hardware Accelerator for Immersive Virtual Reality 3D Audio. In *Proceedings of the 21st ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games* (San Francisco, California) *(I3D '17)*. Association for Computing Machinery, New York, NY, USA, Article 20, 2 pages. doi:10.1145/3023368.3036842

[13] Thirsa Huisman, Axel Ahrens, and Ewen MacDonald. 2021. Ambisonics Sound Source Localization With Varying Amount of Visual Information in Virtual Reality. *Frontiers in Virtual Reality* 2 (2021). doi:10.3389/frvir.2021.722321

[14] Doug L. James, Jernej Barbič, and Dinesh K. Pai. 2006. Precomputed acoustic transfer: output-sensitive, accurate sound generation for geometrically complex vibration sources. *ACM Trans. Graph.* 25, 3 (jul 2006), 987–995. doi:10.1145/1141911.1141983

[15] Craig Jin, Anna Corderoy, Simon Carlile, and André van Schaik. 1999. Spectral Cues in Human Sound Localization. In *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K. Müller (Eds.), Vol. 12. MIT Press. https://proceedings.neurips.cc/paper_files/paper/1999/file/b29eed44276144e4e8103a661f9a78b7-Paper.pdf

[16] Craig Jin, Anna Corderoy, Simon Carlile, and André van Schaik. 2004. Contrasting monaural and interaural spectral cues for human sound localization. *The Journal of the Acoustical Society of America* 115, 6 (06 2004), 3124–3141. doi:10.1121/1.1736649 arXiv:https://pubs.aip.org/asa/jasa/article-pdf/115/6/3124/8094078/3124_1_online.pdf

[17] Craig T. Jin. 2020. A tutorial on immersive three-dimensional sound technologies. *Acoustical Science and Technology* 41, 1 (2020), 16–27. doi:10.1250/ast.41.16

[18] Timothy W. Leishman, Samuel D. Bellows, Claire M. Pincock, and Jennifer K. Whiting. 2021. High-resolution spherical directivity of live speech from a multiple-capture transfer function method. *The Journal of the Acoustical Society of America* 149, 3 (03 2021), 1507–1523. doi:10.1121/10.0003363 arXiv:https://pubs.aip.org/asa/jasa/article-pdf/149/3/1507/13847603/1507_1_online.pdf

[19] Song Li and Jürgen Peissig. 2020. Measurement of Head-Related Transfer Functions: A Review. *Applied Sciences* 10, 14 (2020). doi:10.3390/app10145014

[20] Leo McCormack, Nils Meyer-Kahlen, David Lou Alon, Zamir Ben-Hur, Sebastià V. Amengual Garí, and Philip Robinson. 2023. Six-Degrees-of-Freedom Binaural Reproduction of Head-Worn Microphone Array Capture. *Journal of the Audio Engineering Society* 71, 10 (2023), 638–649.

[21] Leo McCormack, Archontis Politis, Thomas McKenzie, Christoph Hold, and Ville Pulkki. 2022. Object-Based Six-Degrees-of-Freedom Rendering of Sound Scenes Captured with Multiple Ambisonic Receivers. *Journal of the Audio Engineering Society* 70, 5 (2022), 355–372.

[22] Nikolaos Moustakas, Emmanouel Rovithis, Konstantinos Vogklis, and Andreas Floros. 2021. Adaptive Audio Mixing for Enhancing Immersion in Augmented Reality Audio Games. In *Companion Publication of the 2020 International Conference on Multimodal Interaction* (Virtual Event, Netherlands) *(ICMI '20 Companion)*. Association for Computing Machinery, New York, NY, USA, 220–227. doi:10.1145/3395035.3425325

[23] Orhun Olgun, Ege Erdem, and Hüseyin Hacıhabiboğlu. 2021. Rotation Calibration of Rigid Spherical Microphone Arrays for Multi-perspective 6DoF Audio Recordings. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. 1–7. doi:10.1109/I3DA48870.2021.9610848

[24] OpenAL developers. [n. d.]. OpenAL Soft. https://www.openal-soft.org/.

[25] Eduardo Patricio, Andrzej Ruminski, Adam Kuklasinski, Lukasz Januszkiewicz, and Tomasz Zernicki. 2019. Toward Six Degrees of Freedom Audio Recording and Playback Using Multiple Ambisonics Sound Fields. *AES Convention* (2019).

[26] Justin Patterson and Hyunkook Lee (Eds.). 2021. *3D Audio* (1st ed.). Routledge, United States. doi:10.4324/9780429491214 Publisher Copyright: © 2022 selection and editorial matter, Justin Paterson and Hyunkook Lee; individual chapters, the contributors. Copyright: Copyright 2021 Elsevier B.V., All rights reserved..

[27] Axel Plinge, Sebastian Schlecht, Oliver Thiergart, Thomas Robotham, Olli Rummukainen, and Emanuë Habets. 2018. Six-degrees-of-freedom binaural audio Reproduction of first-order ambisonics with distance information. In *2018 AES International Conference on Audio for Virtual and Augmented Reality*.

[28] Iana Podkosova, Michael Urbanek, and Hannes Kaufmann. 2016. A Hybrid Sound Model for 3D Audio Games with Real Walking. In *Proceedings of the 29th International Conference on Computer Animation and Social Agents* (Geneva, Switzerland) *(CASA '16)*. Association for Computing Machinery, New York, NY, USA, 189–192. doi:10.1145/2915926.2915948

[29] Paul Power, Chris Dunn, William J. Davies, and J. Hirst. 2013. *Localisation of Elevated Sources in Higher-Order Ambisonics*. Technical Report. British Broadcasting Corporation.

[30] Schuyler R. Quackenbush and Jürgen Herre. 2021. MPEG Standards for Compressed Representation of Immersive Audio. *Proc. IEEE* 109, 9 (2021), 1578–1589. doi:10.1109/JPROC.2021.3075390

[31] Jaime Sánchez, Mauricio Sáenz, Alvaro Pascual-Leone, and Lotfi Merabet. 2010. Navigation for the Blind through Audio-Based Virtual Environments. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (Atlanta, Georgia,

USA) *(CHI EA '10)*. Association for Computing Machinery, New York, NY, USA, 3409–3414. doi:10.1145/1753846.1753993

[32] Konstantin Semionov and Iain McGregor. 2020. Effect of various spatial auditory cues on the perception of threat in a first-person shooter video game. In *Proceedings of the 15th International Audio Mostly Conference* (Graz, Austria) *(AM '20)*. Association for Computing Machinery, New York, NY, USA, 22–29. doi:10.1145/3411109.3411119

[33] Unity. [n. d.]. Audio Spatializer SDK. https://docs.unity3d.com/Manual/AudioSpatializerSDK.html.

[34] Valve Corporation. [n. d.]. Immersive Audio Solutions. https://valvesoftware.github.io/steam-audio/.

[35] DeLiang Wang and Guy J. Brown. 2006. *Binaural Sound Localization*. 147–185. doi:10.1109/9780470043387.ch5

[36] Elizabeth M. Wenzel, Frederic L. Wightman, and Scott H. Foster. 1988. Development of a three-dimensional auditory display system. *SIGCHI Bull.* 20, 2 (oct 1988), 52–57. doi:10.1145/54386.54405

[37] Elizabeth M. Wenzel, Frederic L. Wightman, and Scott H. Foster. 1988. A Virtual Display System for Conveying Three-Dimensional Acoustic Information. *Proceedings of the Human Factors Society Annual Meeting* 32, 2 (1988), 86–90. doi:10.1177/154193128803200218 arXiv:https://doi.org/10.1177/154193128803200218

[38] Frederic L. Wightman and Doris J. Kistler. 1989. Headphone simulation of free-field listening. I: Stimulus synthesis. *The Journal of the Acoustical Society of America* 85, 2 (02 1989), 858–867. doi:10.1121/1.397557 arXiv:https://pubs.aip.org/asa/jasa/article-pdf/85/2/858/11398484/858_1_online.pdf

[39] Jing Yang, Felix Pfreundtner, Amit Barde, Kurt Heutschi, and Gábor Sörös. 2020. Fast synthesis of perceptually adequate room impulse responses from ultrasonic measurements. In *Proceedings of the 15th International Audio Mostly Conference* (Graz, Austria) *(AM '20)*. Association for Computing Machinery, New York, NY, USA, 53–60. doi:10.1145/3411109.3412300