THESIS

THE SHAPE OF SOUND: RENDERING INTERACTIVE SIX-DEGREES-OF-FREEDOM AUDIO IN SOFTWARE

Submitted by

Daniel Rehberg

Department of Computer Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2024

Master's Committee:

Advisor: Francisco Raul Ortega

Sanjay Rajopadhye
Laura Malinin

ABSTRACT

THE SHAPE OF SOUND: RENDERING INTERACTIVE SIX-DEGREES-OF-FREEDOM AUDIO IN SOFTWARE

Six-degrees-of-freedom (6DoF) audio is an area of growing interest in interactive software, but it has faced several challenges: it does not easily conform to object-based rendering when achieved with arrays of ambisonics microphones; prior studies rely on subjective metrics 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); and the ambisonics technique requires specialized equipment and recording space, as well as audio engineering expertise for setup and calibration to work properly. These factors limit the accessibility of 6DoF audio to be implemented in research experiments or within commercial products like videogames. My work has involved taking an interdisciplinary approach to design, prototype, and validate (with human subjects) an inherently object-based 6DoF rendering method. This method exploits computational geometry techniques and follows a rendering paradigm inspired by the programmable graphics pipeline to create 3D audio meshes which can be transformed in real time to dynamically render monaural audio samples – meaning the output of the method can still be input into contemporary audio filtering and spatialization functions/tools, like a head-related transfer function. This work includes two studies performed with human subjects as well as a breakdown of the rendering method and its prototype implementation. The results of the human-subject studies indicate clear advantages to localizing a spatial sound in 3D space compared to the contemporary three-degrees-of-freedom approach.

ACKNOWLEDGEMENTS

During my bout in graduate school, I have gotten to interact with excellent professors and students that I would like to acknowledge. These interactions, undoubtedly, helped shape the work in my thesis. I would like to thank Dr. Francisco Ortega for supporting my pursuit of this topic on audio rendering and interaction. This allowed me to continue garnering knowledge from disparate fields to work on a multidisciplinary problem. I would like to thank my committee members Dr. Sanjay Rajopadhye and Dr. Laura Malinin – Sanjay has always been supportive in my academic pursuits and Laura oversaw the first augmented reality research project I was involved in. I am grateful for you both taking the time to be on my committee. Thank you to Dr. Anil Ufuk Batmaz and Dr. Adam Coler for assisting me on my written research. I would like to thank Amelia Warden for advancing my understanding of experimental design within Human Factors. Your perspectives and skills have been immensely practical for my work with human subjects. I would also like to thank Dr. Shrideep Pallickara and Dr. Nathaniel Blanchard for giving me opportunities to grow my presentation and public speaking skills through teaching – I have become a more competent and comfortable presenter as a result. I want to thank my friends and peers who showed support and interest in my research inquiries, both inside academia and in the skateboarding community. I want to thank my family for showing ongoing support in my interests, and my Mother's initial support of my exploration in Interactive Computer Graphics from a young age. Lastly, I would like to thank Dr. Brandon Touchet for being an ever-available mentor – your questions allowed me to continue exploring a playground of ideas and kept me focused as I pursued a graduate degree. Thank you all.

DEDICATION

This thesis is dedicated to my wife, Zoë, thank you for your perpetual support in my endeavors.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

What it means to have six-degrees-of-freedom (6DoF) audio is currently an open question, with a comprehensive review of practices from the last decade summarized by Garath Llewellyn and Justin Paterson [1]. Part of the questions around 6DoF sound stem from a required innovation to provide this type of dynamic experience to a listener using software. Current compute power tentatively exists to produce real-time 6DoF sound, but it is important to question if attempting to make this interactive sound is even meaningful for a listener. This section will cover a few things:

- How do humans perceive “3D sounds?”
  - Why is the head-related transfer function important for interactive software?
  - What problems exist with human hearing?
- Why is ambisonics emerging in XR and how it can improve research?
  - How does ambisonics connect to 6DoF sound?
- What desirable features are missing with current 6DoF sound?

A novel audio rendering method and the human-subject studies in Chapter 3 and Chapter 4, respectively, are motivated by this background and the literature review found in Chapter 2.

1.1 Background

Interactive audio utilizes spatialization techniques to inform listeners of sound positions. Two predominant choices are used to either account for individual body characteristics of a listener or the characteristics of a room for ambient sound. Use of head-related

transfer functions (HRTF) in which a monaural sound clip has been modified to provide localization cues and ambisonics, putting a listener at a fixed location within a sound field. HRTF is well studied as its constituent parts define the primary audio cues for human spatial perception [2–5]. Ambisonics has grown recently for its use in virtual reality (VR) enhancing experiences such as 360-degree videos [6, 7], but prior to this was used to record sounds for periphony spaces [8]. Ambisonics interest has expanded beyond stationary VR to more general interactive surround sound and extended reality (XR) applications given the potential to produce 6DoF audio when utilizing higher-order ambisonics (HOA) [9, 10].

An existing question is posed as to what 6DoF audio should represent to a listener compared to the original three-degrees-of-freedom (3DoF) experiences defined by HRTF and ordinary ambisonics [1]. To explore human performance with rotatable sounds in isolation, this study considers individual 6DoF qualities of a sound source by defining and utilizing a method to reconstruct a 3D sound surface for object-based rendering. These rendered features describe what changes should be perceived given audio that is dynamic to both translations and orientation about a listener. However, merely developing a rendering method is insufficient; instead, 6DoF audio should be tested for its objective effects on human perception and performance – e.g., sound guidance. Without a justification to the tool, it is difficult to suggest use cases for real-time applications, as there is computational overhead to rendering sound. For that reason, it is important to understand current practices and how human hearing is described and translated for software – thus defining a merit for 6DoF audio tools and their use cases.

HRTF has been used as a standard approach to provide spatial audio for interactive 3D experiences [11–16] including human-computer interaction and psychoacoustic studies [1, 4]. It is an egocentric technique concerned with modifying a monaural sound based on the relative position of the sound about a listener [9]. Interaural time (ITD) and level (ILD) through panning is an earlier, computationally less expensive approach, but

lacks spectral modifications from body derived (e.g., pinnae, head, and torso) interference [5, 17] – instead focusing on the time (ITD) sound is heard between the ears or its amplitude (ILD). Such studies indicate that spectral cues regarding the head-torso configuration and pinnae shape are meaningful to human perception – specifically more so than with panning alone. Studies of HRTF exclusively indicate that there is a cone-of-confusion for which humans have a hard time discerning sound locations when limited interaural disparity is present [2–5, 18]. This cone-of-confusion is one such place to ask if 6DoF audio could be beneficial, i.e., can more audio information assist a listener in locating a sound inside the interaural confusion space?

Ambisonics and periphony localization studies use more audio information than HRTF. Ambisonics differs from the playback of monaural audio because an entire soundscape is captured from a coincident microphone array [8]. This means several recordings simultaneously capture different directions of a soundscape from a shared and fixed location. Each audio recording can be mapped directly to a single physical sound channel in a periphony or general surround sound system [8, 9, 19]. Playing HOA recordings in periphony studies has shown humans can more accurately pinpoint the direction of sounds compared to lower/first-order ambisonics (FOA), in both the horizontal [20] and the vertical plane [21]. This capability helps overcome some of the issues associated with the cone-of-confusion, as humans can cue into an initial sound wave, a phenomenon known as the precedent effect [22]. Ideal reproduction values for HOA resolution have been studied [23] and are reinforced by human subject experiments in the periphony studies [20,21]. Reproduction of HOA for periphony studies requires large spaces, hence limiting the number of researchers that can explore HOA as a tool compared to using simulated HRTF in devices like desktop computers or XR HMDs.

With commercial availability of headtracking HMDs, ambisonics/periphony experiences are now being simulated in VR and, more broadly, XR [1,9]. The desire to have headtracking technology for interactive audio experiences can be found as far back as the 1980s

within NASA research [24, 25]. Because VR can be a stationary experience, ambisonics recordings can be used to dynamically switch which recording to play for a listener as they rotate in place. This works because an ambisonics device, again, records a soundscape from multiple directions in a fixed location. In software, this means that a user's forward vector can decide how to change/interpolate between directional recordings. This interaction allows a rotational experience for listeners from a fixed location in 3D space defining a 3DoF audio experience. The use of headtracking and XR devices has sparked interest in 6DoF audio experiences, currently pursued through HOA tools [1].

HOA approaches have been explored over the past few years to enable both translational and rotational degrees of freedom. This includes manually adjusting amplitude gain in FOA rotational VR experiences [10] and by using multiple HOA devices to map multiple listener perspective throughout a room [26–28]. Some 6DoF ambisonics studies have even occurred with human subjects [10, 26, 28]. While these techniques can be simulated in XR, they still have substantial overhead in equipment and space to acquire initial recordings – again, limiting the number of researchers that can perform human-subject studies with ambisonics. The current 6DoF studies also rely on a subjective metric [10, 26] compared to HRTF and periphony studies that have objective data – meaning HOA 6DoF tools have not been fully realized for human performance. Lastly, recording an entire soundstage does not lead to easy object-based rendering – an approach that allows media elements to be arbitrarily instantiated in software.

Our study demonstrates an alternative recording and 6DoF rendering method to current HOA practices. The practice defined requires substantially lower overhead in both space, equipment, and technical understanding to help maximize the number of researchers and game developers that could utilize 6DoF sound. Our prototype of this technique is demonstrated in a human-subject experiment using tasks with objective metrics to help illustrate why 6DoF audio can be a useful tool to extend psychoacoustic, accessibility, and HCI studies. The tool ensures that sounds are inherently usable for object-based

rendering, and can work alongside existing HRTF applications by dynamically producing a monaural sound that can then be transformed by usual HRTF application programming interface (API) calls. The experiments are inspired by the gains of additional audio information found in ambisonics periphony studies, and are compared to standard monaural HRTF to verify performance differences – such performance differences were found favoring the proposed 6DoF technique.

Chapter 2

Related Work

2.1 Limits of HRTF

While an HRTF describes the qualities of egocentric listener perspective when hearing sounds, it is not without notable limits. One such limitation occurs when monaural HRTF rendering is used in a navigation task. Larsen et al. [17] performed a simulated navigation study in a virtual environment where human subjects (48 total) were guided along a piecewise linear path using sound. Two versions of the virtual environment were used, one with FMOD audio – a panning-based audio system (only ILD and ITD) – and the other using Diesel Power Mobile – an audio system with a full generalized HRTF. Endpoints to the current linear segment had a 3D sound that would play to help guide the participant to the end of the current linear path. Once they reached the endpoint (using a keyboard and mouse to move), the current 3D sound would be disabled, and the next 3D sound would begin playing to guide them to the next destination. Once all destinations were reached the trial ended. The positions of the participants in this virtual space were tracked continuously to see how far they deviated from the ideal linear paths. While the HRTF method illustrated fewer deviations over the panning only audio, there are still major discrepancies in the deviations found when HRTF was used. This illustrates an example of the limitations of human hearing when there is not a strong discrepancy in left/right ear disparity for ILD/ITD or spectral modifications.

While the navigation deviations occurred with both the panning and HRTF from Larsen et al. [17], it could be argued that this is a result of not having a specialized HRTF for each listener. However, the cone-of-confusion is an ongoing blight in HRTF studies seeking to maximize human localization performance in the front/back and vertical domains [3, 17, 18, 29–32]. This is an important aspect of sound localization because audio nav-

igation could be used for the situations where the visual domain is overwhelmed [33] or fixated on an important task like driving [34]. Exploiting what is known about egocentric perception can mean developing audio tools for visually impaired individuals to navigate new places [35] or to navigate GUIs like web browsers [36, 37]. However, the limitations of HRTF navigation and localizing vertical sounds are likely to crop up regardless of the real or virtual domain suggesting other means of audio cues are relevant to explore for localization tasks.

2.2 Object Shape and Accurate Sound Waves

An initial motivation to replicate 6DoF sound arises from understanding the fundamental mechanics of audio – sound travels through space as mechanical waves in three dimensions. This implies that if observers are positioned at various locations around a sound event, they might perceive slightly different sounds due to the typical wave interference interactions explained by physics (diffraction, reflection, refraction, diffusion). In audio fields this can be referred to as an impulse response (IR) and is commonly discussed with regards to a room IR (RIR) – where sound propagation can be simulated to map out an entire allocentric space [38]. The interactive changes with RIR have been shown to provide meaningful cues to listeners in virtual [35, 39–41] and augmented [42–44] spaces – where RIR causes reflections and reverberations that indicate to a listener their proximity to walls and objects.

The RIR refers to allocentric sound interactions, yet theoretically, a sound should interact with the material it originates from before propagating in a 3D environment. A visualization and example closely related to this type of sound production can be found in James et al. [45], where the shape of the object altered the outgoing sound waves that were originally sourced from that object in a simulation. Their work visually demonstrates the non-uniform nature of sound waves as they interact with a surface – illustrating

that a sound event would be perceived differently between observers located at different orientations about the sound source even before allocentric or RIR interferences occur.

An HRTF's spectral modifications are defined by a head-related IR (HRIR) [4, 5], but again a simulated HRTF uses a constant monaural sound [9, 29] sample and lacks the additional audio spectral changes provided by an object's own interference to its sound events let alone a RIR. A research benefit of monaural HRTF is that it serves as a tool in software, isolating the information necessary for testing human performance. In contrast, Ambisonics records an entire real-world soundscape, which will inherently contain both object-source interference and allocentric interference like RIRs, but are not easily decoupled to study a listener's perception of an object-source IR in isolation. That is, HRTF in software can utilize object-based rendering, whereas ambisonics recordings of a soundscape can not. The use of object-based rendering provides researchers, audio engineers, fx artists, game designers, and the like with more control over their software-based experiments.

An exception for object-based 6DoF audio exists in McCormack et al. [28], which utilizes beamformers to isolate sounds that can then be subtracted from the rest of the environment. The researchers were able to exploit the characteristics of beamformers, which filter wave signals, to attempt to isolate HOA 6DoF sound events for object-based rendering. This technique involves filtering out pieces of information that could be useful in allowing portions of sound events to be removed from a prerecorded HOA soundscape. Possessing isolated 6DoF sound objects would allow object-source audio interference to be used for human-subject studies. The research indicates a "realism" measure as perceived by human participants in a VR study, but the rendering method lacks a source direction component for isolated sounds that may convey less realism to the 3D surface audio representation [28] and the precedent effect [22, 39] where humans can identify the origins of sounds.

2.3 Ambisonics, Periphony, and Sound Localization

Periphony studies with interactive ambisonics recordings preserve the direction sourcing of sound information given the independent set of speaker arrays uniformly spaced around a listener. Moreover, the order of ambisonics recordings, precedent effect, and interaural cueing all affect how a listener will localize a sound. Clapp et al. [23] analyzed localization errors using human-derived estimates of interaural time and distance differences as defined in HRTF, but with ambisonics recordings. In their work, they focused on the alteration of estimated interaural cues as they increased from FOA to varying orders of HOA. Their work indicated a reduction in error for binaural interpretation of interaural differences as the ambisonics order increased for lateral (horizontal) localization. This was not performed on human subjects and instead represents a quantification of timing (ms) and amplitude (dB) error that a listener may, theoretically, perceive between ambisonics orders.

Actual human-performance results were found by Huisman et al. [20] as they tested 21 participants with a blindfold and an HMD within a room of loudspeakers. Their results indicated that localization errors were significantly lower when moving beyond FOA [20]. These errors are angle deviations when identifying the location of a sound within a -90 to 90 degree horizontal space. This increase in order aligns with the reduced error evidence and suggestion of interaural performance posed by Clapp et al. [23]. However, the significance of Clapp et al. [23] error estimates was not as clear without human subjects – compared to Huisman et al. [20], whose work indicated that an ideal amount of cueing information exists at 3rd-order ambisonics as only marginal increases existed with 5th and 11th orders [20].

An earlier study by Power et al. [21] confirms the effectiveness of 3rd-order ambisonics over lower orders. Moreover, their work indicates that 3rd-order ambisonics is a starting place in HOA to begin observing benefits in both vertical and horizontal localization abilities within human subjects. This experiment also utilized a soundstage with a series of

loudspeakers, but in a spherical arrangement to test for localization of sound sources in the vertical plane within a range of -35 to 35 degree. The periphony arrangement also covered a full 360 degrees horizontally about participants. Their results indicate that representation of real-world sounds reproduced in a controlled environment can provide vertical cues to listeners when HOA is utilized. The use of HOA over FOA indicates that the accuracy of these cues could be baked in the spectral nuances of captured sound rather than just the placement of the speakers in the room.

Each of these studies indicates that more audio information can be useful to increase the perception of a sound location, including in the vertical domain and around the cone-of-confusion. It also illustrates the importance of conducting human-subject studies to empirically identify the upper limits of usefulness with audio tools. Huisman et al. [20] and Clapp et al. [23] both point to better localization as order increases with ambisonics, but studies with human subjects help indicate where the growth of benefits by order start to decrease.

2.4 6DoF Developments with Human Subjects

The combination of HRTF and at least 3rd-order ambisonics are tools that have been shown to reveal how humans can localize sounds. However, the 6DoF HOA reproductions do not currently offer ample objective metrics to further illustrate these qualities in interactive software.

Plinge et al. [10] demonstrate an approach to allow slight translations within a simulated VR environment using FOA with headphones. They allow the listener to be disjoint from the location in the scene where the ambisonics device would have been located during an initial recording. Translations were allowed by altering the direction-of-arrival (DoA) of sounds from the ambisonics recording set. Their experiment included three independent variables: a baseline synthesized object-based rendering of the original soundscape, a normal 3DoF FOA playback, and a “6DoF” playback allowed with the DoA enabled

translations. The sounds in the experiment were typical sound events heard in daily life. A MUSHRA scoring system was used to indicate how 3DoF and 6DoF compared to baseline. The 6DoF playback was closer to the baseline than the 3DoF. These results promote a motion towards 6DoF sound, but also indicate that synthetized object-based rendering might offer a more natural experience over an encapsulating soundscape.

Patricio et al. [26] also devised and performed a MUSHRA test for "6DoF" experiences using sounds heard from normal daily routines. The researchers utilized an array of ambisonics devices throughout a room to capture real-world perspectives of disparate parts of a soundscape. The MUSHRA ratings included two tests for FOA and HOA at 3rd-order ambisonics. Participants were asked to consider how natural sound localizations perceived and how natural were distant sounds as users moved around. The 3rd order system ranked close to 80 (out of 100 to the reference) on average in both tests whereas FOA was closer to 70 on average. While this study indicates arrays of HOA devices as a potential means to the production of 6DoF sound, it points out the issue of doing so – requiring planar layers of HOA devices for true 6DoF. That is, one plane of HOA devices provides more than 3DoF, but ultimately requires stacks of these planes to capture observable samples of a soundscape's entire volume to achieve full 6DoF.

These 6DoF studies [10, 26, 28] all indicate that ambisonics can be a driver for rotational and translational degrees of freedom. However, each study uses subjective measures to determine how natural an interactive space feels to determine the merits of 6DoF audio rendering – qualitative comparisons are meaningful but lack consideration of human performance metrics (additional objective metrics), which could aid in discovering use cases for 6DoF audio interaction. Additionally, ambisonics and arrays of HOA devices pose a large overhead in space and cost, limiting the number of researchers or interactive designers that can utilize these tools for 6DoF sound. Our study is motivated to pursue both the human performance advantages that might be revealed when given

additional degrees-of-freedom with sound, but also to build a simplified recording and rendering model for 6DoF interactive audio.

Chapter 3

6DoF with Triangulated Audio Meshes

For this study, a novel approach to recording and playback was taken to simplify 6DoF interactive playback and to test human-performance with 6DoF audio. This approach involved recording sounds in an outside-in fashion to isolate a sound signature inside a microphone array rather than capture an entire soundscape. Our technique was designed to overcome some disadvantages of HOA by having a single sound event captured from multiple angles. As described in this chapter, isolating sounds with outside-in microphones means that the recordings can be used in an object-based rendering approach. The set of recordings were used to create a 3D local coordinate system for the sound objects by placing microphones around the recording subject as if they existed on traditional Euclidean axes: two recordings per axis where one microphone was recording from the positive direction and another from the negative direction totaling six recordings per object. Visually, this outside-in recording would have all microphones axis-aligned and facing the origin. Each microphone represents a unique listener's perspective about the recording subject. These recording samples were then used to create a connected octahedron as a triangulated mesh. This shape was exploited to determine what final sound to play for a participant in $O(1)$ time.

For more explicit information, included in Sections 3.1 and 3.1.1 are the theoretical bases and for the audio rendering method, outside of the prototype implementation used for the human-subject studies.

3.1 Constraints and Objectives

The domain of Computer Graphics has already built practical solutions in order to not just provide visual objects with 6DoF transformations, but specifically alter how surfaces and pixels look only after the coordinate transformations. This effectively means there are

two spatial relations to consider: how does the object fit into a coordinate system, and then how does a user perceive that object within the shared (global) coordinate system? The first question is not concerned with delivering information to an end user until it is in its transformed configuration because that reveals the most relevant information for the user once visuals have been produced. The first stage relies on a shape represented by a graph – a triangulated mesh – which is something that current object-based audio representations do not use.

A concise architecture is already outlined to do this very thing within the domain of graphics APIs through a process called the graphics rendering pipeline [46]. Figure 3.1 is an abridged version of this pipeline to highlight the most relevant information to consider when adapting it to audio. My goal is to use this as a model to be adapted to audio because it takes into account the manner in which an object is placed in a coordinate system to dynamically determine what information is (at any given moment) relevant to deliver to an observer. The relevant information is what pixels should an observer see (Rasterization), which has a prerequisite step, what surfaces are facing an observer (Vertex Shader)? It is only after these steps that the pipeline becomes useful to color pixels (Fragment Shader), either simplistically with a few textures and Lambert's Cosine Law or even with complex ray-tracing lighting models. But, this requires knowing the shape of objects first.

The closest representation to a shape for sound is found in volumetric audio rendering [47]. This solves a problem when using HRTF to spatialize sound sources. If a graphical element, say a river, extends over a large region but a monaural sound for the river exists only at the local origin of the river's mesh, then a listener could be immediately adjacent to the river but the river could sound far away depending on their distance to the river's local origin given an inverse-square attenuation. Complex solutions to this problem could be computing optimal positions for placement of a monaural sound in multiple locations [47], or more classically a single monaural sound could be moved along the volume

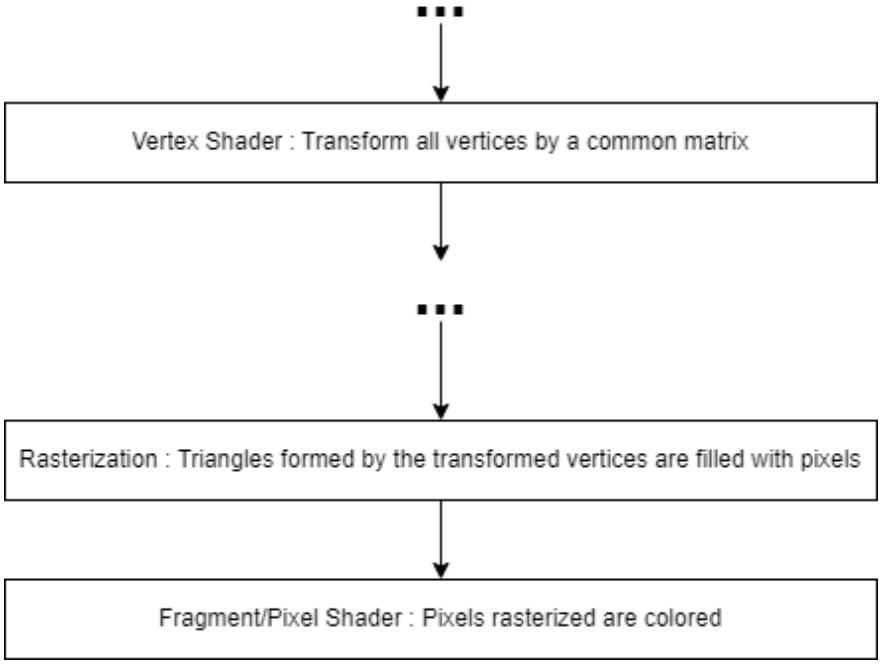

Figure 3.1: Reduced representation of a flowchart describing the graphics pipeline to render computer graphics.

boundaries of a graphical object to be placed in a location closest to the listener's position at all times – by using an algorithm like Gilbert-Johnson-Keerthi (GJK) [48] to find the closest point between the boundary volume and the listener.

This volumetric representation tells us nothing about the shape of the sound object from a sound source. Its primary concern is similar to HRTF and ambisonics – how should a sound be perceived for a listener as it exists inside a 3D space. This is analogous to skipping the transformation and rasterization stage of the graphics pipeline and trying to generate a coherent image – which is possible with modern generative AI, but is not well tuned for accuracy or coherency in real-time and even offline generation.

The shape of the sound is ultimately important because depending on a location about a sound source, a listener will hear a shared sound event differently – just as an observer on the opposite side of an object will visually see different features of the common object. This is illustrated in Figure 3.2 which is a recording of human speech from six different locations around the speaker, where the distinctly colored waveforms are overlapped by

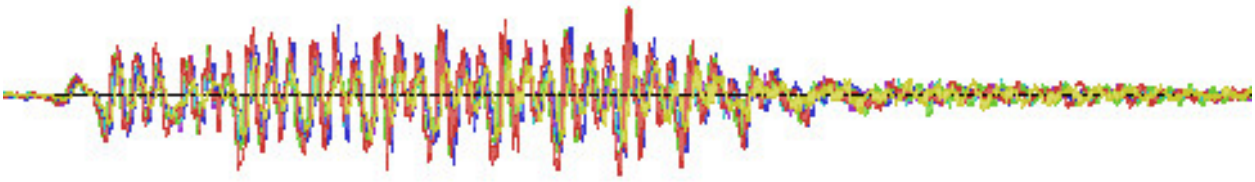

Figure 3.2: Waveforms from six simultaneous recordings around a speaker's head at a common radius from the speaker's mouth, shown in distinct colors, demonstrate that a sound source generated from an object is flavored by the object it emerges from.

a time parameter. The different peaks and valleys of each waveform are a direct result of spectral changes as the sound waves diffracted around the speaker's head.

To retrofit the stages of the graphics pipeline to audio – to yield an object-based basis for 6DoF interaction – it is first important to understand how to represent an audio object as a 3D shape in a graph. Once we have a 3D shape, a means of generating a single monaural sample dynamically (akin to rasterization of pixels) can occur so that existing filters, RIRs, and HRTFs and the like from existing APIs and hardware accelerators can still be used and provide the final "colored" sound to a listener.

3.1.1 Shape of Sound as a Time Parametric

In an ideal empty space, the shape of a sound heard is some wave as perceived from some orientation about the sound source. Amplitude of the waveform will attenuate with distance, but that can be simulated using an HRTF, so a primary concern is not about the scale of the waveform but how it varies at a consistent time based on an orientation about the sound source. However, to keep the scale consistent – i.e., how much natural attenuation occurs – an important constraint is to consider observations of a sound event at a common distance from that sound event. This effectively means the shape of concern is a hypersphere (for simulating audio in arbitrary dimensions), where every possible waveform generated by a sound event could be heard on the hull of this hypersphere.

If an infinite number of microphones could be arranged about a sound source at a common radius, then a graph of an infinite number of vertices storing audio recordings could be created. This describes an outside-in recording approach which lets each recording converge towards a single sound event. This is in contrast to an ambisonics recording which is inside-out which is the perspective of hearing a divergent set of sounds from a fixed location. This outside-in quality described means what is recorded will be of a single sound event, and therefore will fit an object-based rendering model.

Given this set of recordings from an infinite number of microphones, the relevant sound an observer would hear is the recording closest to them in this graph, just as a microphone from the original recording is an observer that only hears the waveform from its location about the sound event. This single vertex would then be the monaural sample to use in existing HRTF implementations to provide egocentric perspective to the sound – including distance attenuation. This graph can be rotated and translated, meaning it represents a sound source that has 6DoF potential.

Of course, this is not practical as microphones cannot be infinitely arranged about a sound source to generate a perfect observation hypersphere. Instead we need to rely on finite-element method in the same way that it is utilized for computer graphics by generating piecewise-linear surfaces to form a shape – such as triangulated meshes. A triangulated mesh is an ideal choice because existing practices have been well defined for interpolating information based on proximity to a triangle's vertices.

This interpolation comes from computing a weighting coefficient from barycentric coordinates as first defined in the 19^{th} Century by August Möbius. This uniform interpolation over a triangle allows us to take fractions of vertex attributes – say position, color, or (proposed here) audio recordings – and mix them together for any region within the triangle. This is effectively how continuous information is generated over discretized shapes using Gouraud shading [49] on a per-pixel level. Barycentric coordinates work for interpolation because the sum of coefficient weights add up to 1.0 as shown in (3.1).

$$\sum_{\mathbf{i=1}}^{\mathbf{3}} \lambda_{\mathbf{i}} = \mathbf{1.0} \tag{3.1}$$

3.1.2 Finite-element Audio Sphere

In this finite-element approach we can compute the barycentric weights of a listener's position within a triangle boundary to generate a monaural sample dynamically between three different audio recordings. By first computing the barycentric weights – given a function that takes in a position L and vertices of a triangle A, B, C with positions and audio samples shown in (3.2), and finally summing the weighted vertices to generate an audio sample S as described in (3.3).

$$\lambda_{\mathbf{1}}, \lambda_{\mathbf{2}}, \lambda_{\mathbf{3}} = \mathbf{barycentricWeights}(L, A_{position}, B_{position}, C_{position}) \tag{3.2}$$

$$\mathbf{S} = \lambda_{\mathbf{1}} \times \mathbf{A_{audio}} + \lambda_{\mathbf{2}} \times \mathbf{B_{audio}} + \lambda_{\mathbf{3}} \times \mathbf{C_{audio}} \tag{3.3}$$

To have a good distribution of audio samples, an ideal reduced triangulated mesh would take the form of an icosphere. As the number of vertices in the icosphere approach infinity, we have the unobtainable audio hypersphere with infinite observation samples. Within the finite-element perspective, as we reduce the number of vertices in the icosphere we will obtain less accurate interpolated monaural samples as the area of a triangle will increase in size, but having microphones places in an icosphere arrangement ensures uniform distance samples are present between the voronoi regions of microphones – as the triangles will have the same area. Three examples of different density icosphers are shown in Figure 3.3.

We can expect that summing together normalized weights of adjacent audio samples will reconstruct an accurate in-between sample. This works because of a coherency between the samples as waves are continuous. The accuracy of this interpolation is,

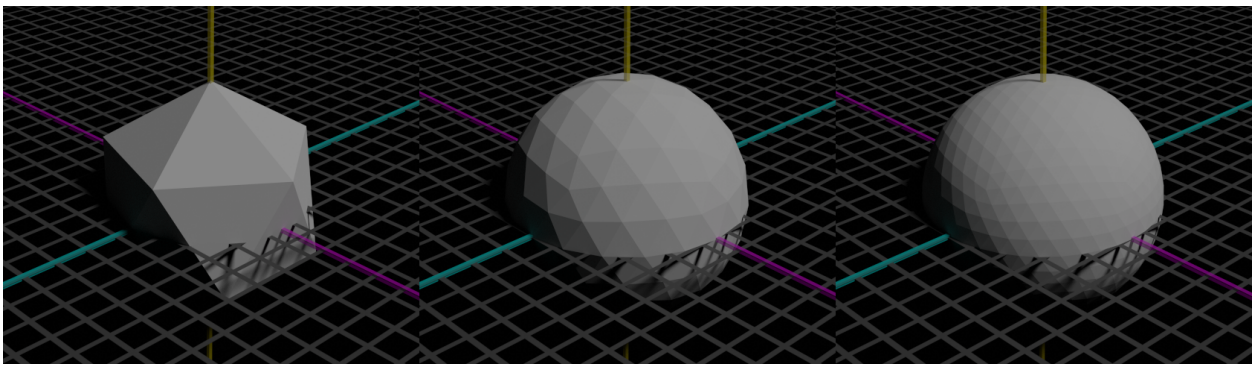

Figure 3.3: Three different icospheres of different vertex counts where the left has 12 vertices, the middle has 162, and the right has 642 vertices.

again, based on on how close the samples are to each other – a numerical method resulting in the usual accuracy to compute (data and/or performance) tradeoff. However, because the observation locations are along a hypersphere, the location L of a listener in (3.2) needs to be a position on the hypersphere's surface. Given the observers's true position O, we can find L by projecting them to the hypersphere given its center position C and its scalar radius r – shown in (3.4);

$$\mathbf{L} = \mathbf{r} \times \frac{\mathbf{C} - \mathbf{O}}{|\mathbf{C} - \mathbf{O}|} + \mathbf{C} \tag{3.4}$$

We also need to know which triangle is closest to this point L before we can use it in (3.2). Because the icosphere is convex, if we store each of its triangles' surface normals in an array N, then we can perform a support mapping function [50] to determine which face index f is closest to L, shown in (3.5) and (3.6).

$$\mathbf{Support}(\theta, \vec{D}, \mathbf{S}) = \vec{D} \cdot \vec{S_\theta} \tag{3.5}$$

$$\mathbf{f} = \arg\max_{\mathbf{i}} f(\vec{L}, N) = \{\mathbf{i} \mid \forall \mathbf{j} : \mathbf{Support}(\mathbf{j}, \vec{L}, N) \leq \mathbf{Support}(\mathbf{i}, \vec{L}, N), \mathbf{j} \in \{0, 1, \ldots |N|\}\} \tag{3.6}$$

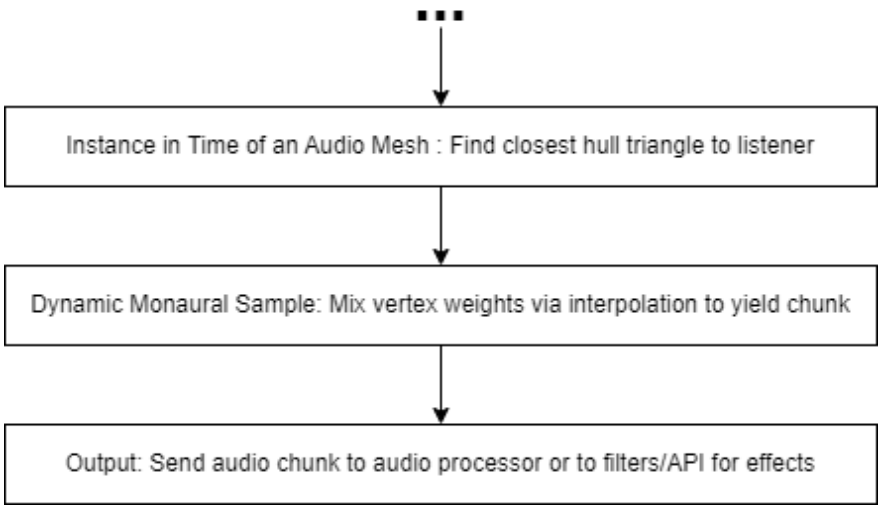

Figure 3.4: A flowchart describing an audio render pipeline to have 6DoF sound through transformable audio meshes, analogous to the graphics pipeline.

3.1.3 A Revised Audio Pipeline

By describing a triangulated mesh shape to represent the different observable sounds that are emitted from a sound source, we can exploit affine transforms in the same way they are used in the graphics pipeline to initially rotate and translate an audio mesh in a shared global space. Once we have this instance of an audio mesh transformed, we can synthesize a monaural sample by first determining which triangle from the audio mesh is closest to a listener, and then utilize barycentric coordinates to mix three distinct audio samples from that triangle. Because this yields a dynamic monaural sample, the information can still be passed into existing audio tools and filters for additional effects, RIR, and even HRTFs. This outline describes an audio rendering pipeline that closely resembles how computer graphics are rendered, shown in Figure 3.4.

For simplicity, this generalized approach was not implemented for human-subject experiments. Instead, a first-order approach (negative and positive axis-aligned recordings) was taken where instead of dynamically mixing audio chunks – the partitioned number of bits to sample from audio files – six audio directions were played simultaneously where their volumes levels were matched to the weights of the barycentric coordinates – re-

sulting in the same effect as the generalized solution. Having six axis-aligned sounds (e.g., $-/+X$, $-/+Y$, and $-/+Z$ microphone placements) ensured a simplified $O(1)$ implementation could be used in the experiment, as described in section 3.3.

3.2 Recording Method

The first step of building object-based audio with 6DoF involved isolating a single sound event during recording, as shown in Figure 3.5. Six standard condenser microphones with a cardioid pattern were placed 0.25 meters from the recording subject along three idealized axes representing three-dimensional Euclidean space. Each microphone corresponded to a positive or negative location along the X, Y, and Z axes. The microphones faced each other acting as a *"listener"* hearing the recording subject from the left, right, top, bottom, front, and back. This meant the recording subject existed at the origin of this coordinate system while all of the microphones recorded in the direction of the origin. This technique isolates a sound event by having all six microphones converge on a single sound source.

Each recording represents what a listener would hear if they were standing in the negative or positive direction of the orthogonal axes of Euclidean space. Each listener's perspective is from a common radius from the speaker and were captured at the same time eliminating the need to calibrate the recordings temporally. In this way each recording represents what would be heard from a given direction, at a constant distance, at a given time. This representation of the sound object means all recordings can playback under a common parametric representation of time. The final six recordings were made into a triangulated mesh as a data structure to represent the local coordinate space of a sound object.

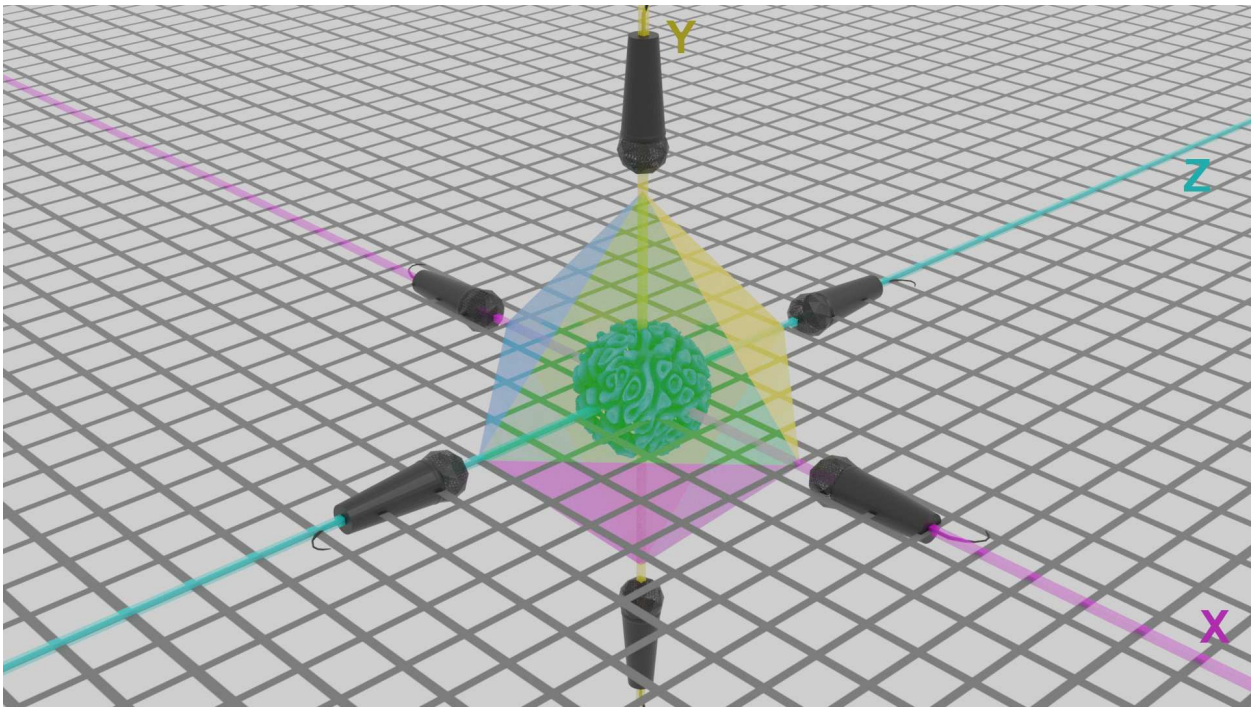

Figure 3.5: Visualization of the outside-in recording method. Microphones are placed at a common radius from the sound event to capture. This ensures a spatial-temporal coherency is present in the separate directional recordings – allowing a parametric representation of time to generate monaural samples in real-time. The recordings are stored as a triangulated mesh, where monaural samples are generated as an interpolation between three recorded samples of a triangle in the mesh.

3.3 Playback Method

The octahedron triangulated mesh provide simplifications to produce a new monaural sound from the six directional recordings. Real-time production of a new monaural sound enables this audio chunk to be used in existing transformers like HRTF and IRR. Because the octahedron is convex the most vertices to consider nearest to a listener's virtual position is three. Therefore, a final mix must then be a weighted combination of the recording samples across the nearest triangle facet of the octahedron. This weighting is calculated using barycentric coordinates to determine how much of each recording sample on a triangle should be mixed to reconstruct missing audio data.

First, it is important to know which triangle is closest to the listener's position. Because the octahedron is formed from axis-aligned vertices in both the positive and negative directions, the nearest triangle is described to be within the same octant as the position of the listener within the local-coordinate space of the sound object. To find the octant the listener is within, the global coordinate of the listener position is transformed into the local space of the sound object to determine which side of each plane (XY, XZ, YZ) the listener is on: resulting in one of eight octants. Due to the convexity of the mesh, this necessitates that the nearest position of the mesh will be on the triangle nearest to the listener, which is the triangle within the same octant as the listener. The process is described in Algorithm 1.

Given the triangle nearest to the listener, a barycentric projection is performed to yield the weights describing the listeners position within the triangle. These weights correspond to being nearest to a vertex, line, or within the face of the triangle. By multiplying the weights to their corresponding vertices and summing the results, a final nearest location to the listener is found. As these vertices are audio samples representing the spectral perspective of a listener at the vertices of the octahedron, these weights also represent how much of each audio sample to sum together to form a new monaural sound. This works because the weights sum to exactly 1.0 meaning that there will never be a generated

audio sample that exceeds the amplitude of the original three recordings being mixed. Algorithm 2 describes the prototype mixing process used in the Unity game engine.

For this study, all six recordings were played simultaneously and simulated mixing to a single monaural sample by adjusting the volume of each playback. This simplified the prototyping of the 6DoF technique by using the barycentric weights to directly modify the volume levels (output amplitude) of the sounds needing to be mixed. This simplification works without further modification, as the volume levels in Unity are normalized on a range of [0, 1.0].

Algorithm 1 findOctant Procedure

Require: input $position$
  $right \leftarrow (1, 0, 0)$
  $up \leftarrow (0, 1, 0)$
  $forward \leftarrow (0, 0, 1)$
  if dot product of $position$ and $forward$ is less than 0 **then**
    if dot product of $position$ and up is less than 0 **then**
      if dot product of $position$ and $right$ is less than 0 **then**
        return 7
      else
        return 8
      end if
    else
      if dot product of $position$ and $right$ is less than 0 **then**
        return 6
      else
        return 5
      end if
    end if
  else
    if dot product of $position$ and up is less than 0 **then**
      if dot product of $position$ and $right$ is less than 0 **then**
        return 3
      else
        return 4
      end if
    else
      if dot product of $position$ and $right$ is less than 0 **then**
        return 2
      else
        return 1
      end if
    end if
  end if

Algorithm 2 Update Sound Object

Require: $volumes = (front, back, left, right, top, bottom)$ such that $volumes_i = [0, 1]$ and each $volume_i$ is a $radius$ away from recorded subject

Ensure: $\sum_{i=0}^{5} volumes_i = 1$

$relative \leftarrow objectRotation^{-1} \times (listener - objectPosition)$
$octant \leftarrow findOctant(relative)$
$u \leftarrow (0, radius, 0)$ ▷ 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**
    $u.y \leftarrow -radius$ ▷ Vertical sound is in the negative position
    $active_0 \leftarrow 5$
end if
if $octant$ is 2 **or** $octant$ is 3 **or** $octant$ is 6 **or** $octant$ is 7 **then**
    $r.x \leftarrow -radius$ ▷ Horizontal sound is in the negative position
    $active_1 \leftarrow 2$
end if
if $octant$ is 5 **or** $octant$ is 6 **or** $octant$ is 7 **or** $octant$ is 8 **then**
    $f.x \leftarrow -radius$ ▷ Depth sound is in the negative 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**
    $sounds_{active_i}$ set volume to $weights_{active_i}$
    $i \leftarrow i + 1$
end while

Chapter 4

Methods

4.1 Software and Hardware

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.

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 their 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 [51] 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 ensure consistency existed between the final sets of audio.

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 successful 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 total 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. Between four randomized blocks of trials, sounds were instantiated at the center of each face on the cube. A 3-second countdown would appear to ready the participant for a trial, at which point one of the sounds on the cube would play. The sound was a short "hello" statement. There would be a 1-second pause, then the sound would play again. Finally, there would be another 1-second pause before playing the sound a third time. 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.

Study 1 had four total blocks of trials. In each block, the six sounds located on the cube's faces were 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 direc-

tional 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 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 rendering method prototype. 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. Using the 6DoF method meant simulating the experience of walking around a person while that person was 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. 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 rendering method for this study was used, or if a monaural sound was used with HRTF. Three blocks used the specialized rendering method with varying conditions. *6DoF Only* was a condition using only the proposed rendering method — no HRTF.

6DoF w/Surfaces was the second block condition which used the 6DoF prototype method but with HRTF. Because the prototyped method relied on playing six simultaneous sounds (rather than mixing the chunks of the sound samples into a monaural chunk), the sound samples could exist in disparate locations to be spatialized distinctly from one another with HRTF. This block condition had each of the six sound samples spread out about the cube so that, while a 6DoF interaction was present, additional interaural cues could be presented to the listener as they moved around the cube.

6DoF Centered was the condition in the third block, which used the 6DoF prototype and HRTF. This condition had all six sounds collocated at the center of the cube so that they were spatialized in a uniform manner. This representation of the purposed 6DoF playback matches what a finalized tool would do by mixing audio chunks into a new monaural sample in real-time.

Monaural HRTF was the condition in the fourth block, which did not use the 6DoF prototype method. This condition provided spatialization in the contemporary manner of having a single monaural sound played with a HRTF to provide a listener with spatial information. The sound was located on one face of the cube to ensure that a listener

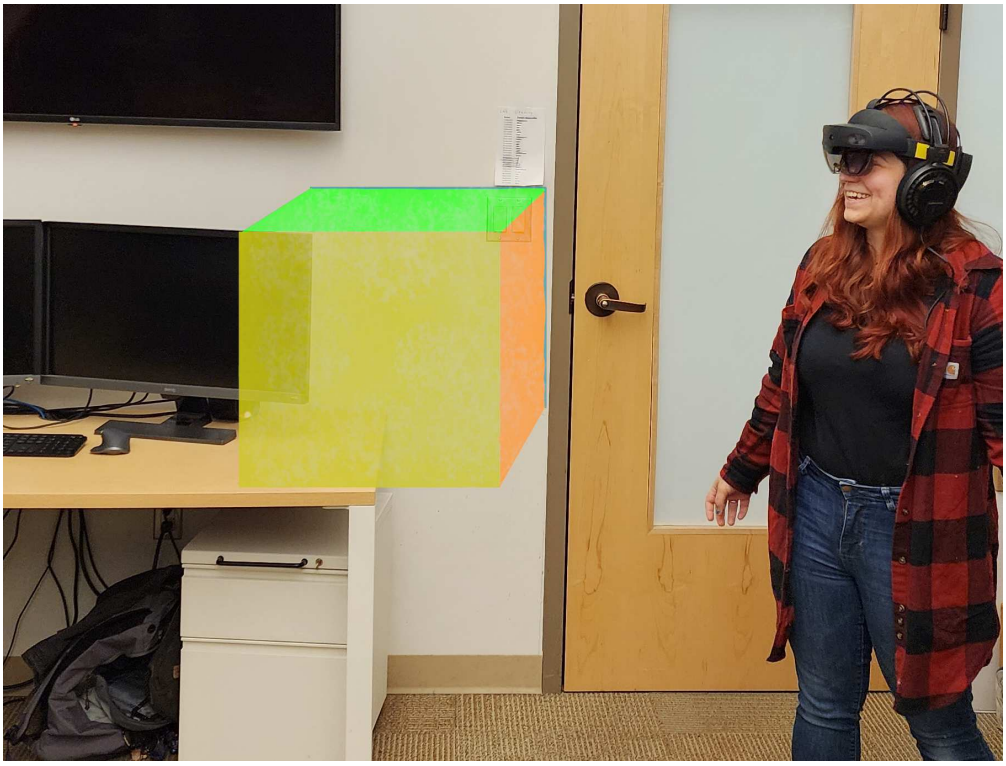

Figure 4.1: Depiction of Study 2's experiment.

could still be directed to a *"front-facing-direction"* of the cube given the observation of audio amplitude changes by proximity to the sound source.

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.

$H1$ 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 focuesed on HRTF spatialization M, each hypothesis is $H1 : D \neq M$.

$H2$ 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 $H2 : D \neq R$.

$H3$ 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 $H3 : X \neq A$. 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, $H4$, 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 $H4 : R \neq D$.

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 whether a block has a greater influence in side-by-

side comparisons to help assess additional interests in performance variance between directional and monaural sounds in Study 1 and 6DoF and HRTF conditions in Study 2.

Chapter 5

Results

The results were collected at the end of each participant's run through Study 1 and 2. The data from Study 1 was recorded from participant responses entered through a keyboard wirelessly connected to the HoloLens 2 and from the relative user position about the fixation object in Study 2 – yielding a directional vector. 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 the 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 rejection or non-rejection for its, respective, null hypothesis given a significance level of α = 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 5.1. 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.

Table 5.1: Hearing Test Trial Order

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

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 Ta-

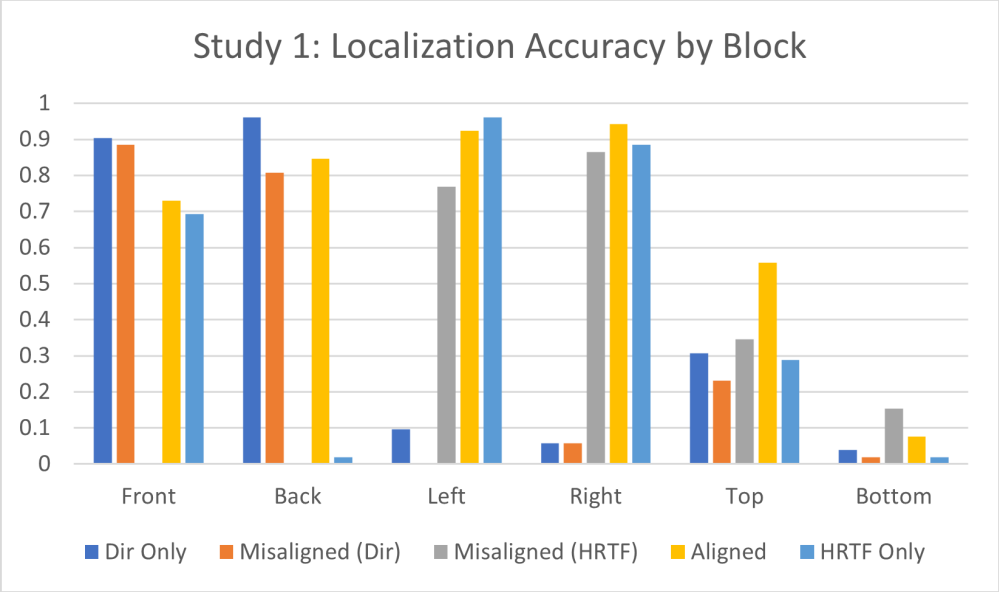

Figure 5.1: Total accuracy from Study 1 including the Dir and HRTF perspectives of the Misaligned block.

ble 5.2 and Figure 5.1 to assist in observing trends based on object-induced spectral cues. The Dir Only block used no contemporary spatialization technique and instead played each of the six directional recordings. The Misaligned and Aligned blocks played the directional recordings, with each one located at a consistent position to be accompanied by spatialization effects through a HRTF. In the Misaligned block, the directional recordings were misaligned with the positions – relative to the listener (Misaligned HRTF and HRTF, respectively) – while the Aligned block had the directional recordings matching a spatial position – relative to the listener. The HRTF Only block used only the front directional sound recording for a contemporary monaural HRTF representation.

Table 5.2: Study 1 Accuracy

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%

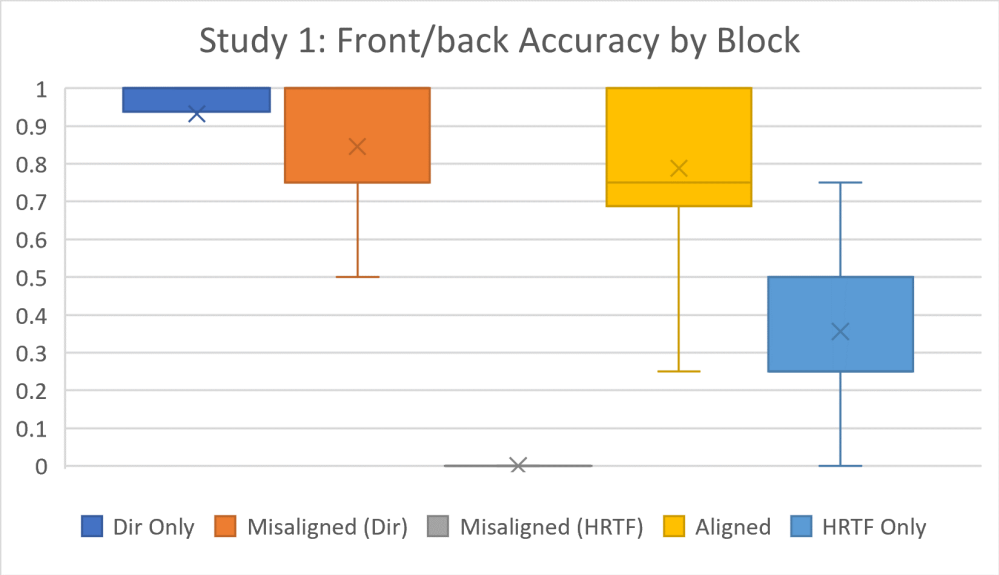

Figure 5.2: Front and Back localization accuracy during Study 1 for $H1$.

Figure 5.2 shows a boxplot for front/back response accuracy between each block before the Wilcoxon comparisons. Two sets of comparisons were performed to assess HRTF localization performance against the directional recordings that represent object-induced spectral modifications of a sound source observed from different locations about a common radius. These comparisons defined the tests for $H1$. Table 5.3 show the results for $H1$ 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.

Table 5.3: Results for $H1$ hypothesis tests in Study 1.

H1 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

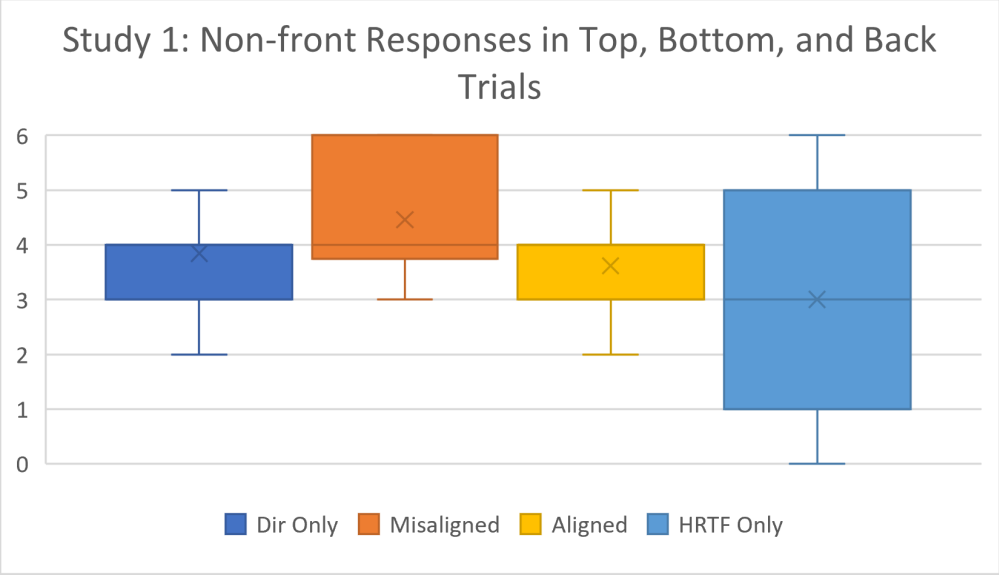

Figure 5.3: Non-front responses within the vertical plane of hearing during Study 1 for $H2$.

Figure 5.3 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 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 $H2$. Table 5.4 show the results for $H2$ testing where a (-) W observed value indicates a trend in favor of monaural HRTF for perception within the cone-of confusion while a (+) W observed value indicates a trend in favor of directional sounds for this perception.

Table 5.4: Results for $H2$ hypothesis tests in Study 1.

H2 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

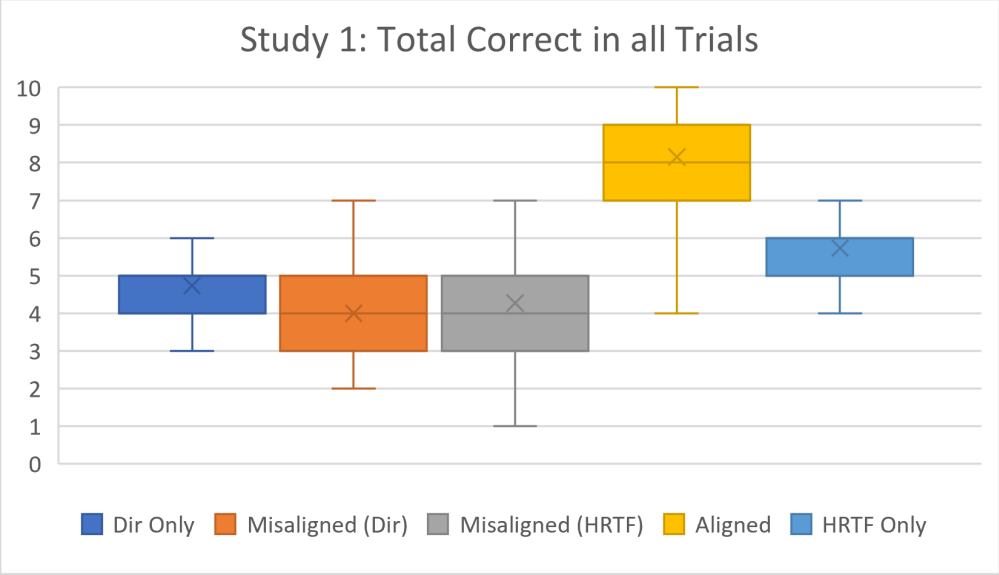

Figure 5.4: Total count of correct responses in Study 1 for $H3$.

Figure 5.4 is a box plot showing the count of correct responses with regard to sound direction or sound location. These results are used in tests are for $H3$. This hypothesis is an extension, and therefore dependent, of $H1$ for front and back accuracy being higher with directional sounds, but includes considerations of greater left/right localization given a HRTF. The final formulation of the hypothesis asks if coherency of vertical position and directional sounds are more meaningful for perception and sound localization. Table 5.5 show the results for $H3$ testing where a (+) W observed value indicates a trend for greater accuracy in the Aligned block condition.

Table 5.5: Results for $H3$ hypothesis tests in Study 1.

H3 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(+)	$p < 0.001$

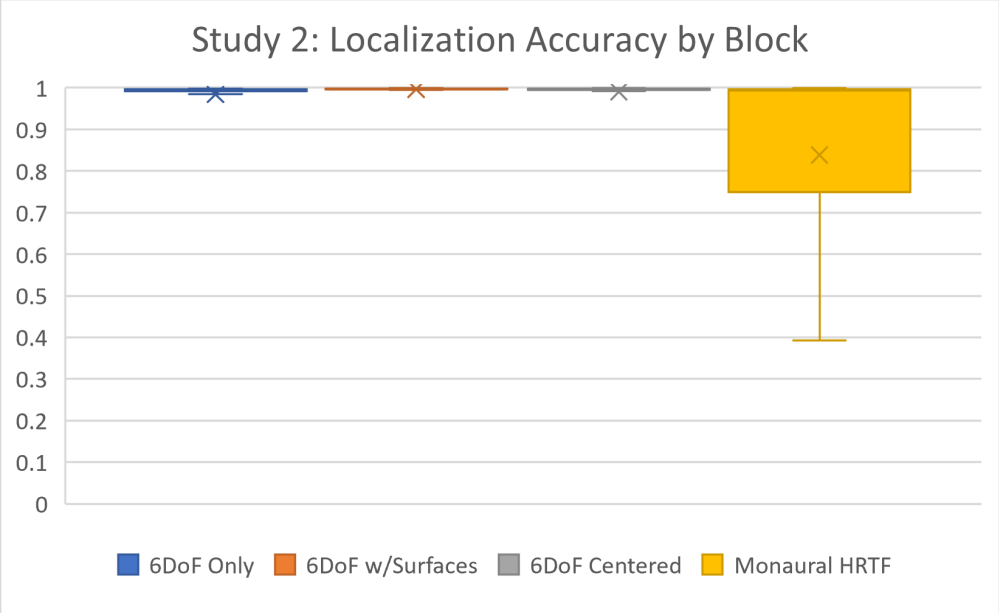

Figure 5.5: Localization accuracy in a mobile task between all blocks of Study 2. Participants walked around a virtual cube, augmented in their space, which was the source of our 6DoF audio method or monaural HRTF audio of human speech. Participants stopped in a position around the cube where they thought they would be facing the human speaker.

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 5.5 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, $H4$, which was that 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 5.6 shows the results for $H4$ testing where a (-) W observed value indicates a trend in favor of the Monaural HRTF block condition for localizing sound

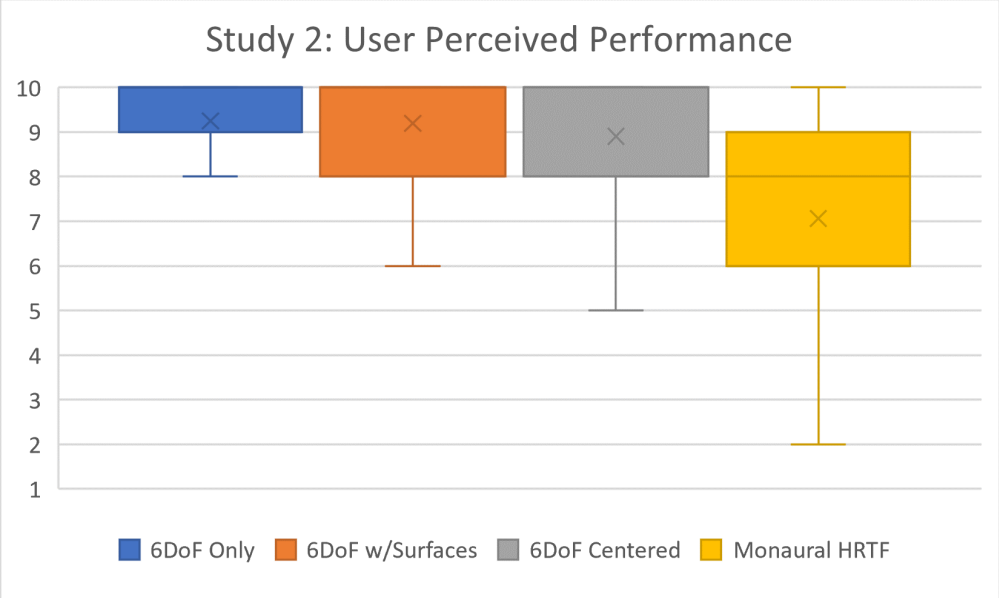

Figure 5.6: Perceived performance by block in Study 2. Participants reported a value on a scale of 1-10 for their level-of confidence to be standing in front of the human speaker based on the interactive audio located at the augmented fixation cube.

about an object in 3D space (+) W observed value indicates a trend in favor of 6DoF information for this type of localization task.

Table 5.6: Results for $H4$ hypothesis tests in Study 2.

H4 Tests	W	p-value
$Monaural\ HRTF\ \neq\ 6DoF\ Only$	90(+)	≈ 0.031
$Monaural\ HRTF\ \neq\ 6DoF\ w/Surfaces$	34(+)	$p < 0.001$
$MonauralHRTF\ \neq\ 6DoF\ Center$	56(+)	$p < 0.001$

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 5.6 shows the average of the two trial responses from each participant. Comparing this graph to 5.5 helps show a correlation between tangible and subjective performance in Study 2.

5.4 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 could be considered 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.

$H2$ 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 $H2$, 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.

Chapter 6

Discussion

In this paper, we outlined an object-based rendering technique for 6DoF sound that was prototyped and tested within human subjects. Our goal was to simplify the process to produce 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 $H1$. These results indicate directional audio was 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 the spectral modification of sound sources behind an object (in this case a human head) can provide audio cues that are more substantial than short-distance attenuation.

For hypothesis $H2$, there is not a definite indication as strongly in favor of this hypothesis compared to the results of $H1$. 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 5.4, comparing HRTF Only to Dir Only and Misaligned blocks indicates results in favor of $H1$ 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 $H3$ was in favor of hypothesis, suggesting coherent directional sound and HRTF spatialization is meaningful for a listener to perceive sound events about a 3D object. Table 5.2 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 5.1 for the Misaligned (HRTF) block. It is within the cone-of confusion where the next most interesting information resides. Front/back accuracy was already shown to be favorable with the direction cues in $H1$, 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 $H4$ uses a mathematical basis in geometry with vectors to assert a metric that could help determine a quantifiable means of assessing performance gains with 6DoF over monaural HRTF. The results indicate favoring $H4$ 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.

Chapter 7

Conclusion

7.1 Summary

In this work, 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 human spatialization. This can provide an example for how 6DoF audio tools could be tested with human subjects. This metric, combined with a 6DoF interactive audio experience, means additional questions on localization could be asked within psychoacoustics, accessibility, and HCI studies.

The proposed object-based 6DoF technique prototyped 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. Similar to ambisonics, it is possible that a higher-order of recordings from the prototyped technique could cue listeners better than the 1st-order axis-aligned recording technique used in this study.

A benefit of this technique as a tool is it minimizes the overhead of equipment and space required to produce a 6DoF interactive audio experience. 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.

7.2 Future Work

Some topics were not exhaustively explored for navigation and immersion as the human-subject research only extended to testing a prototype – to validate the use of 6DoF sound. For general use cases, this approach to 6DoF rendering could mean new experiments could be designed to see if audio navigation can be improved beyond the traditional use case of a monaural sound with HRTF. This type of question can be measured objectively by examining metrics like, time to complete or deviations from an ideal linear path. It is possible that my method could be used with artificial or idealized sounds to indicate paths in the vertical domain without having to rely on the error prone vertical localization of natural sounds with a listener's natural HRTF. Beyond this, subjective studies can be further expanded to see how a user responds to being in a virtual world that uses this technique for every sound – a question of immersion.

Additional numerical analysis could be useful to expand the theory behind this rendering technique. Questions to explore could relate to estimated accuracy loss, and attempts to balance memory usage (uncompressed audio footprints are large) to accuracy or user preference for a final rendered sound. This could mean using an non-uniform recording setup about a sound source – in the event that certain observer positions are more important than others. It is possible that this approach to interpolating directional recordings

could be meaningful in the domain of volumetric rendering as well as this technique of interpolation could equally ensure audio information is spread over spatial regions.

While this method was outlined for use in real-time software, this could be used in offline settings such as rendering animated films. If a high-density set of audio sphere recordings were captured, then an artist could place 6DoF sounds within regions of a scene to incorporate a more phenomenal audio experience to the film. Recording audio in real-life of a complex soundscape naturally captures directional nuance from objects within a scene. While films do have tracks like speech dubbed over for listener clarity, other sound events are most natural when kept to the sounds recorded when filming a scene. Having an audio object that encodes its own directional audio waveforms would allow animation studios to streamline dynamic audio within segments of their films.

Bibliography

[1] Justin Paterson and Hyunkook Lee, editors. *3D Audio*. Perspective on Music Production. Routledge, United States, 1st edition, July 2021. 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.

[2] Craig Jin, Anna Corderoy, Simon Carlile, and André van Schaik. Spectral cues in human sound localization. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

[3] Craig Jin, Anna Corderoy, Simon Carlile, and André van Schaik. Contrasting monaural and interaural spectral cues for human sound localization. *The Journal of the Acoustical Society of America*, 115(6):3124–3141, 06 2004.

[4] Jens Blauert and Robert A. Butler. Spatial Hearing: The Psychophysics of Human Sound Localization by Jens Blauert. *The Journal of the Acoustical Society of America*, 77(1):334–335, 01 1985.

[5] Frederic L. Wightman and Doris J. Kistler. Headphone simulation of free-field listening. I: Stimulus synthesis. *The Journal of the Acoustical Society of America*, 85(2):858–867, 02 1989.

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

[7] Enda Bates and Francis Boland. Spatial music, virtual reality, and 360 media. In *2016 AES International Conference on Audio for Virtual and Augmented Reality*, 2016.

[8] Michael A. Gerzon. Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society*, 21, feb 1973.

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

[10] Axel Plinge, Sebastian Schlecht, Oliver Thiergart, Thomas Robotham, Olli Rummukainen, and Emanuë Habets. 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*, 2018.

[11] OpenAL developers. Openal soft. https://www.openal-soft.org/.

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

[13] Valve Corporation. Immersive audio solutions. https://valvesoftware.github.io/steam-audio/.

[14] Facebook. Audio v47 reference guide. https://developer.oculus.com/reference/audio/v47/o_v_r_audio_8h.

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

[16] Unity. Audio spatializer sdk. https://docs.unity3d.com/Manual/AudioSpatializerSDK.html.

[17] Camilla H. Larsen, David S. Lauritsen, Jacob J. Larsen, Marc Pilgaard, and Jacob B. Madsen. Differences in human audio localization performance between a hrtf- and a non-hrtf audio system. In *Proceedings of the 8th Audio Mostly Conference*, AM '13, New York, NY, USA, 2013. Association for Computing Machinery.

[18] V. Sundareswaran, K. Wang, S. Chen, R. Behringer, J. McGee, C. Tam, and P. Zahorik. 3d audio augmented reality: implementation and experiments. In *The Second*

IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings., pages 296–297, 2003.

[19] Schuyler R. Quackenbush and Jürgen Herre. Mpeg standards for compressed representation of immersive audio. *Proceedings of the IEEE*, 109(9):1578–1589, 2021.

[20] Thirsa Huisman, Axel Ahrens, and Ewen MacDonald. Ambisonics sound source localization with varying amount of visual information in virtual reality. *Frontiers in Virtual Reality*, 2, 2021.

[21] Paul Power, Chris Dunn, William J. Davies, and J. Hirst. Localisation of elevated sources in higher-order ambisonics. Technical report, British Broadcasting Corporation, 2013.

[22] DeLiang Wang and Guy J. Brown. Binaural sound localization. In *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, pages 147–185. Wiley-IEEE Press, 2006.

[23] Sam Clapp, Anne Guthrie, Jonas Braasch, and Ning Xiang. Evaluating the accuracy of the ambisonic reproduction of measured soundfields. In *EAA Joint Symposium on Auralization and Ambisonics 2014*, 2014.

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

[25] Elizabeth M. Wenzel, Frederic L. Wightman, and Scott H. Foster. A virtual display system for conveying three-dimensional acoustic information. *Proceedings of the Human Factors Society Annual Meeting*, 32(2):86–90, 1988.

[26] Eduardo Patricio, Andrzej Ruminski, Adam Kuklasinski, Lukasz Januszkiewicz, and Tomasz Zernicki. Toward six degrees of freedom audio recording and playback using multiple ambisonics sound fields. *AES Convention*, 2019.

[27] Bartłomiej Mróz, Marek Kabaciński, Tomasz Ciotucha, Andrzej Rumiński, and Tomasz Żernicki. Production of six-degrees-of-freedom (6dof) navigable audio using 30 ambisonic microphones. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pages 1–5, 2021.

[28] Leo McCormack, Archontis Politis, Thomas McKenzie, Christoph Hold, and Ville Pulkki. Object-based six-degrees-of-freedom rendering of sound scenes captured with multiple ambisonic receivers. *Journal of the Audio Engineering Society*, 70(5):355–372, 2022.

[29] Armando Barreto, Kenneth John Faller, and Malek Adjouadi. 3d sound for human-computer interaction: Regions with different limitations in elevation localization. In *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '09, page 211–212, New York, NY, USA, 2009. Association for Computing Machinery.

[30] Michele Geronazzo, Simone Spagnol, and Federico Avanzini. A head-related transfer function model for real-time customized 3-d sound rendering. In *2011 Seventh International Conference on Signal Image Technology & Internet-Based Systems*, pages 174–179, 2011.

[31] Sascha Dick and Juergen Herre. Investigation of the impact of spectral cues from torso shadowing on front-back-confusion and perceived differences along cones of confusion. In *Audio Engineering Society Convention 155*, Oct 2023.

[32] Michele Geronazzo, Simone Spagnol, and Federico Avanzini. Do we need individual head-related transfer functions for vertical localization? the case study of a spectral notch distance metric. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(7):1247–1260, 2018.

[33] Irvin Steve Cardenas, Kaleb Powlison, and Jong-Hoon Kim. Reducing cognitive workload in telepresence lunar - martian environments through audiovisual feedback in augmented reality. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '21 Companion, page 463–466, New York, NY, USA, 2021. Association for Computing Machinery.

[34] Tomi Nukarinen, Roope Raisamo, Ahmed Farooq, Grigori Evreinov, and Veikko Surakka. Effects of directional haptic and non-speech audio cues in a cognitively demanding navigation task. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, NordiCHI '14, page 61–64, New York, NY, USA, 2014. Association for Computing Machinery.

[35] Jaime Sánchez, Mauricio Sáenz, Alvaro Pascual-Leone, and Lotfi Merabet. Navigation for the blind through audio-based virtual environments. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, page 3409–3414, New York, NY, USA, 2010. Association for Computing Machinery.

[36] Ketki A. Dhanesha, Nitendra Rajput, and Kundan Srivastava. User driven audio content navigation for spoken web. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1071–1074, New York, NY, USA, 2010. Association for Computing Machinery.

[37] Stuart Goose and Carsten Möller. A 3d audio only interactive web browser: using spatialization to convey hypermedia document structure. In *Proceedings of the Seventh ACM International Conference on Multimedia (Part 1)*, MULTIMEDIA '99, page 363–371, New York, NY, USA, 1999. Association for Computing Machinery.

[38] Chengcun Gu, Mengyao Zhu, Haofeng Lu, and Benoit Beckers. Room impulse response simulation based on equal-area ray tracing. In *2014 International Conference on Audio, Language and Image Processing*, pages 832–836, 2014.

[39] Iana Podkosova, Michael Urbanek, and Hannes Kaufmann. A hybrid sound model for 3d audio games with real walking. In *Proceedings of the 29th International Conference on Computer Animation and Social Agents*, CASA '16, page 189–192, New York, NY, USA, 2016. Association for Computing Machinery.

[40] Konstantin Semionov and Iain McGregor. 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*, AM '20, page 22–29, New York, NY, USA, 2020. Association for Computing Machinery.

[41] Dukki Hong, Tae-Hyoung Lee, Yejong Joo, and Woo-Chan Park. 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*, I3D '17, New York, NY, USA, 2017. Association for Computing Machinery.

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

[43] Johannes M. Arend, Sebastià V. Amengual Garí, Carl Schissler, Florian Klein, and Philip W. Robinson. Six-degrees-of-freedom parametric spatial audio based on one monaural room impulse response. *Journal of the Audio Engineering Society*, 69(7/8):557–575, July 2021.

[44] Nikolaos Moustakas, Emmanouel Rovithis, Konstantinos Vogklis, and Andreas Floros. Adaptive audio mixing for enhancing immersion in augmented reality audio games. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ICMI '20 Companion, page 220–227, New York, NY, USA, 2021. Association for Computing Machinery.

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

[46] OpenGL Wiki. Rendering pipeline overview — opengl wiki,, 2022. [Online; accessed 21-April-2024].

[47] Carl Schissler, Aaron Nicholls, and Ravish Mehra. Efficient hrtf-based spatial audio for area and volumetric sources. *IEEE Transactions on Visualization and Computer Graphics*, 22(4):1356–1366, 2016.

[48] E.G. Gilbert, D.W. Johnson, and S.S. Keerthi. A fast procedure for computing the distance between complex objects in three-dimensional space. *IEEE Journal on Robotics and Automation*, 4(2):193–203, 1988.

[49] H. Gouraud. Continuous shading of curved surfaces. *IEEE Transactions on Computers*, C-20(6):623–629, 1971.

[50] Christer Ericson. Chapter 3 - a math and geometry primer. In Christer Ericson, editor, *Real-Time Collision Detection*, The Morgan Kaufmann Series in Interactive 3D Technology, pages 23–73. Morgan Kaufmann, San Francisco, 2005.

[51] Acm code of ethics and professional conduct, 2018. https://www.acm.org/code-of-ethics.

Appendix A

Unity C# Partial Class Code

```
public void updateSound(int octant)
{
    //In close proximity, cam need not be projected to sphere
    Vector3 cam = Camera.main.transform.position;
    soundTop.volume = 0.0f;
    soundBottom.volume = 0.0f;
    soundFront.volume = 0.0f;
    soundRear.volume = 0.0f;
    soundLeft.volume = 0.0f;
    soundRight.volume = 0.0f;
    int i = 2, j = 5, k = 0; // ijk initialized to case 1
    switch (octant)
    {
        case 2:
            i = 2; j = 4; k = 0; break;
        case 3:
            i = 3; j = 4; k = 0; break;
        case 4:
            i = 3; j = 5; k = 0; break;
        case 5:
            i = 2; j = 5; k = 1; break;
        case 6:
            i = 2; j = 4; k = 1; break;
        case 7:
            i = 3; j = 4; k = 1; break;
        case 8:
            i = 3; j = 5; k = 1; break;
    }
    Vector3 u = globalPositions[i];
    Vector3 r = globalPositions[j];
    Vector3 f = globalPositions[k];
    Vector3 vol = barycentricWeights(u, r, f, cam);
    if (i == 2) soundTop.volume = vol.x;
    else soundBottom.volume = vol.x;
    if (j == 4) soundLeft.volume = vol.y;
    else soundRight.volume = vol.y;
    if (k == 0) soundFront.volume = vol.z;
    else soundRear.volume = vol.z;
}
```

```
public int inOctant()
{
    Vector3 soundPosition = this.transform.position;
    Vector3 viewPosition = Camera.main.transform.position;
    Vector3 r = viewPosition - soundPosition;
    //X+ is left and X- is right - akin to facing someone
    if (Vector3.Dot(hyperplanes[2], camRelative) < 0.0f)
        if (Vector3.Dot(hyperplanes[0], camRelative) > 0.0f)
            if (Vector3.Dot(hyperplanes[1], camRelative) < 0.0f)
                return 1;//Front Top Right
            else
                return 2;//Front Top Left
        else
            if (Vector3.Dot(hyperplanes[1], camRelative) > 0.0f)
                return 4;//Front Bottom Right
            else
                return 3;//Front Bottom Left
    else
        if (Vector3.Dot(hyperplanes[0], camRelative) > 0.0f)
            if (Vector3.Dot(hyperplanes[1], camRelative) < 0.0f)
                return 5;//Rear Top Right
            else
                return 6;//Rear Top Left
        else
            if (Vector3.Dot(hyperplanes[1], camRelative) < 0.0f)
                return 8;//Rear Bottom Right
            else
                return 7;//Rear Bottom Left
}

void Update()
{
    if (!soundFront.isPlaying)
    {
        soundFront.Play();
        soundRear.Play();
        soundBottom.Play();
        soundLeft.Play();
        soundRight.Play();
        soundTop.Play();
    }
    updateSound(inOctant());
}
```

A.1 Description

This code can be used in a Unity script for a class storing six $AudioSource$ objects, an array of three $Vector3$, and an array of six $Vector3$.

The first array is for storing axis-aligned normalized vectors (e.g., Vector3.right). As seen in the $inOctant$ function, hyperplane is storing these directions to determine if the listener is to the left or right, top or bottom, and front or back sides of the sound object. If the audio object is rotated, then these vectors could also be rotated.

The second array is storing the global positions of each sound object, as some radius distance and direction from the center of the parent (i.e., $this.transform.position$) object which has these six sounds as field variables. These are used to determine the barycentric weights to alter the volume of relevant directional sounds in a given frame.

This represents a fixed orientation for the 6DoF audio object – pulled from the Study 2. Additional transformations can be applied to take this code and allow for the 6DoF object to be rotated. An inverse rotation could be applied to find the relative listener's position in the local coordinate space of the 6DoF object to determine which octant the listener is in and to compute the barycentric weights. Using the inverse rotation would mean avoiding to have to transform each hyperplane and sound position. Of course, those transformations could instead occur at the start of the update loop to keep the inOctant and upSound methods unchanged.

Appendix B

Experiment Documents

B.1 Screening Requirements

To participate you must meet the following criteria. If you meet these criteria then you are a member of the population of interest for this research, which is that of persons over 18 that can use head mounted displays.

• Be age 18+

• Have normal vision or corrected-to-normal vision.

• Have normal or corrected-to-normal hearing. If you normally wear contact lenses or hearing aids you will need to wear them to participate.

• Have no known history of seizures

• Have a self-reported understanding of the English language (text and verbal prompts are in English)

B.2 Questionnaire

Have you heard of spatial or 3D sound before? Do you have an idea of what it means?

Have you used VR or AR devices before?

Do you play video games, and if so how often?

If you play games, how much do you rely on audio?

B.3 Debrief

Audio Position Debrief (A)

Thank you for taking the time to volunteer in this study on audio perception in augmented reality. This experiment will take approximately 30 minutes to complete. You will be wearing a Microsoft HoloLens 2 and an open-back pair of headphones for this exper-

iment. After you have finished reading these instructions, these devices will be fitted to you in order to start the experiment. If at any time you feel nausea or discomfort and want to end the experiment, let us know and we will stop immediately. You might experience moments of discomfort from peaks in sound volume, but all sound levels have been tested and calibrated to be within safe listening volumes for the duration of the experiment.

Trials

The start of the experiment includes an audio test to help ensure you will be able to hear the sounds used during the trials. Eight sounds will be played, one at a time, from either the left or right headphone speaker. After it has played, please respond if you heard it from the right or left side, and let us know if you could not hear it at all. After the sound test, you will go through two experiments of trials. Both involve listening to spoken recordings. The sounds are played back using spatialization techniques, and your responses to questions during the trials will help us gauge which techniques are better and when.

Both experiments have a single practice trial to help you get ready for the real trials. Each practice is played before the first experiment of corresponding real trials.

One experiment will ask you to be seated. You will view a cube in front of you. Please keep your head still as moving around might vary how you hear sounds. After an initial 3 second countdown these trials will play a sound three times and you will be asked where around the cube you think the sound came from: the back, front, left, right, top, or bottom of the cube. This is a localization test and is meant to see which spatialization technique cues you into the location of a sound in the space in front of you.

The second experiment will ask you to stand. You will see a cube with several colored faces. When these trials start playing sound, you will hear a continuous dialogue. You are asked to walk around the cube and stop in front of the face (one of the four faces perpendicular to the ground) that seems to be where the sound is coming from. Specifically, think about walking around a person that is talking, and try to stop where it sounds most

like you are standing in front of them – i.e., facing their face as they talk. You will be asked how confident you feel that you are standing in front of this person, on a scale from 1 to 10. Please stand in place until your confidence value is recorded.

The researcher controls the progress within trials. Please wait for all sounds to finish playing before responding to questions. These questions will appear within your field of view to remind you of what the question was, and to let you know what time a response is required.

The experiments will play twice, alternating between each type of trial (except the practice). The experiments will start in a random order. If you have any questions or confusion, please ask them now.