8-8-2023

Visual and spatial audio mismatching in virtual environments

Zachary Lawrence Garris
Mississippi State University, Zlg25@msstate.edu

Visual and spatial audio mismatching in virtual environments

By

Zachary Lawrence Garris

Approved by:

James Adam Jones (Major Professor)
J. Edward Swan, II
Michael S. Pratte
T.J. Jankun-Kelly (Graduate Coordinator)
Jason M. Keith (Dean, Bagley College of Engineering)

A Thesis
Submitted to the Faculty of
Mississippi State University
in Partial Fulfillment of the Requirements
for the Degree of Master of Science
in Computer Science
in the Department of Computer Science and Engineering

Mississippi State, Mississippi

August 2023

Name: Zachary Lawrence Garris

Date of Degree: August 8, 2023

Institution: Mississippi State University

Major Field: Computer Science

Major Professor: James Adam Jones

Title of Study: Visual and spatial audio mismatching in virtual environments

Pages in Study 42

Candidate for Degree of Master of Science

This paper explores how vision affects spatial audio perception in virtual reality. We created four virtual environments with different reverb and room sizes, and recorded binaural clicks in each one. We conducted two experiments: one where participants judged the audio-visual match, and another where they pointed to the click direction. We found that vision influences spatial audio perception and that congruent audio-visual cues improve accuracy. We suggest some implications for virtual reality design and evaluation.

DEDICATION

Dedicated to:

My wife: Jennifer Marie Garris

My parents: Bo Garris and Brandi Garris

My mother-in-law: Teresa Lampe

ACKNOWLEDGEMENTS

I would like to express my gratitude to Dr. J. Adam Jones for his support and guidance throughout the journey of completing my Master's at Mississippi State University. His kindness, patience, and immense knowledge throughout the process were invaluable to my success. His decision to invite me to become his research assistant as he transferred from the University of Mississippi has played a major impact in my life. I gained a large amount of knowledge from him and am grateful for his kindness and patience in teaching me, and I appreciate the opportunity he has provided me.

I am also grateful to my committee members for their valuable suggestions for my research work. Their valuable guidance helped shape my work.

I want to express my deepest gratitude to my beautiful wife Jennifer Garris, who has been by my side since high school and throughout this entire process. She is my inspiration and one of the most passionate people I know. She always encouraged me to keep going when I felt stuck or frustrated. Even though life has been hectic as we both finished graduate school, I look forward to starting a new chapter with her. As she graduates from Pharmacy school and I earn my master's degree, I can't imagine anyone else I'd rather share my life with. I love you and I cannot wait to see what my future holds with you! As Robert Browning once said "Grow old along with me! The best is yet to be."

I would like to thank both my parents, Bo Garris and Brandi Garris, for their support throughout my journey through college. They raised me with the knowledge to thrive. You have

taught me the value of hard work, perseverance, and kindness. You encouraged me to pursue my dreams through graduate school. You raised to know right from wrong and have shown me a model in which to base my life. Thank you for your years of supporting me through it all.

I want to express my gratitude to my wonderful mother-in-law Teresa Lampe, who is like a second mother to me. She has welcomed me into her family and treated me with kindness and respect. She is a remarkable and inspiring woman, just like her daughter, and has dedicated more than three decades of her life to educating the younger generations in science, technology, engineering, and mathematics (STEM). She has always encouraged and supported me in every step of my journey and for that I thank her. I am honored to be part of her family and I wish her all the best in her future endeavors.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1.1 Spatial Audio/Binaural Audio

In virtual reality (VR) systems, users can immerse themselves in a simulated environment that can be different from the real world. VR systems aim to provide realistic and natural experiences for users by stimulating their senses of vision, hearing, touch and sometimes smell and taste. Hearing is essential for creating a sense of presence and spatial awareness by helping users locate sound sources, identify objects, communicate with others, and navigate the virtual environment. To successfully mimic the real world, VR systems must provide spatial audio that simulates how sound waves interact with the user's head, ears, and surrounding environment.

Spatial audio can be delivered through different devices, such as headphones, earbuds, speakers, or bone conduction transducers. Earbuds are the most accurate, as the sound source is closer to the eardrum. Bringing the sound cue closer to the recording point provides more accurate sound localization cues.

1.2 Binaural Rendering

Binaural rendering is a technique that creates a stereo sound signal for each ear by applying filters that mimic the effects of the head and pinnae on the incoming sound waves. Head Related Transfer Functions HRTFs are mathematical models that describe how the head and pinnae modify the sound spectrum depending on the direction of the sound source. Ambisonics is a technique that encodes the sound field in a spherical coordinate system and then decodes it for a specific speaker or headphone configuration. Object-based audio is a technique that treats each sound source as an object with a unique position, orientation, and properties. Then it renders it according to the user's perspective and preferences.

1.3 Experiment One

This experiment aims to investigate the role of vision in spatial audio perception. I created a Unity VR application for the Vive Pro Two with models of locations around the campus of Mississippi State University. Each virtual representation included a real-world audio recording using in-ear binaural microphones to capture a series of clicks produced by a dog training device. Clicks were also recorded in the Raspet Flight Center's anechoic chamber. This was done to have an approximation of a near infinite acoustic environment. Each participant was placed in a virtual room and given a Vive Wand controller. The controller was programmed to receive yes or no inputs from the user. The seated participant was then instructed to play the audio through their earbuds by pulling the controller's trigger. After each click was played, the participant would determine whether they thought that click was recorded in this room and answer by selecting yes or no.

The clicks heard by participants were grouped by room as follows: small, medium, large, or infinite, and anechoic chamber. Five real recorded click locations at -90, -45, 0, 45, and 90 and four interpolated noise locations at -60, -30, 30, and 60 correspond to each real-world location. The participant heard each possible click combination twice and was transported to the next virtual environment. To avoid the sharp change of silence to click, each room had the ambient noise of the actual room played in its virtual counterpart.

1.4 Experiment Two

Experiment Two used the same hardware, locations, and click recordings as Experiment One. However, the participants were asked to determine the location of the clicks instead of deciding whether or not the clicks fit the visual environment. The clicks were not grouped by room, and the participant did not answer yes or no. Once again, the participant was given Vive Wand controllers, programmed to input their location in space along with the Vive headset. The participants were instructed to play the clicks again; however, in test two, they could be from any four locations at any given time, meaning they were no longer bound to play in room groupings. Each time a click was heard, the participant responded by extending their arm fully in the direction of the click and inputting the controller's location by button press. The participants would once again experience every click possibility two times before being transported randomly to one of the four-room locations until all rooms were visited.

CHAPTER II

RELATED WORK

Spatial audio (SP), which allows for “3-D audio” effects, is a powerful tool in the human perception of surroundings and a crucial mechanism for survival. It is an evolutionary advantage of many species that allows them to locate the source of an audio stimulus around them. Two primary essential component methods exist for the perception of localized audio. [14] The first method is the Interaural Time Difference (ITD) which is the time for an audio source to reach each ear. The second method is the Interaural Level Difference (ILD), the audio source's volume on each ear. For example, the head creates an acoustic shadow when an audio source hits the right ear, the head creates an acoustic shadow. The presence of the head between the audio source and the left ear is detectable in volume. This change in volume or level is detectable and is a key factor in locating the source of the audio. Manipulating these effects in the audio playback of headphones can create the illusion of directionalized sound. However, it is important to understand that audio sources typically originate with some environmental context, like vision. In most situations, the audio source and the environment are inseparable. The effect of the environment is usually expressed in the form of a sound's reverberation. Reverberation is effectively the result of overlapping echoes as the waves of sound from an audio source bounce around the environment before being heard by the observer.

2.1 Interaural Time Delay

One key contributing factor to the perception of a sound's location is how long an audio impulse takes to reach each ear from the source. The width of the human head is large enough to introduce a slight but noticeable difference in the arrival of an audio cue.

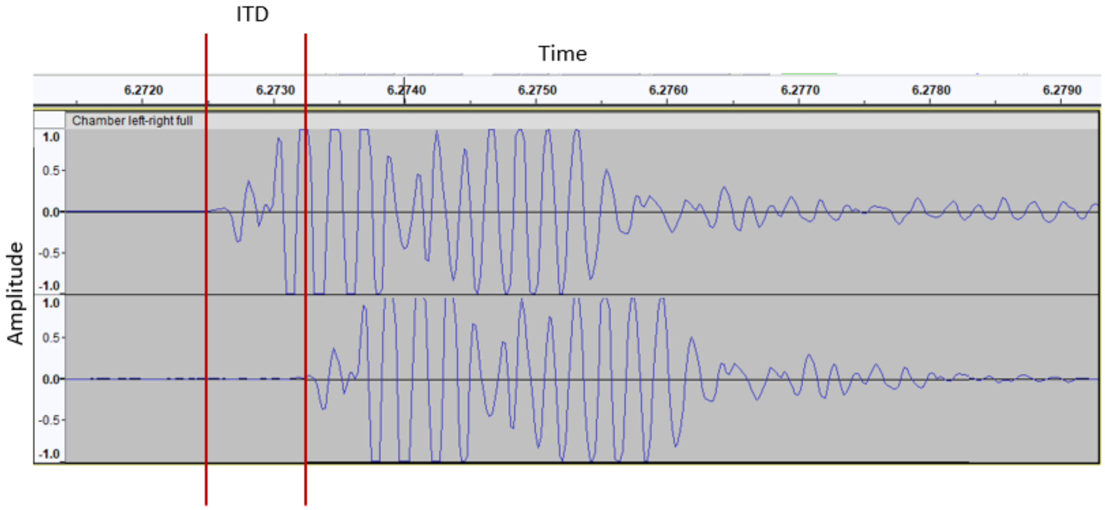

Figure 2.1 An Example Waveform of Interaural Time Delay from Raspet Anechoic Chamber

Interaural time delay plays a role in the perceptual accuracy of a sound source's horizontal and vertical location. Bronkhorst investigated humans' audio localization abilities for real versus virtual audio sources [2]. These virtual audio sources were implemented using a head-related transfer function (HRTFs are complicated functions of frequency and spatial variables used to mimic binaural audio sources).

Bronkhorst [2] created audio recordings with two Microtel M40 microphones equipped with probe tubes inserted into the ear canal until they made contact with the eardrum. A loudspeaker in different positions emitted a 70db impulse that was played through a set of Sennheiser HD 530 over-the-ear headphones because the audio source did not come directly from the ear canal. Bronkhorst [2] used a Finite Impulse Response Filter (2.1). The filter was meant to compensate for the headphone-to-ear canal transfer.

$$T(\theta, \emptyset) = \frac{F(\theta, \emptyset)P}{HS} \tag{2.1}$$

Given that F(θ,ϕ) and H are the Fourier Transforms of the Waveforms Recorded in the Free Field (For a Source at Azimuth θ and Elevation ϕ) and Under the Headphone, Respectively, the Fourier Transform T(θ,ϕ).

Bronkhorst [2] observed that both virtual and real audio impulses provided a mean offset of 5 deg on the horizontal axis. Bronkhorst's findings in the vertical plane differ by 7° with the virtual impulses, which were perceived at 8° and the real source at 13°.

Bronkhorst [2] hypothesizes this is due to the region in the center of the audio playback arc around the observer. Participants perceived little or no interaural time delay when a virtual sound was played from a source directly in front of the participant. Bronkhorst also found that the number of front-to-back reversals, where an audio source was played from in front, but perceived from behind, was nearly twice that of real sources, 11% to 6%.

Schwartz [10] found that, for example, for a speaker in a "cocktail party" with many sound queues overlapping, "ITD alone does not enable accurate segregation of sound sources from mixtures." The researchers found that relying on spatial cues alone can sometimes result in inaccurate sound segregation, mainly when the sounds are spectrally similar. They conducted a series of experiments using stimuli designed to mimic the complex acoustic environment of a cocktail party. They measured listeners' ability to segregate sounds based on spatial cues alone or a combination of spatial and spectral cues. Their results showed that interaural time differences alone were insufficient to produce accurate sound segregation when sounds had similar spectral content. Instead, combining spectral and spatial cues was necessary to accurately segregate sounds from their source.

2.2 Interaural Level Difference

When sound travels to the head from an audio source, it will encounter one ear closer to the sound source and then the head. The head impedes the wave on its way to the second ear causing an Interaural Level Difference (ILD). ILD occurs primarily due to sound waves bouncing off the head and into the ears and the attenuation of sound waves traveling a greater distance. The head acts as a baffle to sound waves. Depending on the size of the head, it can block a substantial portion of the sound waves. An observer can detect this difference in sound wave volume between the ears and use that information in conjunction with ITD to determine the direction and relative distance of a source of a sound origin. [2]

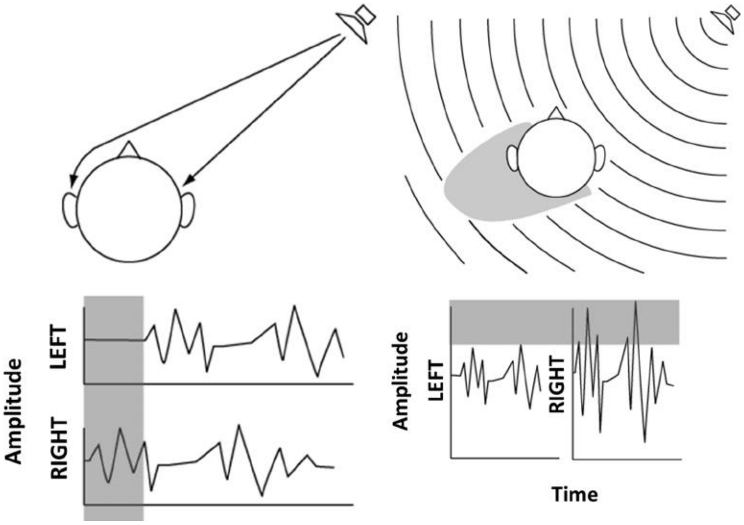

Figure 2.2 Acoustic Shadow Illustration [13]

ILDs are dominant for signals above 1500 Hz, ITDs are dominant for frequencies below 1000 Hz [1,12], and the average human head is ~7.638 inches in width [5], and a waveform of 1500hz is ~9.04 inches depending on atmosphere conditions. Under these conditions, an observer will lose the ability to differentiate the acoustic shadow since the wave is as long as or longer than the width of the head. The observer cannot detect discernible air pressure perturbation once waveforms are as large as the human head.

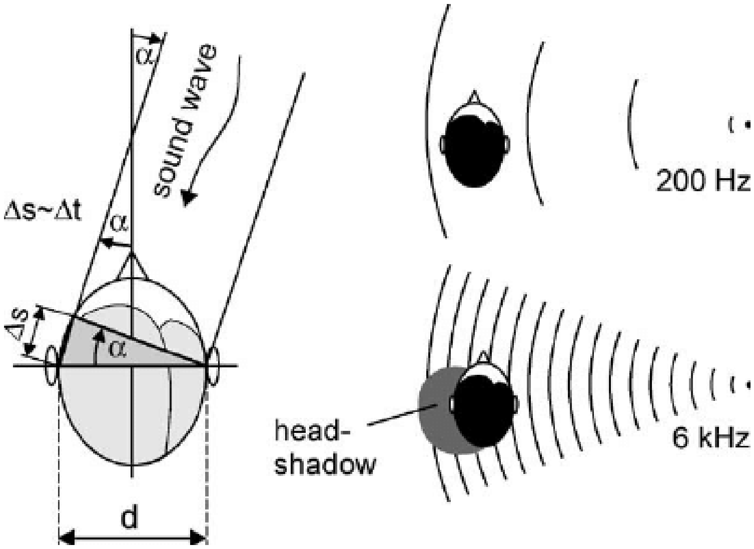

Figure 2.3 An Illustration of the Role of Acoustic Shadow with Large Waveforms [13]

Wightman and Kistler carried out localization studies utilizing unique Head Related Transfer Functions (HRTF) [2]. HRTFs are a signal processing technique used to simulate how sound is heard in a three-dimensional environment. HRTFs are based on how sound waves interact with the human head, ears, and torso. When sound waves enter the ear, they are filtered and shaped by the head, ears, and torso before they reach the eardrum. While HRTFs are believed to mimic the core components of “3D Audio” they lack the accurate reverberation that a room provides, and many, if not most, do not model the pinnae, torso, and head. While mimicking the core components of spatial audio, Interaural Time Delay, and Interaural Level Difference. HRTFs lack the reverb profile of recording space. Therefore they cannot duplicate the geometry of a room's reverb profile and its decay time.

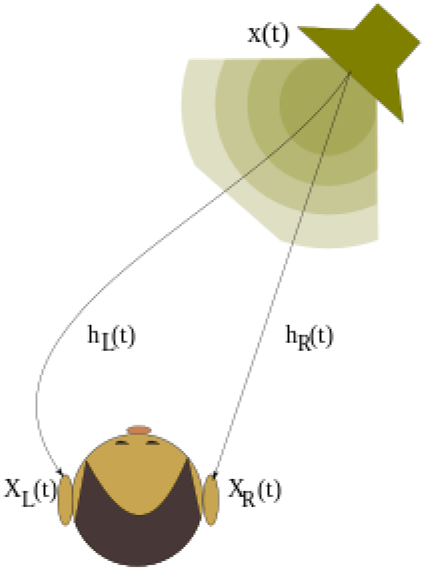

Figure 2.4 Head Related Transfer Function Illustration [I2]

Bronkhorst [2] found that the proportion of front-back reversals for virtual sources created with HRTF was approximately two times higher than for four non-virtualized sources (11% vs. 6%). Wenzel et al. [14], who previously ran the experiment with non-individual HRTFs, discovered 31% vs. 19% (front-back reversals) and 18% vs. 9% (vertical disparities) [2]. The “dead spots” in our perception create a cone of confusion in which spatial perception becomes inaccurate. Bronkhorst ran an experiment where participants were asked to identify the quadrant short stimuli originating using head pointing to mitigate the cone of confusion and investigate the possible causes of Wenzel et al.'s large difference in reversals. The large quadrants are located around the center of the region, not close to the edges. If a participant's spatial resolution was not accurate, the test would only show the results determined by pure confusion rather than limited spatial resolution.

2.3 The Effects of Reverb

When sound reverberates around a room, listeners will hear both the direct and indirect reflections of that sound off surrounding surfaces. These reflections are only able to be separated approximately by the brain. McDermott studied sounds collected from 271 locations around Boston, Massachusetts, by recording an impulse response (IR) with a loudspeaker and microphone to obtain a frequency response waveform.

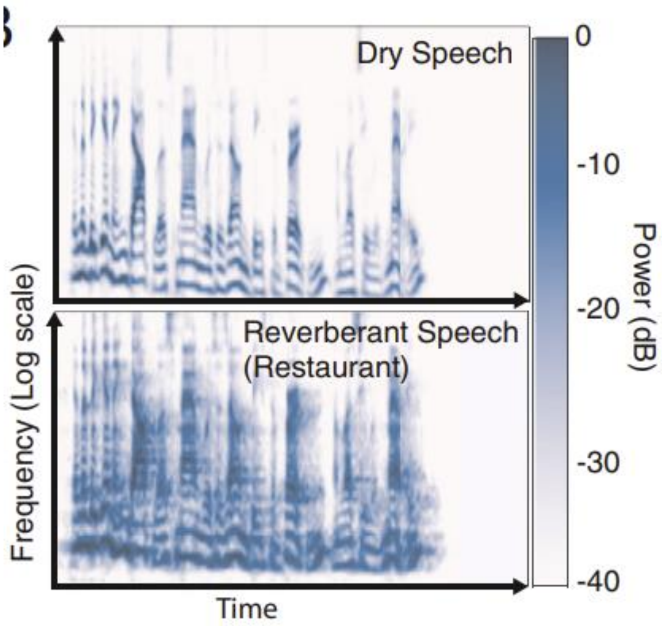

Figure 2.5 Reverberation Distortion of the Structure of Source Signals [11]

To examine the waveforms, they used “cochleagrams,” a graphing technique that filters the waveform to mimic the frequency selectivity of the cochlea.

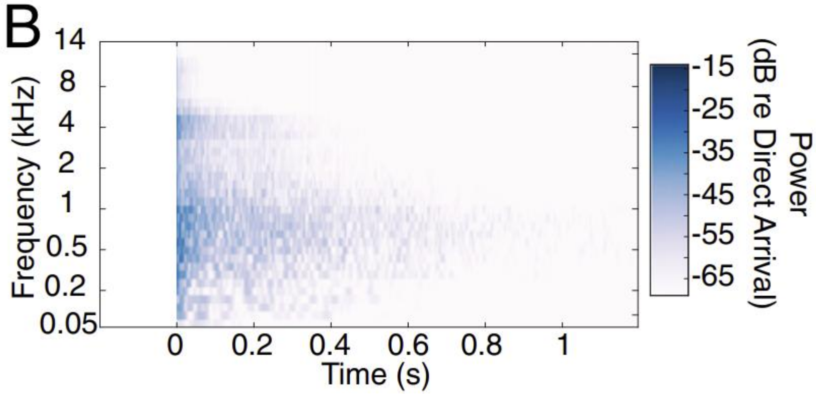

Figure 2.6 Cochleagram of the Restaurant Impulse Response from Fig. 2.5 [10]

Analysis of the amplitude from the filter revealed that the tail of the IR waveform decayed with a small number of high-amplitude echoes separated by brief periods of relative quiet. This phenomenon happened with considerable regularity across all locations. “The overall conclusion of our IR measurements is that real-world IRs exhibit considerable regularities. The presence of these regularities raises the possibility that an observer could leverage them for perception.” [10]. Unfortunately, McDermott only ran this experiment with mono recording equipment to isolate the tails of impulse responses. They found the tail exhibits the most significant distortion with the most regularity. When the tails of the waveforms were changed and participants went through the trials again, the early reflections of the sound were much less salient to results than the manipulation of the tail. “Collectively, these results suggest that the features revealed by our analysis of real-world IRs—a Gaussian tail exhibiting exponential decay at frequency-dependent rates—are both requisite and sufficient for the perception of reverberation.” [10] HRTFs do not encode these unique tails as they have no environment. This results in a wave with no scene geometry recorded into the waveform, only the directionality provided the equations used to generated by the HRTFs. The authors found the human brain separates the statistical predictableness of the environment from reverberant sound sources. “Collectively, our results suggest that reverberation perception should be viewed as a core problem of auditory scene analysis, in which listeners partially separate reverberant sound into a sound source and an environmental filter, constrained by a prior on environmental acoustics“[10].

2.4 Precedence Effect

An effect unique to spatial audio is the precedence effect, also known as the law of first arrival. The precedence effect refers to the tendency of the auditory system to prioritize the perception of the first sound in a series of reflections over the subsequent reflections, as shown in Figure 2.7. A few milliseconds difference can cause a bias to the first sound. The precedence effect is thought to be caused by the fact that the brain uses the first sound to arrive as a reference for the second sound and is thought to be a key component for localizing sound. The precedence effect is thought to be the reason an observer can localize sound sources inside a reverb-dense environment. If two audio impulses equally loud are played simultaneously, the precedence effect does not work, and the brain experiences a singular sound. This is known as audio fusion. The effect is prevalent with short delays in small enclosed spaces, which can lead to dimensioned audio localization accuracy.

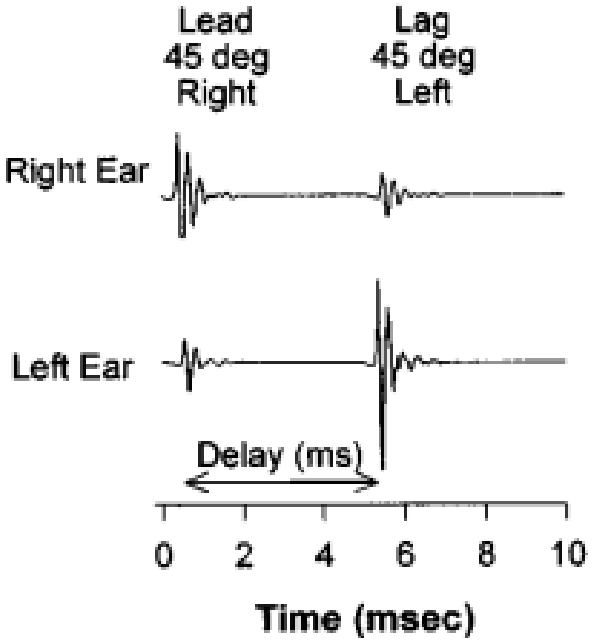

Figure 2.7 The Precedence Effect of Two Impulse Responses Reaching the Ear [4]

When a leading audio impulse is played with delays followed by a lagging source from 0-2 milliseconds through both headphones and loudspeakers, the leading source is interpreted as the source of the audio 80%-90% of the time, with the lagging source being interpreted as the source only 10%-20% of the time [4].

2.5 Spatial Updating

Researchers have found that participants can use language and 3-D audio to guide their actions through space with the same precision as if they were using their vision. Loomis et al. [6] were specifically looking at the task of spatial updating, which refers to the ability of observers being able to update their perceived position in an environment when moving through it. The study found that people can update an internal stimulus representation using spatial language (e.g., "2 o’clock, 16 ft”) and 3-D audio. The researchers found that while both effectively guided blind and blindfolded participants, 3-D audio provided nearly double the information rate of the target landmark locations. This was possible as Binaural audio cues could be played in much more rapid succession than spatial language cues could be given.

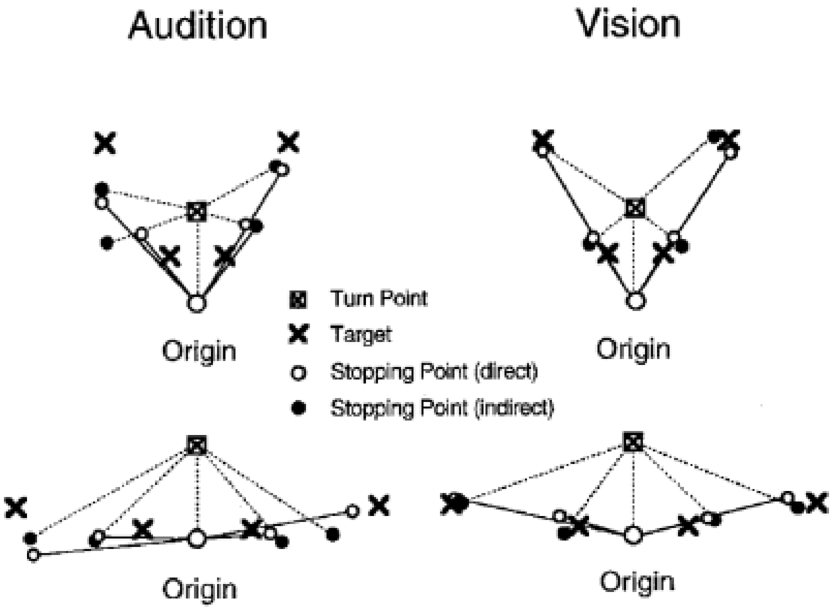

Figure 2.8 Stimulus Layout and Results of an Experiment on Spatial Updating of Visual and Auditory Targets. [6

Researchers Bronkhorst [2] and McDermott [10,11,1] have found ILD and ITD cannot provide adequate spatial information directly in front/behind and high/low target acquisition. Spatial language can be used as a useful supplementary information source. Figure 2.8 depicts that visual distance perception was more accurate than auditory distance perception. This technique could provide a quick and accurate view of one's surroundings while preventing information loss and reversals.

2.6 Synthetic Binaural Audio in Reverb Environment

The synthesis of Binaural audio in reverb-dense environments is necessary for use in some applications. Going to every location to record is impractical or impossible if a space is digitally produced. To test if synthesized audio of a reverb-dense environment is easily detectable, researchers Traer and McDermott [11] imposed different types of energy decay on noise filtered into simulated cochlear frequency channels. Three source types, impulse, spoken sentences, and synthetic modulated noise, were chosen for the experiment. Listeners were asked to identify which of the two sounds played were recorded in a real space. The task should be difficult if the synthetic IRs replicate the perceptually significant reverberation effects. Participants could not detect the synthetic IRs when the source was more complex, such as speech and synthetic modulated noise. "In some cases, the subjective impression was striking." IRs with spectrally inverted aspects seemed to contain two distinct sounds, a source with moderate reverb and a "hiss." The auditory system is apparently unwilling to interpret high frequencies that decay more slowly than low frequencies as reverberation. [11]

CHAPTER III

BACKGROUND

3.1 Binaural Recording Methods

In Binaural audio recordings, listeners rely on their headphones to accurately reproduce the recorded sounds. This ensures proper recreation of ILD, ITD, precedence, etc. There are three major methods for recording binaural audio. Method one uses two directional cardioids microphones angled ~110 degrees from one another, as shown in Figure 3.1. Cardioids are preferred, as the anti-phase lobes of hyper cardioids tend to give exaggerated width requiring a smaller space. The microphone pair is placed approximately but no more than 17 cm apart. This ensures the time delays are not too substantial and closest to that of an actual head.

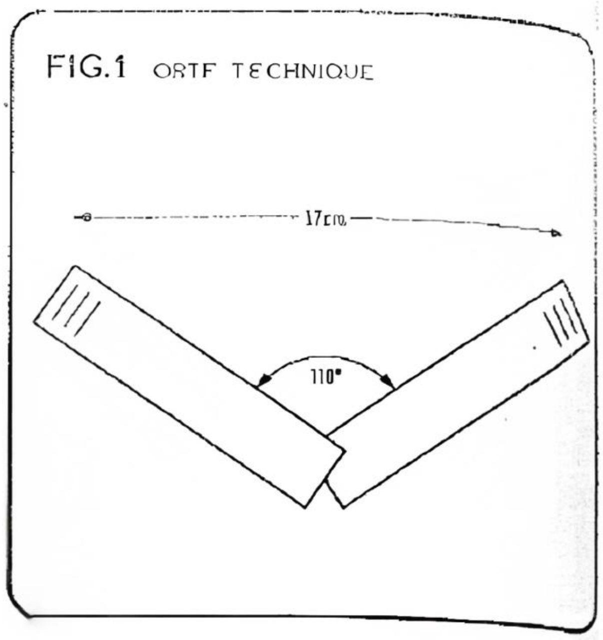

Figure 3.1 110° Cardioid Microphone Set [3]

Method two uses a dummy head with in-ear microphones. The ears of the dummy are shaped to match the population’s average pinnae and ear separation, providing the best possible result for most people. However, these heads can be costly and only approximate a real head. Ideally, each sound would be recorded in the ear of the head the audio would be played back in. This is to capture every aspect of the person recording the audio.

3.2 Free Field Recording Methods

Anechoic audio recording aims to provide little to no reverb. To have no reverb, a space must be infinite in size to provide no reflection sources. Historically there are two established methods for recording near free field audio, audio with little reverb. The researcher can travel to a tall free-standing building with no audio sources nearby. This method is simplistic and cost-effective but is often not useful for research settings where a recording must be clean of outside audio sources. The method allows the sound wave to propagate unopposed outwards with little to reflect the sound back to the originating source. The floor is the only reverb source, which can be mitigated using the sound-absorbing material, as in Figure 3.1. This method attempts to maximize the delay between arrival and direct sound.

Reduction of reverb can be accomplished in an open field with fresh snowfall or a similar material acting as a sound dampener.

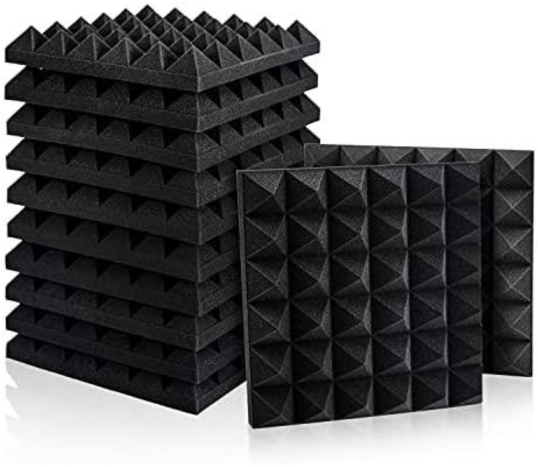

Figure 3.2 Acoustic Paneling

This material style is known as acoustic paneling and is used to “capture” audio waves, as shown in Figure 3.2. The purpose of the material is to take an oncoming sound wave and convert it into heat by moving a material denser than air, such as foam or fiberglass until the wave has lost its energy. If done properly, the waveform cannot return to the source and diminishes to an imperceivable level, which is used in an anechoic chamber, as shown in Figure 3.3.

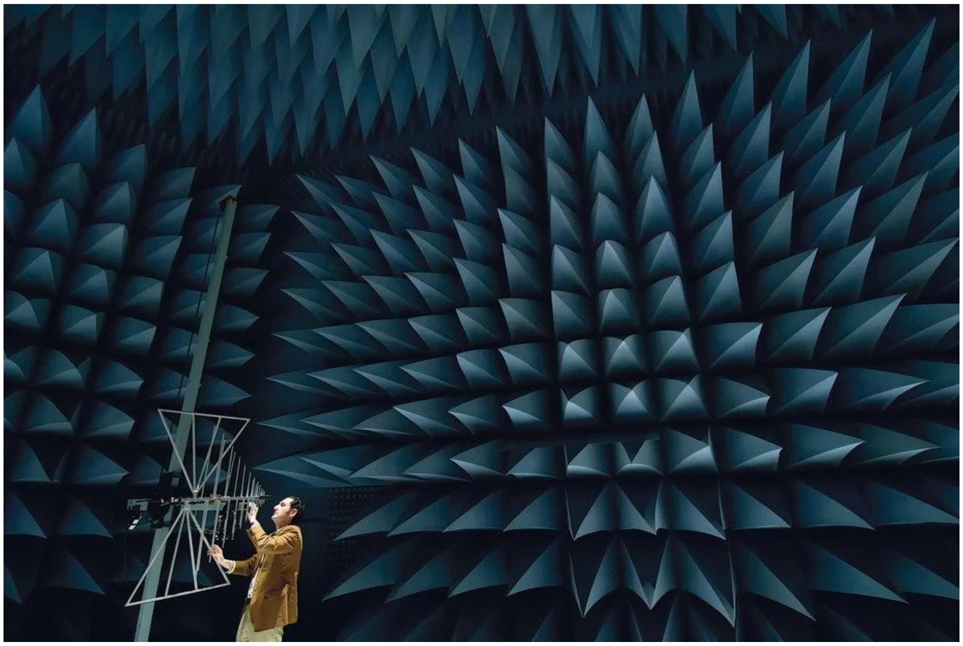

Figure 3.3 Anechoic Chamber

These chambers are ideal for research settings. A clean audio sample can be taken inside a chamber with minimal reverb and foreign audio sources. In a well-built chamber, the reverb will be suppressed to a nearly imperceivable level.

3.3 Vive Tracking System

The HTC Vive VR system uses an IR light to track its position in space [7]. The system includes a minimum of two base stations that emit the tracking IR light. These beams are picked up by sensors on the Vive headset and controllers, allowing the system to track its location in space. The base stations emit the IR light in a sweeping pattern, which is then detected by the onboard sensor of the Vive equipment.

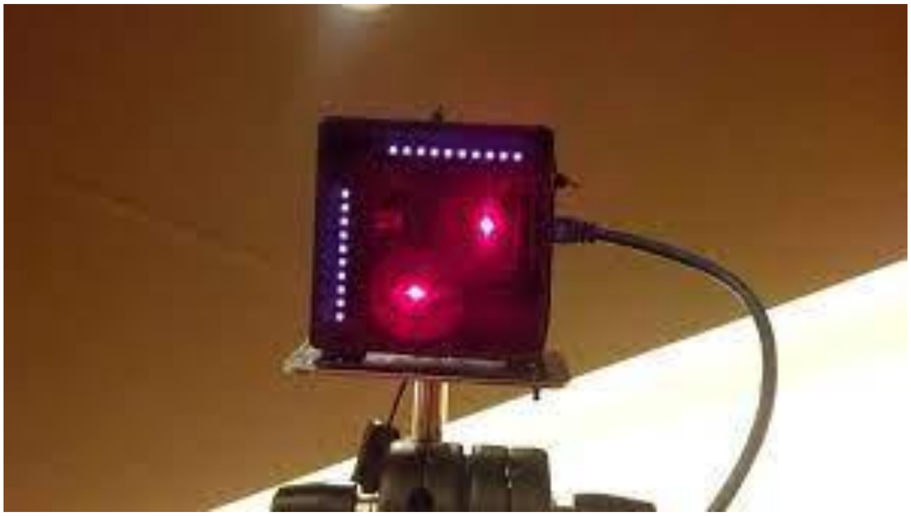

Figure 3.4 The Inside of Vive Lighthouse

These sensors then use the timing of the sweeping pattern to triangulate their position in the room at approximately 60Hz. Base stations also emit a second modulated IR signal that synchronizes the timing of the base station’s IR laser sweeps. The Vive lighthouse tracking system is a very accurate tracking system with high precision. However, the precision may vary depending on the specific environment and setup. Luckett et al. found that positional measurements can be submillimeter accurate in a smaller area with two base stations. [7] This precision is used to capture the location of the participant's responses by tracking the head and hands.

CHAPTER IV

EXPERIMENTAL OVERVIEW

This research aims to separate and manipulate aspects of audio and visualization perception variables through different scenes and reverb audio sources. Informal testing has shown that mismatched auditory and visual spaces in the real world could cause the observer to struggle to perceive the audio's originating location. Observers struggled to localize a previously recorded binaural audio track when played back in a different environment. This led to the hypothesis for my research.

4.1 Hypothesis

The unexplained phenomenon experienced in preliminary testing would hold over different environments in both natural and virtual worlds. The mismatching of the visual space with the auditory scene would cause perception issues with audio localization. Audio scientists have found that an observer localized audio with reverb with tools like the precedence effect [9,8]. These experiments were conducted in the same environment as the originating sound. This research aims to mismatch the environment with the audio to see if vision plays a role in spatial audio perception if so, to what degree.

4.2 Research Goal

Virtual Reality could provide an opportunity to conduct potentially impossible experiments in the real world. Matching and mismatching different audio and visual environments in real-time, tiny spaces like the inside of an SUV or small room is complicated. Thus, VR could be used to pursue the interlink between vision and spatial audio. This research is important because it represents a novel investigation into the impact of mismatched audio and visual representation on the localization of spatial audio for both VR and the real world.

Figure 4.1 HTC Vive Pro

By using the environment model in the Unity software coupled with recordings taken in varying environments with varying reverb, the goal is to gain an insight into how humans may be using vision to localize the audio in which they hear.

4.3 Variables

4.3.1 Independent Variables

The independent variables in this study are the room size, the reverb, and the click location. These are manipulated by the researcher to create different audio-visual conditions for the participants. The room size refers to the dimensions of the virtual environment, which range from small to infinite. The reverb refers to the amount of sound reflection and decays in the auditory environment, which can match or mismatch the visual environment. The click location refers to the direction of the sound source relative to the participant, which can be either real or interpolated. These independent variables are hypothesized to affect the accuracy of spatial audio perception in virtual reality.

4.3.2 Dependent Variables

The dependent variable in this study is the accuracy of spatial audio perception in virtual reality. This is measured by two different methods: the percentage of correct responses in Experiment One and the angular error in degrees in Experiment Two. The accuracy of spatial audio perception reflects how well the participants can locate and identify sound sources in virtual environments. It is expected to vary depending on the independent variables.

4.4 Experiment Logistics

4.4.1 Participants

Participants were recruited from the Mississippi State University population and the surrounding community. However, participants with a history of visually induced seizures or severe motion sickness were excluded from the study. Twenty participants were recruited in total 16 males and four females, with a mean age of 29). All participants were provided with an

MSU approved IRB-informed consent document before beginning the study. Participants were required to provide consent to participate.

4.4.2 Apparatus

Participants wore an HTC Vive Pro Two, with provided earbuds connected via an extension cable and 3.5mm connector, as shown in Figure 4.2.

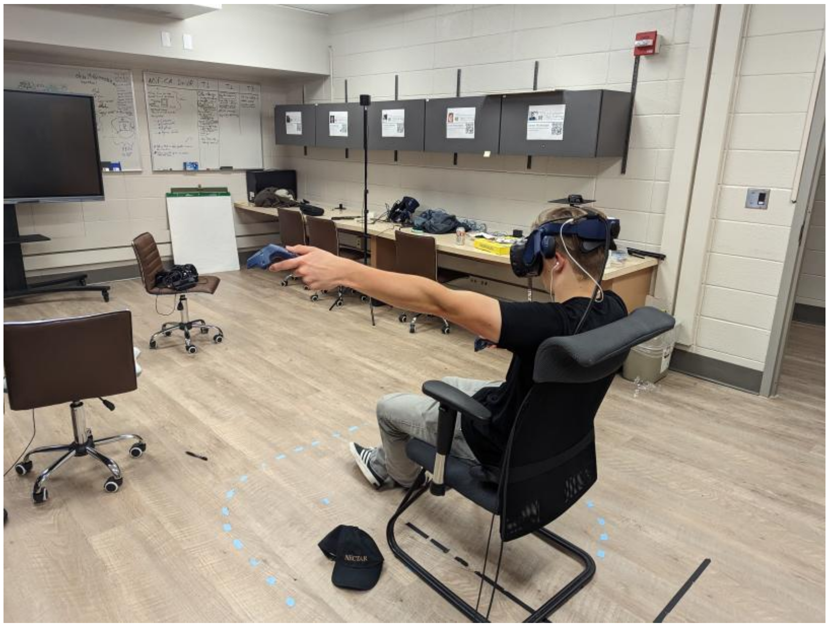

Figure 4.2 Participant Conducting the Experiment

In addition, each participant was given two HTC Vive Wand controllers for data input from each hand. These controllers, along with the headset, were monitored by a Lighthouse 2.0 tracking system shown in Figure 4.3.

Figure 4.3 HTC Lighthouse 2.0

4.4.3 Virtual Environments

Participants were presented with both matched and mismatched combinations of visual and auditory spaces based on four real-world locations:

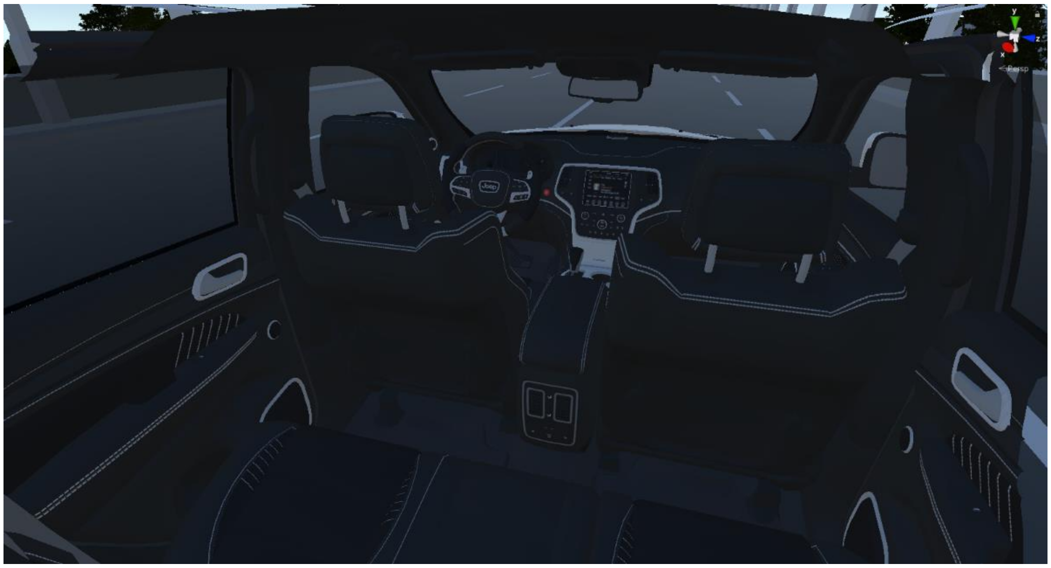

Figure 4.4 Virtual Cab of a Jeep Cherokee SUV

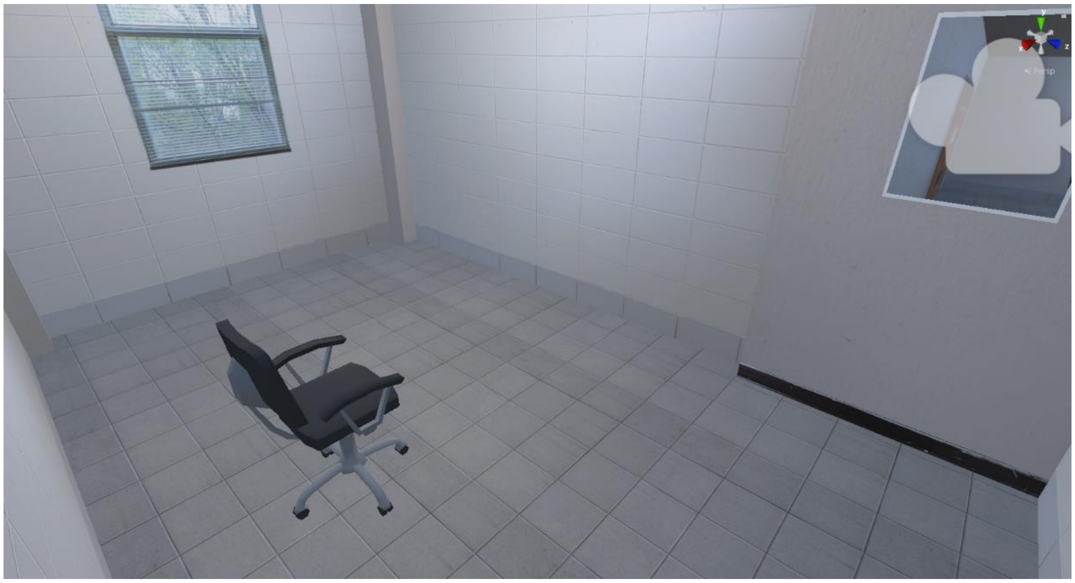

Figure 4.5 The Virtual Former Location of the Hi5 Lab in Rice Hall

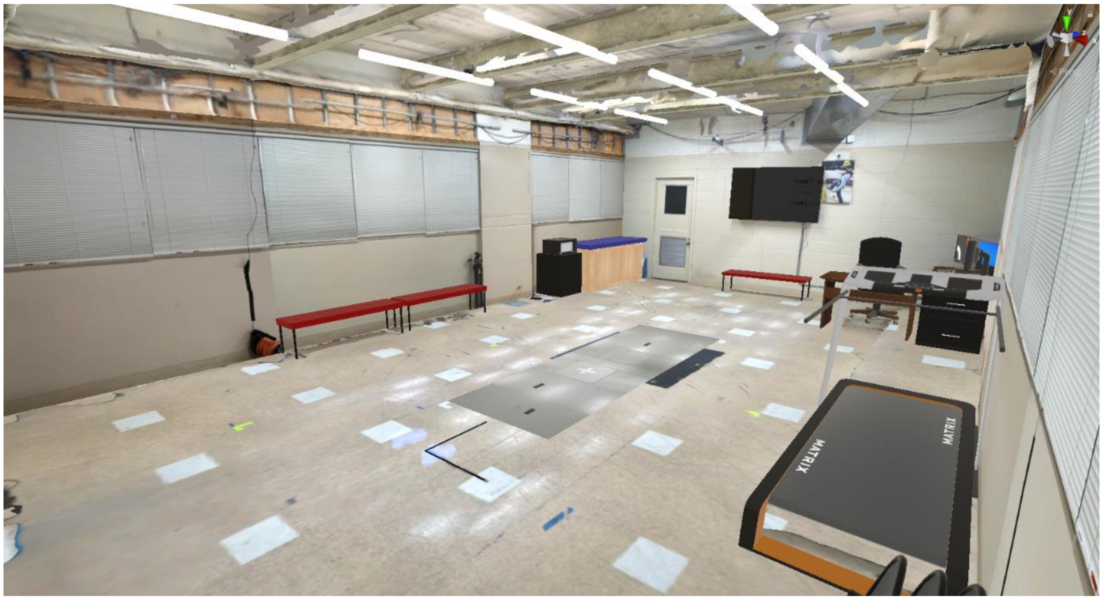

Figure 4.6 The Virtual Neuromechanics Lab on Mississippi State University’s Campus

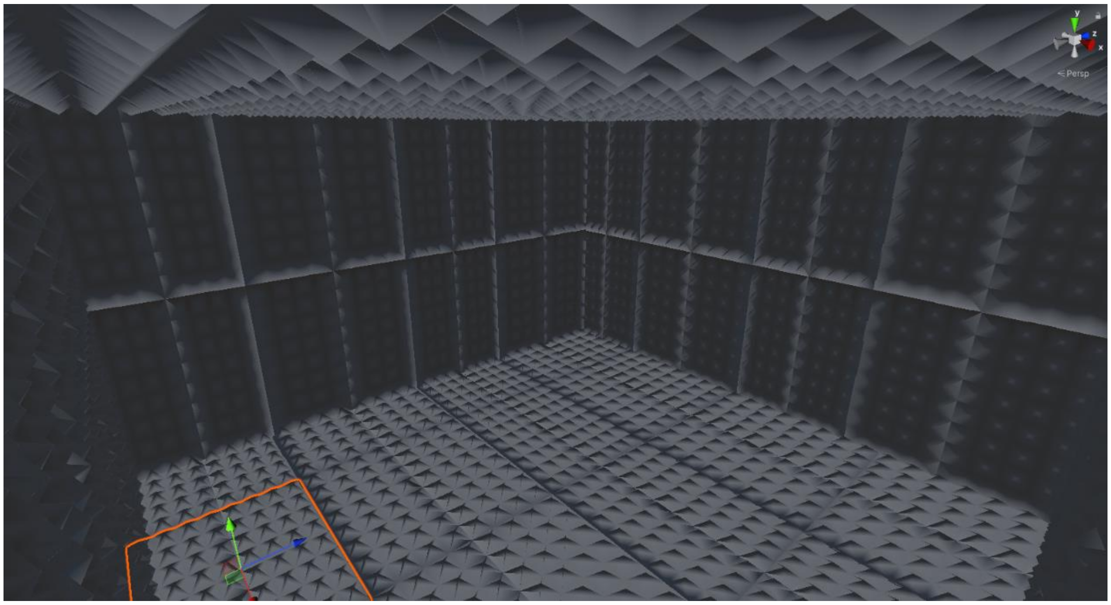

Figure 4.7 The Virtual Anechoic Chamber at Raspet Flight Labs

These locations were selected because they correspond to small, medium, large, simulated near infinite reverberant spaces and non-reverberant (acoustically infinitely large) spaces, respectively. Each space has corresponding binaural audio recorded from the perspective of a human head in all four locations in a high-fidelity virtual environment. These conditions will be presented using an HTC Vive Pro Two virtual reality headset and a pair of wired earbuds connected to the presentation PC via a 3.5mm audio cable.

4.4.4 Binaural Recordings

Each recording was taken in one of the four varying size rooms. Also, five angled locations around the human head were recorded, -90, -45, 0, 45, and 90. Four additional noise audio source locations were generated for each room at positions -60, -30, 30, and 60. Each recording was taken 3.5ft from the sound professions #11846, as shown in Figure 4.8. An assistant would hold the clicker to their chest at marked locations along the floor to maintain the proper distance and angle from the recording device.

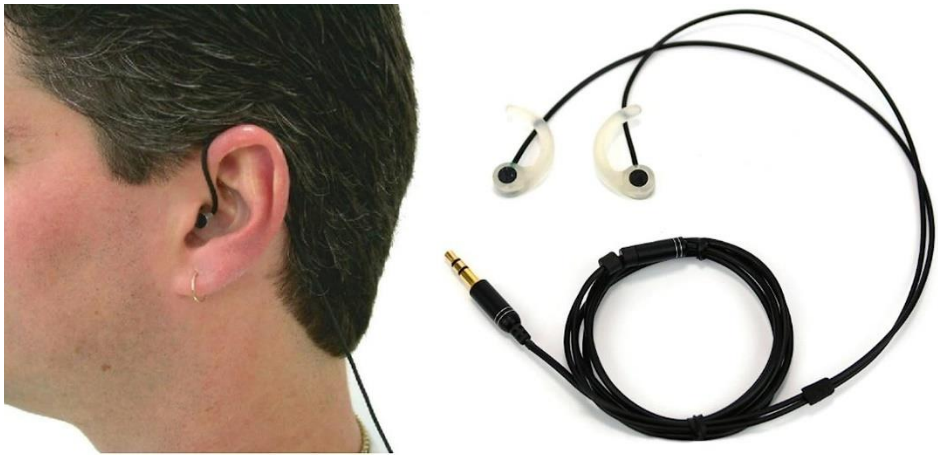

Figure 4.8 Master Series #11846, by Sound Professionals

All recordings were done with myself wearing in-ear microphones to maintain consistency across recordings. The in-ear microphones were connected to the Tascam DR-05X recorder, as shown in Figure 4.9.

Figure 4.9 Tascam DR-05X Stereo Handheld Digital Audio Recorder

To avoid the ceiling effect, noise locations were interpolated from the waveforms. However, this data was unreliable as it did not reflect the actual reverb of the room geometry. McDermott’s study showed that reverb decay, measured by the RT-60 time (the time for the sound level to drop by 60 dB), was a key cue for separating sound and space perceptually.

CHAPTER V

RESULTS AND DISCUSSION

This chapter describes the experimental results and discusses the experimental conditions and hypotheses. As well as the use of Repeated Measures Analysis of Variance on the data collected in experiments one and two.

The first hypothesis states that the unexplained phenomenon experienced in preliminary testing would persist in different environments in both real and virtual environments. The mismatching of the visual space with the auditory scene would cause perception issues with audio localization and provide dimensioned performance in localizing audio targets.

The second hypothesis states a positive correlation exists between the reverb disparity and the error rate in virtual reality. The more the reverb differs from the visual cues, the higher the error rate.

5.1 Front-Back Reversals

Of the participants in the trial, 29% experienced reversals. This finding is similar to that of Wenzel et al. [14], who had a reversal rate of 31%. When asked why, participants often indicated the absence of a visual queue capable of producing sound in front of them led them to believe the sound must have come from behind them. Likely the cause is the ILD and ITD being the same both forward and backward, and the lack of visual stimuli, the participants hear a reversal. Since this study was not a test of reversals and the ILD and ITD are the same, the back-facing angles were flipped along the axis of the participant's shoulders. The cue type may also have been a cause for reversal. An assistant experiment talked at each angle location during each recording session. Before inducing every impulse, the helping experimenter would announce each angle's location for the recording. To members of the HI5 Virtual Reality lab, this voice was much easier to locate than the click impulse used in testing, therefore, marked for future research. A probable cause is that the click rapidly covers a broad audio spectrum, whereas the people may be more in tune to pick up on others speaking.

5.2 Participant Judgments of Audio-Visual Manipulation in VR Environments

Experiment One examined the effect of environment size on audio-visual matching. We hypothesized that larger environments would make distinguishing between matched and mismatched conditions easier based on the differences in reverberation profiles. The profiles for the anechoic, small, medium, and large environments were 17ms, 36ms, 330ms, and 350ms, respectively. Figure 5.1 shows a "matched" response probability for each environment size and 95% confidence intervals. The figure shows the probability of a "matched" response and includes 95% confidence intervals. We can see that generally matched and mismatched environments significantly differed (F(2, 38) = 13.501, p<0.001). Our results supported our hypothesis: no significant difference existed between matched and mismatched responses for the small environment. The difference increased significantly as the size disparity between the matched and mismatched environments increased. However, we only tested this effect using impulses recorded in the anechoic chamber, limiting our findings' generalizability. These results may be due to differences in the reverberation profile since that is the part of the sound that contains information about the room's size.

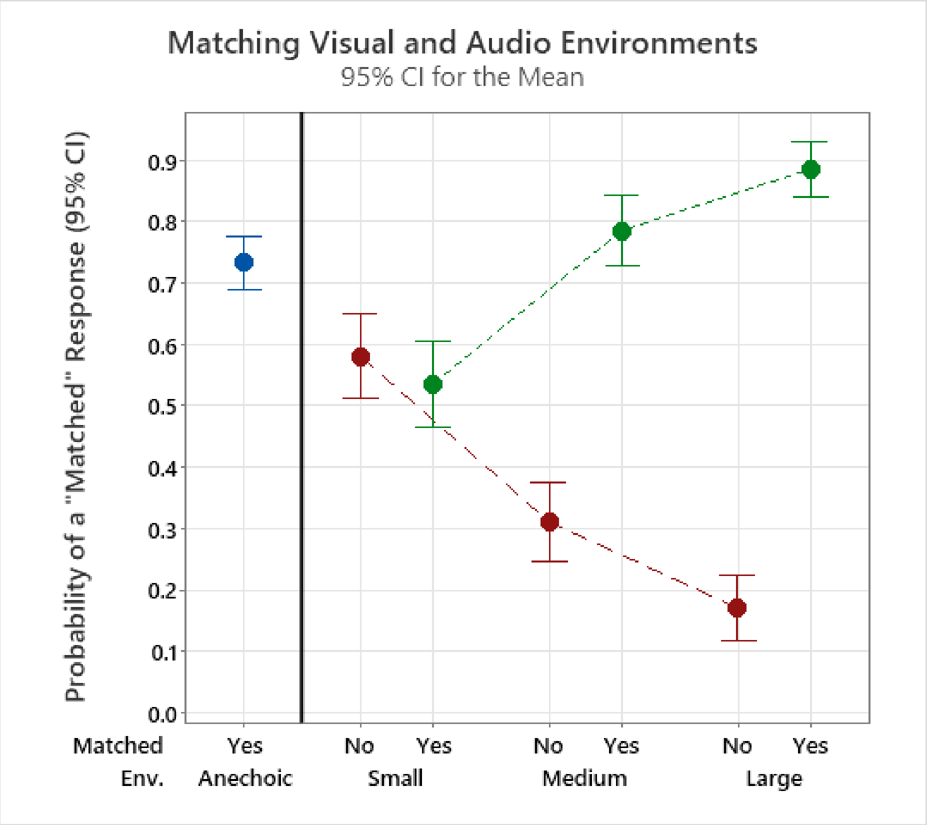

Figure 5.1 Graphical Representation of the Probability of a "Matched" Response for Each Visual Environment When Matched or Mismatched with the Acoustic Environment. Error Bars Represent 95% Confidence Intervals.

The benefit of matched audio environments increased as a function of environmental size. The smallest environment, the inside of a vehicle, had a very short reverberation profile, much like that of the anechoic chamber.

The matched audio environments were those that had the same reverberation profile as the anechoic chamber, while the mismatched audio environments were those that had different reverberation profiles. The unsigned error was the absolute difference between the actual and perceived sound source locations. The results indicated that the unsigned error was lower in the matched audio environments than in the mismatched audio environments, indicating that the participants were more accurate in locating the sound sources when the visual and auditory environments were matched. This effect was more pronounced in larger environments than in smaller ones, as seen in Figure 5.2. Unsigned error significantly differed between matched and mismatched visual environments (F (3, 57) = 5.321, p=0.003).

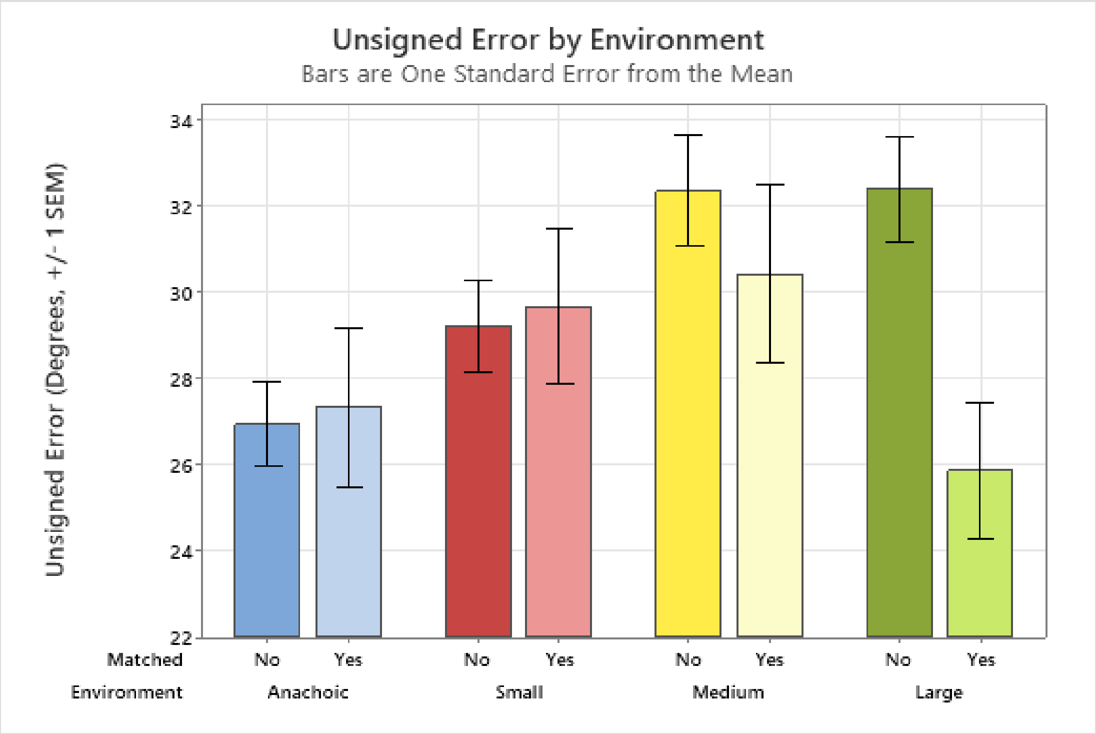

Figure 5.2 Graphical Representation of the Unsigned Error in Degrees from the Various Virtual Environment Room Sizes

5.3 Intended and Perceived Locations of Audio Cues

We conducted Experiment Two to test the effect of audio-visual congruence on participant accuracy. We expected participants to perform better when the visual and auditory environments matched. Our results confirmed this hypothesis: matched environments led to significantly smaller unsigned errors than mismatched environments (F(1, 19) = 6.882, p=0.017). Figure 5.3 shows the distribution of perceived sound source locations across the 20 participants and the 5,760 data points. However, we had to discard 44% of these data points as they were noise trials from the surrounding click locations, and their primary function served to avoid the ceiling effect. They are not illustrated in Figure 5.3.

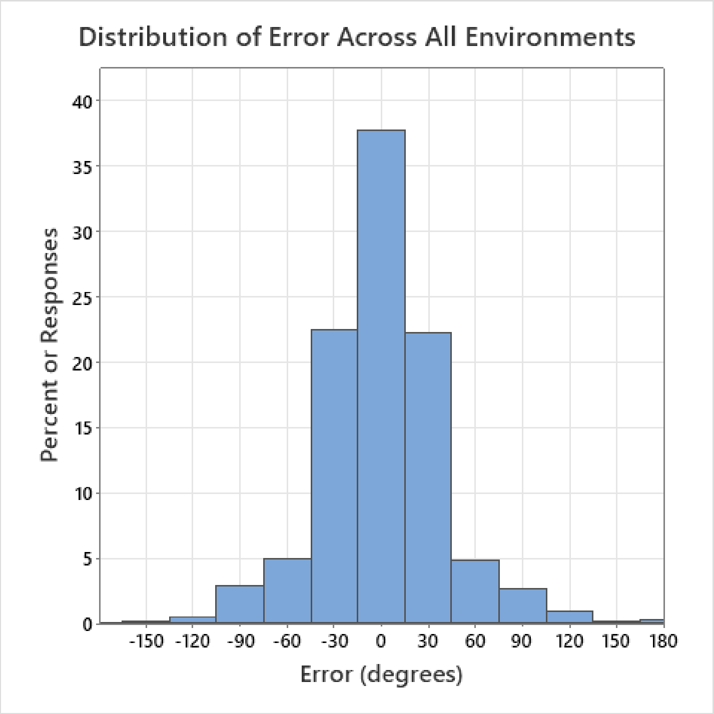

Figure 5.3 Distribution of Error Across All Environments

The mean difference between mismatched environments was 1.9 more error than that of the matched environment, as shown in Figure 5.4. This indicates a high likelihood that room visuals play a part in the perception of spatial audio.

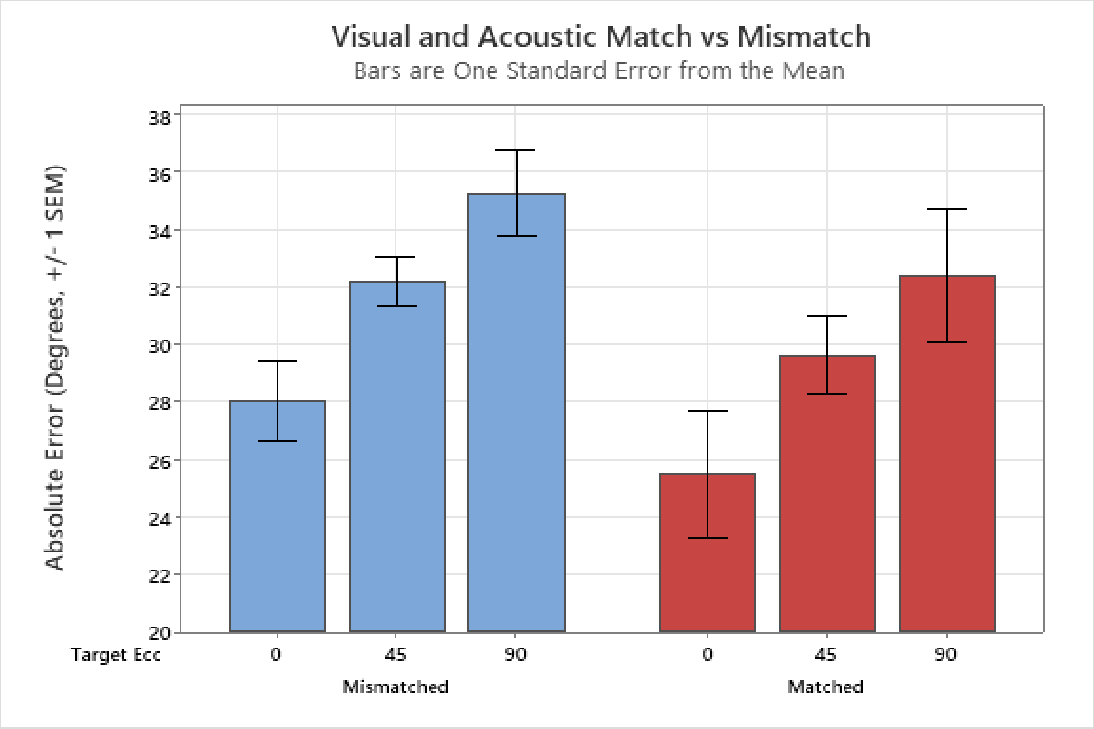

Figure 5.4 Absolute Error of Matched and Unmatched Environments

The intended locations vs. the mean average location of the spatial cues across all environments can be seen in Figure 5.5. Participants were accurate within ~2 across the average except for +/- 90°, where the participants had a mean average of ~62°. Targets at 90° had the highest mean error average, likely due to biomechanics.

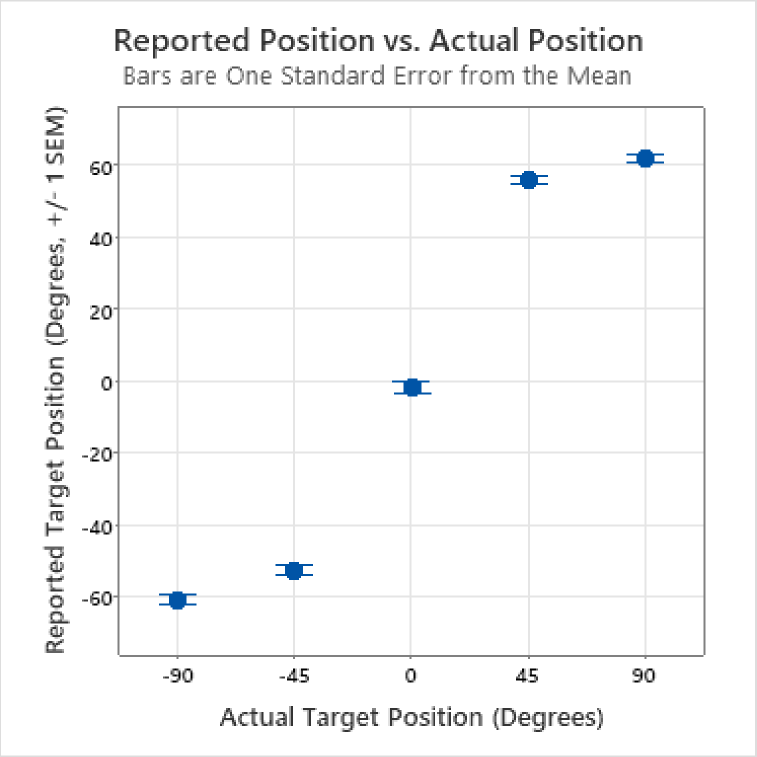

Figure 5.5 Visual and Acoustic Match vs. Mismatch, Experiment One

CHAPTER VI

CONCLUSION

In this paper, we presented two experiments that investigated how spatial audio perception in virtual environments is affected by the congruence or incongruence of auditory and visual cues. We manipulated the reverb of the auditory environment to match or mismatch the visual environment. We measured the sound localization accuracy in both conditions and used repeated analysis of variance to measure the results.

The first hypothesis of this research is that the mismatching of the visual space with the auditory scene would cause perception issues with audio localization and provide dimensioned performance in localizing audio targets. We found that participants made more errors when the reverb did not match the visual cues, indicating that audio-visual congruence is important for accurate spatial audio perception. These results suggest that designers of virtual environments should aim to match the auditory and visual cues as closely as possible to create more immersive and realistic experiences that rely on spatial audio.

The second hypothesis predicts that the reverb disparity, which is the difference between the reverb of the auditory environment and the reverb of the visual environment, positively affects the error rate in virtual reality. The error rate measures how far the perceived sound source location deviates from the actual one. According to this hypothesis, the larger the reverb disparity, the more errors participants will make in locating the sound sources. This implies that the auditory environment's reverb should match the visual environment's reverb to minimize the error rate and improve spatial audio perception in virtual reality. We found that matched environments led to significantly fewer unsigned errors than mismatched environments. Reverb disparity created an increase in error rate, which suggested that matched conditions led to more accurate spatial audio perception than mismatched conditions.

We found that vision played an important role in modulating these factors and shaping the experience of the virtual environment. This research matters because it advances the understanding of how spatial audio perception in virtual environments is influenced by the match or mismatch between auditory and visual cues. Spatial audio is a crucial component of immersive and realistic virtual environments, as it provides information about the location, distance, and movement of sound sources in the virtual space. However, spatial audio perception can be affected by various factors, such as the reverb of the auditory environment, the availability and consistency of visual cues, and the users' individual differences. By manipulating these factors and measuring their effects on sound localization accuracy, we can gain insights into how multisensory integration works in virtual environments and how to optimize it to enhance user experience. This research also has implications for designing and evaluating virtual environments that rely on spatial audio, such as virtual reality games, simulations, training, education, entertainment, and social interaction. By understanding how audio-visual congruence affects spatial audio perception, we can provide guidelines and recommendations for creating more effective and engaging virtual environments that leverage spatial audio.

6.1 Limitations and Future Work

The current experiments had limitations in both operation and information. Firstly, the study only had 20 participants, all of which are young adults with an average age of 26. Therefore, the generalizability of the results across populations, such as older adults, is limited. Future studies could incorporate a broader range of participants. Older participants or participants with hearing loss may rely more heavily on the environment for cues.

Interestingly, while preparing the stimuli, the researchers noticed that the human voice was much easier to locate than the clicker sound, even when recorded in the same room. This suggests that the human voice may have some unique features or cues that facilitate sound localization, such as pitch, timbre, or prosody, that humans are accustomed to picking out. A possible follow-up study could investigate this phenomenon by comparing the performance and experience of sound localization with human voice versus clicker impulses in different VR rooms. This could provide more insights into the role of sound source characteristics and environmental factors in sound localization.

Furthermore, the VR environment used in this study may have created a mismatch between the visual and auditory cues for sound localization. The participants were presented with a VR representation of an anechoic chamber, a small and finite room with walls covered with sound-absorbing material. However, acoustically, an anechoic chamber is supposed to simulate an infinite acoustic environment where no sound reflections occur. Therefore, the participants may have experienced a conflict between what they saw and heard, which could affect their performance and perception of sound localization. A better way to create a VR anechoic chamber would be to use a black void as the visual scene, matching the acoustic scene more closely and avoiding any visual distractions or biases.

As seen in Figure 5.1, as the reverberation time of the waveform tail increased, so did the accuracy of participants' ability to distinguish the room size. This experiment was not originally designed to measure this phenomenon. A better design for the experiment would be to test mismatches of varying tail lengths and reverb lengths of all the different environments. This was avoided due to the length of testing needed. This study was exploratory; therefore, some aspects of the experiment were simplified for the sake of brevity. We only compared the VR rooms with the anechoic chamber audio. In the pilot testing, we played two audio samples of each click position of every room, resulting in a very long experiment of nearly double the length. We reduced this to just a room and the chamber in each room. A random room audio sample and the anechoic chamber audio sample were each played twice per angle. We omitted the longer style of testing to avoid participant fatigue and discomfort. This shortened the experiment time significantly, limiting the amount and variety of data we could collect.

REFERENCES

[1] J. Blauert, “Spatial hearing,” The psychophysics of human sound localization, 1996. doi:10.7551/mitpress/6391.001.0001

[2] A. W. Bronkhorst, “Localization of real and virtual sound sources,” The Journal of the Acoustical Society of America, vol. 98, no. 5, pp. 2542–2553, 1995.

[3] M. Gerzon, “Dummy head recording,” Studio Sound, vol. 17, pp. 42-44, 1975.

[4] E. M. Laitinen and V. Pulkki, “Binaural reproduction for Directional Audio Coding,” 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2009. https://doi.org/10.1109/aspaa.2009.5346545

[5] J. Lee, S. H. Shin, and C. L. Lstook, “Analysis of Human Head Shapes in the United States,” International Journal of Human Ecology, vol. 7, no. 1, Jun. 2006.

[6] J. M. Loomis, Y. Lippa, R. L. Klatzky, and R. G. Golledge, “Spatial updating of locations specified by 3-D sound and spatial language,” Journal of Experimental Psychology: Learning, Memory, and Cognition, vol. 28, no. 2, pp. 335–345, 2002.

[7] E. Luckett, T. Key, N. Newsome, and J. A. Jones, “Metrics for the evaluation of tracking systems for Virtual Environments,” 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), 2019. https://doi.org/10.1109/vr.2019.8798374

[8] R.Y Litovsky, H.S Colburn,W.A Yost, and S.J Guzman, "The precedence effect," The Journal of the Acoustical Society of America, vol .106, no .4, pp .1633–1654,1999.

[9] William. Mills, A.Auditory Localization, which forms chapter 8 (pages303-348)of Foundations of Modern Auditory Theory (ed .J.V.Tobies)Volume II, Academic Press, New York,1972. Mills’article is the best survey that I know of concerning how stereo hearing works.

[10] J.H.McDermott, A. Schwartz, and B.Shinn-Cunningham, "Spatial cues alone produce inaccurate sound segregation: The effect of Interaural Time Differences, "The Journal of the Acoustical Society of America, vol .132, no .1, pp .357–368,2012.

[11] J.Traer and J.H.McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space, "Proceedings of the National Academy of Sciences, vol .113, no .48,2016.

[12] S.S.Stevens and E.B.Newman, "The localization of actual sources of sound, "The American Journal of Psychology, vol .48, no .2,p .297,1936.doi:10 .2307/1415748

[13] J.Willert, Volker, et al." A probabilistic model for binaural sound localization."IEEE Transactions on Systems, Man, and Cybernetics, Part B(Cybernetics)36 .5(2006):982-99

[14] E.M. Wenzelel, M.Arruda, D.J.Kistler, and F.L.Wightman, "Localization using non-individualized head-related transfer functions, "The Journal of the Acoustical Society of America, vol .94, no .1, pp .111–123,1993 .https://doi.org/10 .1121/1 .407089