Orientable Audio

Dan Rehberg*
Colorado State University
CS 567

ABSTRACT

While audio has not received as much attention in the space of virtual environments as graphics, the range of spatial and 3D audio effects have found success in presenting an immersive environment that attempts to render realistic audio alongside high graphical fidelity presentations. However, these effects are rather static when compared to the diverse manner in which audio can propagate in the real world. While the result of head-related transfer functions, interaural intensity and time difference, binaural audio, and ambisonics have improved the representation of sound in real-time rendering, the current use cases employ these techniques against static representations of soundscapes or point source sounds. In addition, without calibrating these traditional playback methods in stereo, the effects tend to fail at demonstrating localization in front/back and top/down scenarios. This research aims to investigate if there are additional changes in perception when hearing a sound source represented from multiple angles, and if experiencing these multiple sound sources further enhance the localization of sound in virtual and augmented environments.

Index Terms: Human-centered Computing [Human Computer Interaction]: Interaction Techniques—Auditory Feedback; Computing Methodologies [Modeling and Simulation]: Simulation Types and Techniques—Interactive Simulation

1 INTRODUCTION

3D audio in real-time interactive environments has a history going back to the 1990's in computer games, and perception of spatial audio extends to the 1930's when stereophonic sound was patented. While the advent of stereo speakers and separate sound channels rapidly deployed to headphones and led to surround sound, these techniques only relate to experiencing static audio sources. That is, an array of speakers playing different mixes to each speaker provides a listener an experience of audio from a limited sound stage. While increasing the number of surround speakers increases the resolution of localized audio, the result does not provide dynamic positions of sounds to be heard in a film - e.g., a viewer cannot request a camera to move 6 meters to the left in a pre-recorded scene. Interactive environments allow users to move around objects, so what should a user hear if they are in front of a train compared to beside it? While head-related transfer functions (HRTF) are used to produce sound at different rates between different speakers - enhancing interaural intensity and time differences with a transformation to account for ear shape and head to torso locations, this is traditionally applied to a single ***point source*** sound and not multi-directional audio. Ambisonics improve this by recording an encapsulating space of sound but still only represent recordings of sound from one location. Ambisonics therefore provide an interaction, but the primary interaction is changing orientation with respect to wherever the original soundscape was recorded.

*e-mail: dan.rehberg@colostate.edu

The interactive space of virtual environments has been extending to augmented and virtual reality with the increase of dedicated commercial products and software adapted to common devices like smartphones. Investigating the effects of additional sensory information may be crucial to find new ways in improving immersion but also general perception of locality within a space that allows natural human interaction. While the HRTF function is most effective when calibrated to an individual's ear pinna and upper body configuration, its generalizable usage has been successful enough to see it implemented in devices at the hardware level - such as in the Microsft HoloLens 2. As a generalized model, the HRTF method offers localization but might falter when manipulating vertical and front/back positioned sounds - with the worst result being a sensation that the sound is positioned inside one's head.

The recording of soundscapes with microphone arrays helps alleviate this problem as a representation of a continuous set of point source sounds are captured from multiple angles simultaneously, thereby also capturing all of the nuance of wave interference in a 3D space. The experience might not be perfectly tailored to vertical and front/back sounds for a given user, but enough information is recorded that a user can acclimate to this space. Binaural recordings can even used molds of ears to offer a simulation of sound heard from the perspective of a given ear shape to attempt to produce greater fidelity in localization. However, the common usage of these recording methods are typically limited in the amount of interactivity they provide with binaural audio being (generally) entirely static - like watching a film - and ambisonics (generally) only allowing the viewer to reorient themselves within the pre-recorded sound stage. Interaural and HRTF effects allow greater flexibility in interactive scenarios by playing separate sound sources with locations in the virtual, or augmented, space and manipulating the sound to provide localization hints to a user. Both the the interaural/HRTF methods and microphone array to produce a soundscape have their lacking qualities: one offers a simulation to apply to separate and arbitrary counts of sound sources in a dynamic real-time scene while the other plays a 3D continuum representation of a sound stage from a static position.

Increases in computational power have allowed audio to be further manipulated to provide more ***realism*** as a result. Practices from high-fidelity graphics rendering have been re-adapted to simulate the travel of sound waves in a virtual scene. These offer a high degree of realism for complicated scenarios like occlusion, diffraction, reverberation and general interference of waves. This differs from HRTF and the like which consider an ideal scenario of an open constant sound medium in which head orientation, sound position, and physiological configurations are known within this space. Interpreting and simulating these effects even mean that ***virtual speakers*** can be placed in a real-world environment by using reverberation analysis to project audio from a front-facing speaker off of the ceiling to produce vertical sources of sound. This additional compute power in the augmented space has been used to dynamically alter the amplitude of audio in order to have it match the sounds of the environment a user is in.

The limited yet diverse background of audio procedures are investigated in this work in an attempt to provide a new method for increasing audible information in fully interactive scenarios. This research questions at which point sound information processing should

begin before utilizing dynamic effects in 3D environments with either HRTF or wave propagation simulation. It is considered that the array microphone configuration offers additional information that is relevant to a person who has interacted with sound actively and passively in their typical life - sounds which vary by orientation and location as an event which propagates waves in a medium are sensed. However, to produce individual ***point source audio*** which can be captured with the full range of positional variance, it is proposed that recording in a uniform sphere around a sound event is a better choice for dynamic manipulation in interactive environments when compared to traditional static microphone array captures like ambisonics and binaural recordings. Moreover, a psychophysics question becomes relevant as more information is provided to a listener interacting in a virtual or augmented 3D space: *do recognizable sounds, like listening to someone talk, offer additional localization cues about position given a location and orientation to the audio being rendered?*

In general, audio has had a fixation on stereo and surround sound applications with fairly static reproduction of pre-recorded sounds. With the increase in compute performance and audio hardware acceleration, further methods and models of audio rendering are beginning to be explored for interactive usage. With this, can methods of microphone array recordings be refactored to provide the additional information found in ambisonics and binaural recordings in immersive interactive settings? The uniform spherical recordings proposed are employed by mimicking a long-standing practice of real-time graphics rendering, storing volumetric data and transforming it by local and global states before applying complicated render details - thereby presenting the most applicable surface of the volume directly to a user.

2 Related Works

Previous research and experiments have demonstrated a plethora of effects found when simulating spatial audio in virtual environments as well as interactive 3D film. Additionally, for several decades software packages have been developed to further promote the utilization and creation of ***3D audio*** such as FMOD and OpenAL with newer mixing software such as Wwise. The description of 3D audio here is synonymous to ***spatial audio*** as the effects produced in these and other complicated systems is meant to increase the ability that a user can discern locality by their presence in a virtual or augmented space and the alteration of audio by manipulating the amplitude, frequency, and supplying a latency of playback between stereo sound systems. These effects result in mimicking some phenomenal aspects of sound waves which then allow the user to intuit where the sound is located. The basic effect is panning which then lends to interaural and HRTF systems in an attempt to extend the dimensionality of localization beyond two-dimensional environments. More complicated systems produce soundscapes via ambisonics and specialized binaural recordings.

Consideration for how these audio effects are used in traditional and novel ways was critical in developing a method for which point-source audio could be used without additional modification of existing audio system. That is, understanding how 3D audio was currently in use was important to engineer a modular procedure for which additional orientation-based audio information would be mixed into traditional single audio spatialized renderings.

2.1 Current Applications

Improvements in computational power have shown that simple software mixing to perform HRTF operations in software no longer impede real-time performance to the point of requiring intrinsic instructions at the hardware level [8]. Of course, promoting the rate at which common audio manipulations can be processed - in hardware - should further promote the complexity of tasks that can occur to modify audio in software. This compute power is useful for scenarios where external sounds need to be combined, or normalized, with sound being rendered in real-time on interactive devices [10]; a scenario where audio is modified to match the outside environment for augmented reality, and avoiding an incoherency between augmented and real-world sounds. In more intensive applications, 3D rendering techniques have been used to produce the same dispersal of wave-like behavior for audio playback typically used for high fidelity graphics when lighting a virtual scene by considering surface level interactions - e.g., diffusion - of waves [7]. This increase in power means interactive scenarios like video-games can have new audio rendering pipelines tested to determine what limitations are still present when throwing new creative design questions at software. Notably, persistent audio for music and soundscapes still face the limitations of total mixable sound chunks per available audio channel in game engines as well as spatializer hardware [3]. These limitations do not typically effect sound clips being mixed as they have short durations and can be more easily culled without a loss in continuity [3]. Even so, hardware level features and computational power are enabling further fidelity delivery even over network communications as stereo microphones have been used to capture binaural speech delivered in real-time to live remote users [9] - incorporating a known immersive audio experience which might be further practical at present with increases in telecommunications in virtual spaces. Binaural speech for telecommunication is a limited example demonstrating a non-static scenario in which a microphone array can exist with interactions; this demonstrates real-time software which emulates life-like conversations with others. Coupling these audio manipulations with hardware acceleration means more intensive processes can occur to test the effects of audio in virtual and augmented environments, whether these are computational simulations of environmental propagation, dynamic playback based on user input, augmented reality soundscapes, or integration in augmented systems to further integrate sensory elements into a mixed-reality [3, 7, 10].

General reasons to offer better audio rendering extend to increasing the perception of the visually impaired with computer interaction and to enhance the amount of sensory data an AI can learn to distinguish events in an environment; as well as generally promoting greater immersion in virtual and augmented environments. For instance, experiments have been produced to utilize machine learning with mobile phones to first identify the type of object in the real-world as well as its distance from the phone and secondly play a spatialized audio cue to guide a visually impaired participant to the object [6]. This is beneficial to a user to identify relevant sets of items in their local environment and minimizes latency by running entirely from a smartphone. Pathfinding by spatialized audio is not a new subject, with an effectiveness demonstrated for HRTF over simpler interaural effects [8], but this augmented implementation is a unique bridge to provide pathfinding to the visually impaired in real-time. For an adept listener, it is wondered whether additional nuances from audio rendering could provide additional information (like orientation) of the identified item. Additionally, precise detection of individual localized sound sources has been explored to promote better environmental awareness in autonomous AI agents [13]. This increases the amount of interpretations an AI can make in a given environment in the event that light-based sensors have artefact data or in the event that sound based information which could be analyzed through light is occluded. If given a machine learning model to train for distinct types of sounds to parse for locality, it is wondered if directional audio information might be beneficial in the training of an autonomous agent.

These scenarios explore how perceptions of phenomenal sound data can be increased by utilizing microphone arrays and general increases in computational power. These applications have relied on the previous work of sound localization from point sources in both dynamic rendering of single sounds or in an encapsulating en-

vironment (soundscapes) of audio. Some of the research offer novel reasons to increase the fidelity of sound, while others indicate an opportunity exists to utilize hardware acceleration and computational power effectively to explore how better audio rendering might occur in former and contemporary (augmented) virtual environments. The breadth of these examples is purposefully wide as none seem to demonstrate a formalism or method to combine the encapsulation of a soundscape with spatialization effects which work with at least six-degrees of freedom in an interactive space. A further breakdown of point source and soundscapes is provided to illustrate strengths and weaknesses as well as to indicate how these two models of audio rendering might be bridged.

2.2 Modifying Point Source Audio

HRTF is an existing procedural practice to produce localizable sound by delaying when that sound is played on different speakers. It is practical as the operation it performs solves a distinct but common problem with producing spatial – or 3D – audio. That is, HRTF is a generalizable approach to problems, but requires some overhead. It is a more complicated variant of interaural intensity and time differences - an effect that changes when speakers start playing back a sound and with what amplitude as well as offering "spatialization" as opposed to 2D discrimination of stereo panning effects. Traditional audio mixing for complicated scenarios happen at hardware levels or involve concurrent operations to decode signals, which is the case of making a multi-purpose 3D audio format for separate channels, ambisonics, or binaural audio [5]. These standardization and formats mean hardware can be engineered to produce rapid effects based on a standard, for instance Dolby Atmos, in order to produce spatial sounds more quickly than software implementations. The reliability of HRTF as a general mechanism is likely a factor in hardware acceleration for this functionality being added in devices such as the Microsoft HoloLens 2 or even Sony's Playstation 5 with its hardware accelerated Tempest 3D Audio. The generalized problem that HRTF solves is not without compromises with limited calibrations. Notably, interaural time and intensity differences do not produce well defined representations of front, back, or vertical sounds, and while HRTF does produce better verticality, these effects in localization vary by individual because their perception of these spaces is based on ear shape among other bodily configurations [6, 8, 11]. Even so, the effect of HRTF in first-person-shooter video-games has been found to enable judgement calls in players who localize sounds of threats based on amplitude and direction while combining visual elements to help indicate whether these sounds are in an open space or behind obstacles [11]. This indicates that in a fast real-time setting with combined graphics, the generalized HRTF is still offering meaningful localization cues.

In a test of precision for front/back errors with HRTF, the effectiveness of a generalized HRTF model seemed to not offer any strong relation of sounds within 180 degress behind a listener compared to no HRTF at all, despite general forward sounds within a 180 degrees having vastly greater precision over no HRTF [8]. Despite this, the effectiveness of HRTF has been found to provide an interface in which users can localize the position of an audio source faster than in stereo panning systems, with a demonstrated experiment having participants navigate a path by sound cues with great efficacy when using a generalized HRTF method [8]. These results were newer but correspond to generalized HRTF testing in augmented reality during the mid 00's where on a known horizontal plane the generalized HRTF model (not specialized for individual user's ears) was highly effective at providing localization but under conditions where the vertical plane randomly elevated sounds provided difficulty in localization [12]. These sound cues were associated with augmented rendered airplanes with a description of verticality being an issue compared to graphical elements combined with front/back sounds being assessed for danger effectively in the more recent first-person-shooter study [11, 12]. Despite this limitation, practice and training with non-calibrated interactions of HRTF modified audio has been found to have similar benefits as found by Larsen et al. in the visually impaired smartphone augmented pathfinding study [6, 8].

2.3 Soundscapes

Binarual recordings, ambisonics, and general coincident microphone arrays are methods employed to capture a ***soundscape***; a form of spatial audio information that puts a listener at the center of a sound stage able to localize sounds due to multiple angles of a encapsulating volume of sounds (capturing nuanced interference) being recorded simultaneously. Binaural audio recordings are further immersive by simulating ear pinna by having ear molds over the microphone diaphragm - only requiring stereo headphones for effective immersion and presence in playback, but binaural audio can also be simulated effectively by using ambisonics where an additional application of a HRTF is effectively shifted from the center of the head to the locations of each respective ear [2]. This suggests that the efficacy of static binaural recordings can be simulated by using the extra information of separated microphones from ambisonics with consideration that these extra microphones can allow for a virtualized translation to ear positions about a listener. Ambisonics generalize well to arbitrary speaker arrangements by utilizing software and hardware processing to reduce or increase the mapping of the microphone array data to a given sound system [1, 2, 5].

The concept of using microphone arrays for localization is found both qualitatively and quantitatively. The quantitative aspect is found in consumer products with multiple microphones to better isolate background noise from a speaker, and includes producing locality in autonomous machines to identify sounds using mathematical models such as inverse-square relationships and the distal separation of each microphone [13, 14].

2.4 A Potential Gap

The result of the soundscape in ambisonic and binaural recordings offers strong experiences of a multitude of oriented audio, and its complex interactions in a shared medium, thus providing a strong context of the volume of space a listener is projected in. Producing manifolds of outward facing microphones faces the same issue of producing single or multi-camera setups for film, the amount of information of a 3D space is increased but it is limited to the placement of the capture device. The quality of capturing more wave development in a continuous medium is shown to be an immersive experience which when executed well can transport an individual to the audio setting recorded from a microphone array [2]. However, this does not work for an interactive environment with the typical six degrees of freedom - primarily because these detailed volumetric representations of sounds are not isolated from other sound events, and the volumetric information is only coherent from one perspective - consider a visual analogy of observing the wrong side of a non-closed topological mesh, the information exists but is meant for being experienced from the other side. The typical HRTF has been the standard baseline effect for spatializing single audio clips in a 3D environment, but does not contain volumetric information and in a hypothetical scenario is collecting one-dimensional information as the size of a microphone diaphragm approaches an infinitesimal surface area. If combined with visual imagery, a user in an interactive environment can combine the information of front/back and vertical sounds with visual imagery to make judgements, but they cannot do so with the sound alone unless the HRTF method is calibrated for the user.

3 Methodology

The initial idea is to represent audio recordings in a similar fashion to how 3D modeling is translated into linearized meshes for real-time rendering. Before any complicated lighting is performed, a

triangular mesh has its vertices transformed based on its local state and its global state in a scene with an orthographic or perspective frustum. This mesh information and this transformation provides visual cues for what shapes are currently attributed to this mesh from the a dynamic arrangement of translational and rotational degrees of freedom allowed in interactive software. This results in cues based on prior experience in the phenomenal world for a user, for instance being able to determine what the forward facing direction is for a humanoid mesh and where in the scene the relative forward-vector is pointing. This research proposes to re-use the microphone array to capture volumetric sound data, but to have the array pointing inward to a sound event rather than outward into a soundscape. Similar to the graphical mesh, a closed surface of information would be captured and transformed based on local and global states to render the most relevant surface to a user before any further complicated procedures are applied to the rendered audio. ***It is hypothesized that this directional recordings will assist in front/back recognition with a generalized HRTF in use, a situation where a non-calibrated HRTF does not excel.***

3.1 Recordings and Equipment

An ideal scenario would be to record a common sound event from multiple angles simultaneously by arranging a microphone array, but this research instead utilized a repeatable sound event through sampling several recordings of speech and by using a percussion instrument driven by gravity. Recordings were produced using a standard digital-audio-converter (DAC) and a standard condenser microphone. The gain on the DAC was kept constant for each sound event to record, and for each angle recorded the microphone was placed at about 6 to 8 inches from the audio source depending on the recording, within half an inch of error. The speech recorded was of a person saying ***hello***. This was recorded eight times in the six directions required for oriented audio mixing - visually, the closest matching sound signatures were made into the six distinct directional clips to use in the primary experiment. The instrument was a glockenspiel where the lowest, highest, and middle A note were played. A rubber mallet was used, and for each recording the mallet arm was positioned at a constant height (about 4 inches) above the glockenspiel and was allowed to swing under the influence of gravity to have the hammer strike each key in a consistent fashion.

The audio experiment was created using Unity 2019.3.1f1 with the Mixed Reality Toolkit (MRTK) provided by Microsoft. The application was deployed to a HoloLens 2 (HL2) augment reality (AR) headset. The HL2 has integrated HRTF hardware to enable spatial audio playback with minimal software overhead. Audio source objects are used in Unity to associate sound files with audio playback. The audio source instances which use the HRTF capabilities of the HL2 have their default settings modified by checking ***spatialize*** and moving the 2D-3D sound slider to 3D – indicated by a value of 1.0. Additional spatialized features were not used, nor would some be applicable to the experiment – e.g., simulated doppler. This version of Unity is using the deprecated extended realty (XR) pipeline with MRTK, and automatically uses the HL2 HRTF hardware as long as spatialized sounds are active with the 2D-3D setting at 1.0.

3.2 Detailed Single Microphone Practices

The audio was recorded using a single standard cardioid condenser microphone, an Audio Technica AT2020. A condenser microphone was chosen to pick up a vast range of sound as gain was increased on the DAC. Because condenser microphones are provided phantom power to amplify the input from there oscillating diaphragm, sounds could be recorded at distances further from the face of the microphone – an atypical practice for standard condenser microphone usage. This ensured that audio would still be captured despite recording from angles not in a direct line of sound waves – for example, capturing the speaker's voice saying hello while the speaker

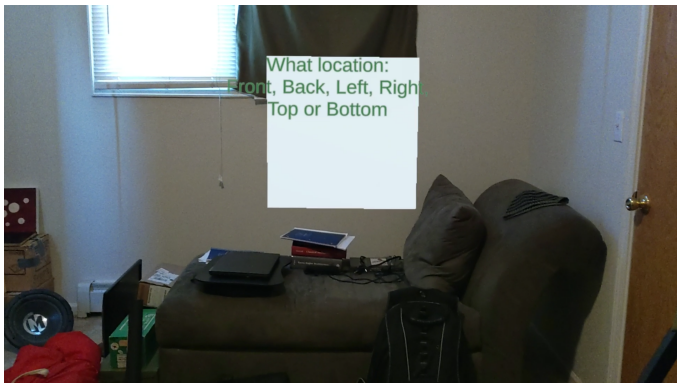

Figure 1: Image of experiment

faced away from the microphone. The cardioid microphone ensured minimal outside noise was picked up in directions away from the face of the microphone. An additional metal deflecting pop filter was used to ensure violent plosives did not agitate the diaphragm excessively while high gains were active.

Within multiple takes, the speaker repeated the word ***hello*** from the six angles previously described. The metal pop filter was always 2 inches from the microphone and the speaker was recorded 6 inches from the pop filter where the distance of the speaker was measured from a center point along the cross-sectional plane for each respective recording angle: front, from nose; back, from occipital lobe, left and right, between the cheek bone and ear; top, center of the top of the head; and bottom, between the neck and chin. For front, back, left, and right recordings the speaker was kept at a consistent height as they spoke from a seated position. Recordings from the top and bottom directions involved the speaker kneeling and rotating their head around the microphone in order to face the upward and downward directions.

The sounds for the observational study were produced by moving the microphone around a glockenspiel which was struck with a rubber mallet. The glockenspiel was chosen because it has a consistent resonance but also has an obvious fallout heard as the initial energy from the mallet strike fades. Three keys were recorded, the highest and lowest note, as well as the middle A note. To capture consistent sounds around the glockenspiel, each recording was of the mallet being held 4 inches above the key to strike which was then accelerated into the key merely by gravity. The microphone was kept 8 inches from each key, respectively, during recordings.

3.3 Experimental Design

Five experimental procedures took place using the speaker recordings. In each experiment the participant was sitting and observing an augmented cube added to their environment. The participants were not trained before the experiment in any way to calibrate themselves to the generalized HRTF effect. The primary purpose of the cube was to let a participant approximate the amount of distance away from them a sound would be playing and to possibly interpret that the cube would muffle sounds in the event that it was occluding an audio source. The cube was one unit length in each dimension within Unity - translating to an approximate cubic meter in real-world scale. Each face of the cube had an audio source object centered to it - therefore the audio sources were half a meter radially away from the center of the cube. The distance of the cube and audio sources were determined to provide ample hints to at least the left and right HRTF rendering. The cube was positioned 2.5 Unity units away from the participant - amounting to about 2.5 meters. Each experiment would iterate through the six audio sources to play for the participant, where each audio source was played three times before iterating to

the next one. The experiment would present a countdown to the participant before playing back audio. A brief pause would occur after the sound was played before playing it a second and third time. The system would pause after the third playback and the participant would verbally tell the researcher which side of the cube the sound was on: either in front, behind, on the left, right, top, or bottom. The researcher would record the response and then press a button to move to the next audio source iteration. The five experiments were defined within five separate cubes, the only difference between the cubes were if HRTF was on and which sounds were located at each of the six faces of the cube. When all six sounds were played, the system would swap out the current cube to the next until all trials were concluded.

The first experiment was the control. The audio sources use only recordings from the forward-facing speaker – i.e., the speaker talking directly into the microphone. These audio sources do not utilize spatial sound and playback with equal volume between the right and left stereo speakers. This ensures no stereo panning would occur during playback while the HRTF was off.

The second experiment had oriented audio sources that also did not use the HRTF. The oriented audio was arranged in static positions for all user, but was blind to the researcher to ensure no accidental cues influenced participants during the active experiment. The actually arrangement of sounds were examined after collecting all experimental data and is shared in the results. Future experiments will consider randomizing the arrangement of oriented sounds to better assess user responses to certain sounds; the current methodology is a meant as a proof of concept to briefly determine if front and back facing orientation of a speaker – without visual cues of the speaker – affect where a participant considers sound to be placed in the scene.

The third experiment uses the same set of forward-facing recordings as the first experiment, but with the spatialized features turned on to utilize the HRTF hardware. The fourth experiment uses the same oriented audio recordings and positions as the second experiment, but with the spatialized audio turned also turned on. Again the arrangement of the directional audio was not known until post experiment. The fifth experiment used the directional audio with HRTF and purposefully had the sound clips aligned to their corresponding faces on the cube, e.g., front to front, left to left, top to top, etc.. This was meant to see if there was any recognizable coherency for participants if the directions seemed to match the side they would be heard from - assuming a speaker talking was facing in each of those face directions while talking. The results collected were analyzed for accuracy in the participant's ability to recognize the location of sound based on no HRTF, HRTF, and directional recordings of sound with and without HRTF.

An observational study was included to qualitatively gauge the level of immersion when playing with the proposed method of observing which set of sounds along a spherical set of recordings should be mixed based on local and global states. The method of determining the sounds to mix and at what amplitudes was based on the number of recordings made with each glockenspiel key. The six recordings were axis-aligned for three-dimensional space, and because the nearest surface was considered the most relevant part to render from, three dot products were examined to determine which octant around the graphical cube the participant was nearest. The dot products being of the normalized head to center of cube vector with the normalized xy, yz, and xz planes respectively. The positive and negative values determine which octant the participant is most likely in, e.g., front top left. The uniform shape of the sound is described as a uniform sphere, but could be more irregular based on the latency between directions for sound to propagate, like under the glockenspiel where more occlusion is present. To linearize a normalized amount to weight against the three sounds corresponding to an octant, the audio locations were treated as points on a triangle

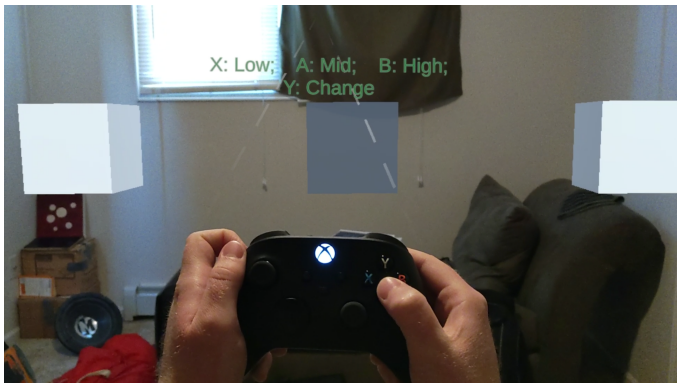

Figure 2: Demonstration following experiment

and a barycentric projection of the head position was made into this triangular. This ensures that, as a normalized value, the linearization of a curve is preserved in the project space. The weights of the barycentric values were then applied to the volumes of these three sound sources, with the remaining sound sources set to 0. This ensured that the sound amplitude never exceeded 1 in the combination of sound sources simultaneously played. The system would then readjust these volumes while the sounds were still playing so the participant could move around and actively hear the mixing of the direction audio.

This demonstration presented three small boxes for the participant to walk up to and around. The left most used the audio from the low glockenspiel key, the middle of the mid A note, and the rightmost of the highest key. The participant was given a portable keyboard to then play each note on a keypress - with one keyboard key per note key. The researcher guided the participant through the experience to ensure the participants understood what they could attempt to experience in this demonstration with engagement and interest based on participants verbal response.

The accuracy results of the five experiments were used to verify whether HRTF spatialization aided in determining audio locations compared to the controlled method. Experiments two, three, four, and five were compared against each other to see if audio orientation had an effect on sound location – in the non-HRTF case – and whether the combination of oriented sounds and HRTF could be found to assist in distinguishing where an audio source was located.

4 Results

This preliminary investigation had four participants, all completing every trial and spending time playing in the demonstration system afterward. Participant 3 was the only person that reported having hearing problems - regarding slight deafness. Participant 4 was the only participant that reported previous experience with a glockenspiel. Table 1 shows the order of sounds that were played and the position where those sounds were located on the cube. The pair is specifically the direction clip followed by cube face. F: Front, O: Rear/Back, B: Bottom, T: Top, L: Left, and R: Right. N/A is present for audio which was not spatialized in any way.

Exp.	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6
0	F,N/A	F,N/A	F,N/A	F,N/A	F,N/A	F,N/A
1	L,N/A	T,N/A	F,N/A	B,N/A	R,N/A	O,N/A
2	F,F	F,B	F,R	F,L	F,O	F,T
3	B,L	O,F	R,T	T,B	F,O	L,R
4	F,F	O,O	L,L	R,R	T,T	B,B

Table 1: Experiment by row, directional clip and cube face

The trials were configured to play in the order of no spatialization to start introducing spatialization in order to validate the effectiveness of HRTF. In this design, the control was always played first and most participants verbally started questioning whether they were hearing the same sound during playback. Table 2 is provided to demonstrate that most participants had a consistency in their guesses of direction during this control stage.

Part.	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6
1	F	F	F	F	F	F
2	F	L	F	B	F	L
3	L	F	F	F	F	F
4	L	T	T	T	F	F

Table 2: Participant response in control

Given the small sample of participants, a total figure is provided in Table 3 showing the overall accuracy of each participant in each experiment. Accuracy here is for matching the directional sound with the stated direction for experiment 0 and 1 because neither of these had any HRTF usage. The remaining experiments are accuracy in response of direction based on the actual spatial location of these sounds because HRTF is played. Table 4 shows all participants accuracy for the fourth experiment by participants' stated directions and the directional audio source played. The average of Table 4 is 37.5 percent. Experiments start in a base zero order to emphasize that the control as the zeroth experimental conditions.

Part.	Exp. 0	Exp. 1	Exp. 2	Exp. 3	Exp. 4
1	100	16.67	16.67	0	66.67
2	50	33.33	66.67	33.33	50
3	83.33	33.33	50	50	50
4	33.33	50	33.33	16.67	66.67

Table 3: Participant accuracy percentage per experiment

Part.	Exp. 3
1	33.33
2	33.33
3	33.33
4	50

Table 4: Participant accuracy in Exp. 3 by directional sound

Table 5 shows the overall average of accuracy as a percentage between all participants in all experiments using the accuracy conditions mentioned for Table 3.

Table 6 demonstrates the front/back accuracy between experiments after the control test.

The demonstration piece had varying responses by participant. The first participant enjoyed the interactive experience and was surprised by its effectiveness to dynamically convey sounds by orientation, but they felt they could not intrinsically walk through the cubes as they felt like real-world objects. They suggested that they would be more inclined to heard sounds between the cubes if they are spaced apart enough to walk between them. The second participant enjoyed the experience but did not like the short-lived resonance of the highest glockenspiel key. The third participant favored the lowest and highest key over the middle note, and was captivated by the experience. The fourth participant stated they thoroughly enjoyed the experience and started trying to play music with just the three notes while moving around the cubes.

5 Discussion

The layout of the experiment seemed to be effective in ensuring that initial HRTF cues did not immediately start to alter the perception of a participant. Future experiments with additional participant should likely warrant randomizing the experiment order, but the data suggests that isolating the HRTF spatialization to after playing a control and directional sounds had an effect on participants. Notably, the participants all seemed to come to a conclusion that in the control they were hearing essentially the same sound. This was further promoted as the case as participants actively questioned aloud whether they could discern localized audio during this control experiment. Immediately after the control, the direction sounds without localization appear to have had some impact on participants. While the accuracy varied between participants, the data seems to indicate that they felt inclined to think different sounds were playing from different locations - despite no HRTF modification to the audio. The third experiment introduced HRTF with the front facing directional recording, and participants continued to be in a position where they started guessing, in general, more directions when compared to the control. The fourth experiment continued to promote a variety of guesses in direction as directional sounds were played with HRTF. The last experiment coherently aligned the direction sounds to their respective cube locations resulting in the second highest overall accuracy after the control.

Experiment	Accuracy
0	66.67
1	33.33
2	41.67
3	25
4	58.33

Table 5: Average of participant accuracy as percentage in experiments

Exp.	F Audio	F HRTF	O Audio	O HRTF
1	50	N/A	50	N/A
2	N/A	50	N/A	0
3	75	0	75	0
4	100	100	75	75

Table 6: Front/Back accuracy by directional sound vs. HRTF

The validation of HRTF as a spatialization technique to promote localization in a 3D space seems to be confirmed by the results of the experiment. While the overall accuracy of the control experiment was high, it was based on the user's guess to the directional sound played and further could not be directly compared as an accuracy because it had no spatial basis. What can then be analyzed is that participants verbally questioned their own abilities to hear spatialized sounds around the cube, and most participants lingered around the same directional guesses during the control. While the second experiment (experiment 1 in tables) offered directional sounds and seemed to provide sensory reasons to start guessing more directions, these without HRTF did worse on average between all participants when compared to the single sound source using HRTF in the third experiment (experiment 2 in tables).

Table 6 seems to suggest the the hypothesis was accurate in that participants had an easier time localizing front/back sounds by the directional audio cures rather than with HRTF using a single audio source. In particular, the third experiment (experiment 2 in tables) had a 0 percent accuracy for correct guesses in the back/rear direction of the cube whereas the second experiment (experiment 1) without HRTF had 50 percent correct guesses for both the front and back directional sounds. The fourth experiment (experiment 3) indicated in both Table 4 and 6 further promote that the directional sounds actually improved front/back accuracy when combined with HRTF. Furthermore, Table 4 indicated that participants were mostly cued into directional sounds rather than HRTF when comparing the 37.5 accuracy based on directional audio guesses compared to the 25 percent accuracy for the same experiment (experiment 3) shown in

Table 3. While it seemed the front/back directional sound cues did provide better representation of front/back sounds during playback, it was unexpected to see that combined with HRTF the localization of front/back was further promoted, and moreover that participants might have been more cued to directional sounds - indicated by the fourth experiment (experiment 3) - than HRTF.

Observing the total averages of accuracy percentage in Table 5 further promote that directional sound sources played a significant role in user perception. While HRTF had a higher accuracy when compared to the second and fourth experiments (experiment 1 and 3) - including the fourth experiments analysis by 3D position and directional sound played - the combination of sound directions with HRTF when aligned in a coherent fashion - coherent soley by position as the cube is uniform and offers no visual cues for direction - seemed to provide the greatest localization among participants.

While the initial front/back hypothesis was found to be correct, the amount of influence directional recordings seemed to have with HRTF seems to indicate that further investigation in this topic should be warranted. In particular, when not coherently aligned by 3D position, participants seemed to make judgements more on the directional sound than in the HRTF, and when coherently aligned had substantial accuracy when determining direction by a uniform graphical reference with remarkable accuracy over using just HRTF with a single audio source.

The level of engagement found in the engineering demonstration with the glockenspiel indicates that there might be something valid in providing a novel method of starting audio rendering from for spatialization. Given the history of 3D audio being designated by methods used against either spatialized single point source sounds and positional static but orientable soundscape recordings, it would be argued that this engineering method combines orientable volumes as well as spatialized translations - six degrees of freedom - to experience audio in an interactive real-time setting, and might therefore fit the descriptor of 3D audio more so than the prior methods. This is a bold claim, but the participants seemed more enthusiastic about the combined volume of sounds heard rather than just HRTF. However, further exploration to test immersion must be performed. The experimental design and this demonstration piece were disjoint, yet the experimental data indicates that participants made guesses based on directional audio rather than HRTF as indicated in the fourth experiment (experiment 3 in tables).

This preliminary investigation indicates promising results, and exploring existing research has indicated potential experimental design to employ in future uses. Additionally, there might be a justification in repurposing the existing usage of psychophysics as it has been used in graphics to determine how auditory sensory information affects the mental perception of humans [4].

6 Limitations

The setup for multi-directional recording would have been better tested with an actual microphone array configuration pointing inward towards the speaker and the glockenspiel. The eight attempts at saying hello from all six angles resulted in obvious visual deviations - in a graphing of the audio - that were hard to determine as being maligned configurations of saying the same word, or actually within some margin of error within each take but observed to differ in their graph representation due to the complex interactions of wave propagation caused by speaking. This is to say, there is no clear indication as to whether a true coherency existed between recordings of speaking towards, away, to the left and right, and up and down from the single microphone because some change in wave propagation should occur due to diffraction around the mouth and head. Because of this, no attempt was made to dynamically mix the spoken words in the experiment in the same way that the glockenspiel was, but given the directional audio seemed to cause some type of cuing to participants it is likely worthwhile to lead this work to eventually testing the volume based methodology proposed with more complicated natural and commonly experienced sounds.

Audio reproduction and capture is similar to graphical reproduction and capture, without being in the presence of the device the sensory information will differ. Any accompanying recordings from this research will lack the qualities of the initial presentation. However, the general concept can easily be reproduced in software that feature spatialized audio mixing - like Unity - to experience the general effect. However, an extreme amount of effort went into getting six sounds that seemed to align even under reproducible conditions with a single microphone, so it is advised to utilize an array configuration if available.

As a preliminary study to examine the merit and prospects of researching this audio topic, some additional information was lacking that would be useful for future experiments. Notably, the setup for the test environment was not made based on standardizations of sufficient audio amplitude ranges for general human participants. The size and spacing of the cube(s) for the experimental trials were sized to ensure that some sense of separation was noticeably present while also ensuring that the audio sources along the faces would be separated well enough to convey left and right HRTF effects well when a participants head was forward facing and idle. This was also constrained by ensuring that while the cube would be large to best highlight the HRTF, it was also not so large that a participant would not be able to see much else of their environment. Under these conditions, the audio was made to a specific amplitude to be heard while wearing the HL2, but investigating the accessible ranges for amplitude by quantity would be a worthwhile venture for future experiments.

How many data points are relevant if this system is meaningful for real-time rendering of audio? A tetrahedron configuration is likely harder to setup with equidistant microphones, but can substantially reduce the memory required to produce these mixed sounds. It should be considered, and explored, whether this reduction has a meaningful effect on this audio rendering. For instance, the tested method essentially is a recording configuration of two tetrahedra, which potentially has more recordings of ***plosives*** - i.e., with more microphone configurations more opportunities exist to have captured a sound with stronger incidence to a notable sound wave. In high density systems, searching for the relevant space on a surface to yield a barycentric projection from three sounds is less of a concern as arbitrary broad phase graph structures could be used to search with minimal computational power, and in a continuous representation should reduce the entire operation to projecting to a point on a sphere without the need to mix but requires an infinite number of sounds captured.

In the discussion, a claim is asserted strongly about the effect directional audio had when combined with HRTF, and without graphical cues to depict visual directions for a participant the directional audio seemed more meaningful for a participant's direction guesses. This was only one trial in the experiment, and with a small sample size of participants. The strong claim is made based on this minimal evidence in the discussion to acknowledge where the preliminary study should be expanded in future research. Of course, increasing the number of trials would better indicate whether this is a hard truth - at present, it is definitely not hard but the evidence is taken still to suggest where to explore next.

This preliminary investigation meant that understanding the best tools and concepts for an experimental setup were lacking. Hindsight suggests that this experiment fits into the field of psychophyscis and could repurpose statistical methods and experimental practices designed from publications on the matter, such as that in SIGGRAPH [4]. With the small number of participants to begin this research, more sophisticated statistical analysis may not have been practical for the time being because general trends are less likely to reveal themselves with such a small sample group.

7 Conclusion

Two things were tested in this research: first, a question of whether front/back values could be better portrayed with directional recordings; and second, if engineering a volume based model of sound sources could be done while remaining coherent, and if it provided meaningful engagement. The former was found to be true with and without HRTF, thereby indicating the directional recordings had an interpretable meaning for participants. This brought an unexpected result in which participants might have been more cued by directional recordings rather than HRTF. The engineered solutions seemed to offer engagement to participants, but further comparative tests should be performed before attempting to assert any definite conclusions about the novel model. However, this model was successful at running in real-time without any noticeable performance degradation and was designed to be employable to existing interactive 3D audio systems as a point to start 3D audio processing rather than forcing existing designs to be changed and overhauled. The investigation into past and present audio research has suggested how future investigations concerned with this model should be explored, including an extension from system modeling/simulation and human computer interaction to psychophysics.

Acknowledgments

The research thanks Dr. Francisco Ortega, Adam Williams, and Xiaoyan Zhou in their support to explore HCI.

References

[1] J. Ahrens, M. Geier, and S. Spors. The interactive soundscape renderer for loudspeaker- and headphone-based spatial sound presentation. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, VRST '17. Association for Computing Machinery, New York, NY, USA, 2017. doi: 10.1145/3139131.3141779

[2] Z. Ben-Hur, D. L. Alon, R. Mehra, and B. Rafaely. Binaural reproduction based on bilateral ambisonics and ear-aligned hrtfs. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:901–913, jan 2021. doi: 10.1109/TASLP.2021.3055038

[3] C. Bossalini, W. Raffe, and J. Andres Garcia. Generative audio and real-time soundtrack synthesis in gaming environments: An exploration of how dynamically rendered soundtracks can introduce new artistic sound design opportunities and enhance the immersion of interactive audio spaces. In *32nd Australian Conference on Human-Computer Interaction*, OzCHI '20, p. 281–292. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3441000.3441075

[4] J. A. Ferwerda. Psychophysics 101: How to run perception experiments in computer graphics. In *ACM SIGGRAPH 2008 Classes*, SIGGRAPH '08. Association for Computing Machinery, New York, NY, USA, 2008. doi: 10.1145/1401132.1401243

[5] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties. Mpeg-h 3d audio—the new standard for coding of immersive spatial audio. *IEEE Journal of Selected Topics in Signal Processing*, 9(5):770–779, 2015. doi: 10.1109/JSTSP.2015.2411578

[6] O. B. Kaul, K. Behrens, and M. Rohs. *Mobile Recognition and Tracking of Objects in the Environment through Augmented Reality and 3D Audio Cues for People with Visual Impairments*. Association for Computing Machinery, New York, NY, USA, 2021.

[7] E. Lakka, D. Brutzman, R. Puk, and A. G. Malamos. Extending x3d realism with audio graphs, acoustic properties and 3d spatial sound. In *The 25th International Conference on 3D Web Technology*, Web3D '20. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3424616.3424709

[8] C. H. Larsen, D. S. Lauritsen, J. J. Larsen, M. Pilgaard, and J. 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. Association for Computing Machinery, New York, NY, USA, 2013. doi: 10.1145/2544114.2544118

[9] D. A. Mauro, R. Mekuria, and M. Sanna. Binaural spatialization for 3d immersive audio communication in a virtual world. In *Proceedings of the 8th Audio Mostly Conference*, AM '13. Association for Computing Machinery, New York, NY, USA, 2013. doi: 10.1145/2544114.2544115

[10] N. Moustakas, E. Rovithis, K. Vogklis, and A. 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, p. 220–227. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3395035.3425325

[11] K. Semionov and I. 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 Conference on Audio Mostly*, AM '20, p. 22–29. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3411109.3411119

[12] J. Sodnik, S. Tomazic, R. Grasset, A. Duenser, and M. Billinghurst. Spatial sound localization in an augmented reality environment. In *Proceedings of the 18th Australia Conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments*, OZCHI '06, p. 111–118. Association for Computing Machinery, New York, NY, USA, 2006. doi: 10.1145/1228175.1228197

[13] I. Trowitzsch, C. Schymura, D. Kolossa, and K. Obermayer. Joining sound event detection and localization through spatial segregation. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 28:487–502, jan 2020. doi: 10.1109/TASLP.2019.2958408

[14] N. Vryzas, C. A. Dimoulas, and G. V. Papanikolaou. Embedding sound localization and spatial audio interaction through coincident microphones arrays. In *Proceedings of the Audio Mostly 2015 on Interaction With Sound*, AM '15. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2814895.2814917