Human Perception of Aural and Visual Disparity in Virtual Environments

Erik Wiker

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Mechanical Engineering

Michael J Roan, Chair
Alfred L Wicks
Pablo Alberto Tarazaga

August 9, 2018
Blacksburg, Virginia

Keywords: Virtual Reality, Spatial Audio, Localization, Ambisonics

Human Perception of Aural and Visual Disparity in Virtual Environments

Erik Wiker

ABSTRACT

With the development of technology over the past decade and the former challenges of virtual environments mitigated, the need for further study of human interaction with these environments has grown apparent. The visual and interaction components for virtual reality applications have been comprehensively studied, but a lack of spatial audio fidelity leaves a need for understanding how well humans can localize aural cues and discern audio-visual disparity in these virtual environments. In order for development of accurate and efficient levels of audio fidelity, a human study was conducted with 18 participants to see how far a bimodal audio and visual cue need to separate for someone to notice. As suspected, having a visual component paired with an auditory one led to biasing toward the visual component. The average participant noticed a disparity when the audio component was 33.7° apart from the visual one, pertaining to the azimuth. There was no significant evidence to suggest that speed or direction of audio component disparity led to better localization performance by participant. Presence and prior experience did not have an effect on localization performance; however, a larger participant base may be needed to draw further conclusions. Increase in localization ability was observed within a few practice rounds for participants. Overall, performance in virtual reality was parallel to augmented reality performance when a visual source biased sound localization, and can this be a both tool and design constraint for virtual environment developers.

Human Perception of Aural and Visual Disparity in Virtual Environments

Erik Wiker

GENERAL AUDIENCE ABSTRACT

Virtual Reality has overcame a large technological gap over the past decade, allowing itself to be a strong tool in many applications from training to entertainment. The need for studying audio fidelity in virtual environments has emerged from a gap of research in the virtual reality domain. Therefore, a human participant study was conducted to see how well they could localize sound in a virtual environment. This involved signaling when they noticed a visual object and a sound split. The average of 72 trials with 18 participants was 33.7° of separation on the horizontal plane. This can be both a tool and a design constraint for virtual reality developers, who can use this visual biasing as a guideline for future applications.

Acknowledgments

I’d like to thank my family and friends for their infinite supply of support. The many who I interacted with during my years in undergraduate and graduate career were those who breathed life into my days. I’d like to thank my professor and advisor Dr. Roan, who let me lead this research with my creativity and his guidance. His support and willingness to turn a stressful situation into a silly one was invaluable. Thanks to my lab mates, Abram and Hyeon for always allowing me to bounce ideas off them. Thanks to everyone who didn’t give up on me, and all the participants who were willing to spend their time to perform these studies. Thank you to my committee members and the staff of the Virginia Tech Mechanical Engineering department for guiding me through this maze called graduate school. And lastly, thank you to Melissa Fairfax, who never stopped believing in me and was there every step of the way.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Introduction

1.1 Introduction

Human computer interaction has been a paramount field of study to better understand and integrate computers into human lives. Smartphones, tablets, and smartwatches have been more common with human users as each day passes. Along with this technology, augmented reality (AR) and virtual reality (VR) have been on the rise. Even though the first virtual reality head-worn display (HWD) was developed in the 1960s, the Oculus Rift was the first major consumer grade HMD to gain popularity with the mainstream. Now a plethora of VR and AR technologies are available, such as the HTC Vive and eventually the Microsoft HoloLens. With the rise of these technologies and the possible benefits to society stemming from them, the need to understand how humans can use them effectively is increasing.

VR applications go far beyond what the typical consumer knows or uses them for. The typical use by current users is for entertainment, a medium for experiences you can't normally do safely or easily. Things like games and movies are a common experience, but VR is also sometimes referred to as the empathy machine. It is a way to experience something firsthand that would otherwise be impossible or unlikely, such as experiencing what it would be like to have schizophrenia. Developers have used this strategy to try and change opinions on things like social issues.

Other applications that have yielded great results are VR therapy for treating phobia (Culbertson et al., 2010), anxiety (Anderson et al, 2013), or PTSD (Hoffman, 2004). VR is also a means of training and simulation, recreating experiences that allow muscles to be trained by stimuli in fields from the military to medical procedures. Others plan to use VR for things like education. The analysis of scientific data is just another use case that could be optimal for VR, like moving around 3D models to have better spatial reasoning. From art to science, VR has so many possibilities that have yet to be fully explored. For these reasons, it is worth trying to understand as much as we can about human interaction with Virtual Environments (VEs).

1.2 Motivation

With the development of VR technology and virtual environment design, the parameters determining the quality of these environments appeared. The work that has gone into the study of VR and understanding the psychology behind it is very comprehensive. However, there appears to be a gap when pertaining to audio fidelity in the VR world. So much research is focused on interaction and display of the environment and how the VEs make the user feel and respond. Immersion and presence qualities both are specific aspects designed for in VR. How real something seems, how you interact, and how it makes you feel encompasses these two qualities.

Audio fidelity, and how it affects interaction, understanding, and feeling of the user is much more ambiguous. There is plenty of work understanding how humans perceive and localize sound, but not how visuals in VEs can affect it. The bimodal perception of a human is complicated, and there is evidence, discussed later, that these effects are different in VR environments.

1.3 Thesis Objectives

The main goal of this thesis is to define how well humans can localize sound in a virtual environment and to what degree visual components affect this skill. It will branch into three main bodies of work: psychoacoustics, spatial audio, and virtual reality. In this, it will point out the gap of knowledge with audio fidelity in current literature and attempt to distinguish the ability of human identification of separation between an audio signal and a visual object. This will be noted by a number of degrees.

This requires a body of participants to study, a series of tests to see whether their ability to hear and localize sound is normal, and how far apart these audio and visual components need to be in order to discern that they are no longer together or incongruent. Secondary objects are to find if there are any trends within this localization ability, such as outside influences providing advantages or disadvantages, or how the disparity displays itself to the participant. This can be in the form of speed or direction of the sound drifting away from the visual component or how tired or experienced a participant is.

Finally, the tertiary objective to see if self-reported presence is a factor when localizing sound in a virtual environment. This will be done by examining the post-questionnaires given after the experiment.

1.4 Thesis Organization

This thesis will be organized as follows. After this brief introduction and motivation, there will be a background section. This will attempt to bridge the three bodies of work previously mentioned and argue why this research is important, due to the gap of knowledge in the literature.

Afterwards, the methodology section will detail how the study was designed and conducted. This will include the scripting and use of different software to connect the audio and VR experience, in addition to detailing how the trials were conducted. Also, this will explain the hearing test and the questionnaires given to the participants prior to and after the study.

Following the methodology, the results and discussion section will analyze all the data that was collected during the experiment and explain the findings. The data will be shown in figures for visualization, and then trends and possible significance of the data will be included in the discourse.

Finally, the conclusion will summarize the discussion of the results, discuss possible improvements that could be made to the experiment, and propose future work. After that, the reference and appendices will follow.

Background

2.1 Literature Review Overview

This chapter’s primary goal is to bring readers up to speed with understanding the main three bodies of work that this thesis will explore and to bridge the gap between those bodies of work. Due to the nature of the questions analyzed in this thesis, namely, “How well can humans perceive a disparity between auditory and visual stimulus?” requires the understanding of three branches of science.

The first of these branches is the psychological aspect. This is necessary to understand how well humans can perceive sounds around them, and to do this, we must dive into how the human body localizes sounds, to what degree they can do this accurately, and under what conditions, and also how the influence of different visuals will affect this.

Within the realm of psychology, the human body’s perception of sound is all detailed in the subject called psychoacoustics: studying the effects of sound on humans and how it is heard. Sound localization will be broken down to give reason for the study’s methodology, in the chapter to follow, and a hypothesis for how well we will perceive sound in virtual environments (VEs). There will also be another topic, about how visuals can affect perception of sound. This hints into complexities of bimodal perception and how we could possibly use this when designing virtual environment applications in the future.

The second body of literature that will be explained is spatial audio. This is another complex topic, in relation to which we will discuss the different ways that we currently understand to recreate sound fields with high fidelity. I will discuss the tradeoff of several spatial audio algorithms and why I chose Ambisonics to reproduce the audio during the experiment. This section will also include a breakdown of how ambisonics works, step by step, and the various limitations and advantages it has.

The third section covered is virtual reality. Although it seems like a somewhat new topic, VR has actually been around for over half a century, depending on definition, with the first HWD invented by Ivan Sutherland in the 1968. Virtual reality has been studied thoroughly and is still a large research topic today.

This section will contain the many uses of VR and how it can be applied effectively. Also, we will discuss the current limits of VR and where research is lacking, wherein lies the aim of this research. The virtual reality discourse will round out with other pertinent aspects of VR, including terms such as presence and immersion as defined by our standards and many other research efforts. Presence and immersion have complex and not-very-well-understood relationships that require more data to see how we can alter these variables in virtual reality applications.

2.2 Psychology: Human Perception and Bimodality

Human perception is a complicated subject involving bodily functions of different organs and the brain. Each sense uses different systems to both absorb the stimulus and then interpret it. This is exemplified by vision, where the eyes have components that take information from the photons of light that they come in contact with, and that information is sent to the brain to facilitate an understanding of our environment. These systems compensate for color, distance, and a wide variety of other information so that we can accurately interact with the world around us.

This is similar with the sense of hearing. The ears are the physical component that have hair cells in which, when sound vibrates the cells, information is sent to the brain to get processed. When the sound is in a particular environment, then eventually in a particular ear in a particular body, it creates a unique signal to the brain. Factors of frequency (how quickly the medium is vibrating) and loudness (the perception of how large the amplitudes of those vibrations are) are both extremely important and not the only variables in sound. The information contained is complex, and reflections off certain artifacts and natural filtering of sound can also give information about one’s environment.

These two senses are very dominant, so much so that if you lose one it can severely impair your understanding of your environment and makes interaction much more difficult. They also affect each other. A study in 2001 discussed that multisensory cells, which are the cells that respond when there are multiple senses at once, increase their firing rate to a point where the sum is much greater than if you had just summed the firing rate of each sensory cell separately (Calvert, Hansen, Iversen, & Brammer, 2001). So, when you have both visual and auditory stimuli, they are more powerful together. This gives larger value to things that are multisensory than at first glance. Furthermore, when the senses counteract each other, it has been observed that the response to these senses significantly drop (Calver et al, 2001).

This is further backed by practical examinations in a study which tasked humans to do a visual search in separate conditions (Bolia, D'Angelo, & McKinley, 2001). The study found that using no audio to find a lit LED resulted in a significantly longer search time as opposed to using spatial audio conditions (Bolia et al, 2001). This shows that adding the component of audio, or some form of listening cue, can increase performance in certain tasks. The study only took into account two different spatial audio forms and zero audio, which polarizes results as there is no gradient or in between. There is no information gained on what level of fidelity of spatial audio, or at what point, is needed for such a large performance improvement. Also, the study's task is centered on locating something, which any sense would aid with to a degree. Particularly hearing and seeing are used to locate things frequently in day-to-day lives. This study is reflective of a 1995 study by Hendrix and Barfield, which studies similar binary components (either completely there or not) in VR that will be touched on in section 2.5.

The ventriloquist effect is the phenomenon that takes place when both visual and auditory components happen congruently, and the brain biases toward the visual stimulus. This is commonly taken advantage of by ventriloquists, where the name originates, when they talk and move their puppets' mouths. This makes it seem as if the puppet were talking, since they do their best to keep their own mouths from moving. This shows an interesting relationship between the two modalities. This effect was also studied in AR applications, which will be spoken to in the section 2.5.

A study investigated this topic’s degree of influence in humans by conducting an experiment where people were exposed to each of the modalities and were tested on how well they could localize them (Charbonneau, Véronneau, Boudrias-Fournier, Lepore, & Collignon, 2013). This study featured an array of speakers and white LEDs that were spread out in hemispherical fashion around the participant at 15° intervals through a 180° frontal space. These were all 75 centimeters from the participant’s head.

In the study, an LED would illuminate with a burst of white noise from one of the speaker pairs, called a probe. Shortly after, a target of either a single modality or both modalities would be sent out to the pair 15° to the left or right of the probe. Sometimes, if it was both stimuli, the pair would be separated with the audio sent from the speaker on the side opposite from where the LED was lit. This case could show biasing toward visual effects.

An important part of the experiment was that the participant had their face forward the entire time, which allowed for analysis of periphery effects as well. Charbonneau et al. found that localization severely decreased in periphery and that biases toward visual cues increased as the targets were more in the periphery. However, when the visual was blurred, the auditory response could dominate. This study provided some useful ideas for designing the experiment for this thesis, and it is also one of the reasons I decided to stay away from the periphery when testing my participants. This would avoid too much complexity, as the experiment needs to tackle the frontal space before it can focus on periphery.

The understanding between the bimodal domain of hearing and seeing is being heavily studied within the brain, but it is out of scope for this research. The gap of practical knowledge starts here between visual and aural stimuli, which helps give purpose to the research done within this thesis.

2.3 Psychology: Psychoacoustics

The majority of understanding psychoacoustics is attributed to Zwicker, who wrote a book encompassing the topic in 2013. The main focus of this section is to detail how the human brain localizes sound, so that I could have a working hypothesis of how well humans can do it in general circumstances, such as when they are not exposed to a virtual environment. Also, this gives insight into how to design the experiment – what kind of audio signals would work best, and to what kind of conditions I should subject participants.

Localization of sound requires the brain to exchange information with both ears, and between them. The information that is exchanged is mainly intensity, latency, and frequency information of the sound itself (Zwicker, 2013). These things are all detected by the ear and compared ear-to-ear to derive from where a sound could come.

The delay between sounds reaching each ear is known as the Interaural Time Difference (ITD), measured in seconds. This is what the human body usually uses to locate a sound if a sound is low frequency, below 1000 Hz. If a sound source is directly to the right or left of someone, or 90° from the front, the maximal ITD occurs. This ranges depending on the person, but it is usually around 0.6 ms (Zwicker, 2013). The smallest notable difference noted by Zwicker is about 5°, which would in turn produce an ITD of approximately 50 µs. This value can range from 30 to 200 µs (Zwicker, 2013). However, this value is disputed since Wang & Brown (2006) claimed it to be 1° with a pure tone sine wave directly in front of the listener. It was not noted if this included a different psychoacoustic effect being used or the frequency of that sine wave. For the series of this paper, 5° will be assumed.

Specific sound frequencies, such as certain narrow band sounds or singular frequency sine waves, can produce really interesting effects, which I wanted to avoid during experimentation. Certain bands or frequencies can cause symmetrical effects, where a sound might be coming from 10° to your left but instead comes from 170° to your left, behind you. This mirroring effect and other confusing localizations occur at these specified frequency bands. This led me to pick a wide range frequency (Zwicker, 2013).

Interaural Level Difference (ILD) occurs when a sound is perceived louder by a listener in one ear than another. This is used at higher frequencies, around 1500 Hz or higher. The difference in the sound level happens mainly by acoustic shadowing, when something is in the way of a sound causing it to have to travel around the room and back to your ear. There is no direct line from most sound sources to both ears, usually being blocked by the listener’s head or body. The body also causes reflections that may not occur equally on either side, which can lead to localization of a sound (Zwicker, 2013).

With this understanding, I could move forward with how to conduct a valid experiment and have a physical understanding of how humans localize sound.

2.4 Spatial Audio

Accurately creating sound, or recreating previously recorded sound, requires a large amount of hardware and processing power. When sounds are produced, they not only carry inherent properties such as frequency and amplitude, but also things like radiation pattern and environmental properties. Whether you listen to a song in a concert hall or anechoic chamber, the number and strength of the reflections you hear rely on the geometry and material of the space around you. These factors are unique to each signal, what the sound is produced by, and the environment that it is being produced in, not to mention where the listener is positioned and their specific head-related transfer function (HRTF).

HRTFs are a sort of complex spectral filtering that is completely exclusive to each individual listener, which is caused by body properties like shape of ear canal, body, and head. This basically affects how the listener perceives the sound and makes it so each individual will hear something slightly different, even if all other variables are constant. HRTFs are important in creating or reproducing sound fields that sound accurate to each listener, and they will be referenced throughout this section.

There is importance to creating sound fields accurately, such as spatial understanding. When one enters a room, they can understand certain aspects like how large the room is just by listening in that room. This is due to the ILDs and ITDs reviewed in the previous section. For example, if there is an echo, how long that echo takes could give you subconscious feedback of how far the walls are away from you. The visual search done by Bolia is proof that we can use spatial audio cues and localization to better our performance of tasks as well. If we wanted to create a feeling of being in a specific environment, auditory sense would be an important feedback from that environment to allow it to be convincing and aid in our understanding of that environment. This lends itself to VR application, which will be talked about in future sections.

In order to reproduce sound fields, some sort of production algorithm must be used. The most basic reproduction of any signal uses one speaker, called monophonic. This speaker can produce spherical waves that eventually can be represented at plane waves at a certain distance. Since all sound is coming from one source, it gives little to no spatial information about the sound produced. One could also easily localize the speaker since the sound is all coming from one spot, depending on environmental factors.

Using two of those speakers, however, gains more resolution and adds the ability to pan the sound between them. The signal could come equally from both or be weighted more to one side. This is called stereophonic, and allows for two channels, or two signals, to be reproduced individually.

Pantophonic, or sometimes known as 5.1 or traditional surround sound, uses 5 loudspeakers and 1 subwoofer placed around the listeners. This adds more resolution and spatial information to the listeners by allowing change of azimuth of the source, or the angle around the listener to the left or right. However, this doesn’t have any ability of reproduction in elevation, up or down, from the listener.

Periphonic is the next natural step, which places speakers all around the listener both along azimuth and elevation. Humans tend to localize worse with elevations as opposed to azimuth, which is further decreased with decrease in intensity of the sound (Su, & Recanzone, 2001). Even with this, the periphonic reproduction allows for much more complex sound field recreation and could aid in spatial understanding and audio fidelity due to an increase in loudspeakers.

The two spatial audio algorithms I will discuss are Vector-based Amplitude Panning (VBAP) and Ambisonics. These are both considered periphonic, with no specific required number of speakers. Both will be compared, and then finally I will discuss which I chose, why, and how it works. This comparison discussion mainly uses Pulkki's paper from 1997 discussing VBAP.

When using periphonics, you can recreate complex sound fields. This can use speaker arrays, which are typically a large number of speakers situated around the listeners, or headphones. Recreating in headphones sounds ideal, particularly when creating spatial audio for HWDs. Since the user of VR will most likely have some head-worn accessory, the headphones can be included in the device. This would allow for the user to not have to worry about the environment they are experiencing VR in, and it can compensate for loudspeaker array and setting up a controlled environment. This is the common way VR companies deal with their audio; Oculus and HTC use these methods on their HWDs.

The downside to using this technology is that it requires the use of HRTFs. To get an accurate HRTF for a user isn't a simple process, and it would have to be done for each individual who wanted to use it. It would involve complicated analysis of each individual's body to create one, and each person's HRTF is unique. It also is very computationally expensive and wouldn't work easily with a system that didn't use HWDs, like Cave systems which use projectors for displays (Pulkki, 1997). Although, most systems now use an estimated HRTF for most users, so some amount of accuracy is retained.

Instead of using headphone-based spatial audio, speaker arrays can also be used. Two techniques used to recreate sound fields with speaker arrays are Ambisonics and VBAP. They both have been studied at length, with Ambisonics being developed in the 1970s and

VBAP by Pulkki in 1990s. These seem to be two suitable solutions for VR applications, which have similar requirements, levels of fidelity and accuracy, and can position virtual sound sources independently from a number of speakers (Pulkki, 1997).

VBAP uses panning between 3 independent loudspeakers to create virtual sources. It allows for an unlimited number of speakers, which can be located anywhere in the room. VBAP uses vectors in combination with standard amplitude panning to create the virtual sources with high computational efficiency. VBAP is considered an object-based method, which means that the audio artifacts or objects are each individually characterized by a single audio track. This is coupled with secondary data, or metadata, to retain things like locational properties of that audio track. VBAP requires low reverberation qualities of the room it is being used in, and the speakers must all be the same distance to the listener.

Ambisonics, on the other hand, decodes a three-dimensional sound field into four separate channels, labeled W, X, Y, and Z. These channels all hold different information, and that information is decoded later. Ambisonics works best when loudspeakers are placed on Cartesian axes (Pulkki, 1997). Even though it can be considered not as flexible and efficient as VBAP, ambisonics doesn’t have to be object-based. This means it isn’t layout-dependent (Artega, 2017). The order of ambisonics used can also be increased to compete with VBAP in terms of spatial resolution and the size of the sweet spot – the area in the room in which the spatial audio is accurately recreated, which is usually in the center (Artega, 2017). Ambisonics is also regularly used in VR technology now; ambisonic tools for recreating audio fields for developers are included in game engines like Unity and audio signal processing software like MaxMSP (Artega, 2017).

With these things in mind, ambisonics was chosen for this study for spatial audio. Due to the restrictions of the equipment and room in the lab where the study was performed, ambisonics fit well with all of the resources available. These reasons are as follows. The sweet spot is not required to be large for the study. Directionality can be poor with ambisonics but improved at higher orders, in comparison to VBAP. Thoroughly studied and reproducible, ambisonics is accepted theory universally. The algorithm can be reproduced regardless of layout of environment. Ambisonics has a fixed number of channels but can be object-based if necessary. Panning and transitions are smooth in comparison to VBAP. It requires many fewer speakers than Wave-Field Synthesis. Finally, it is already used in current VR applications.

As discussed, ambisonics uses four separate channels to hold the sound field information. 0^{th} order ambisonics contains pressure field information, 1^{st} order contains acoustic velocity at origin, and 2^{nd} order and higher add higher order derivatives. As order increases, so does the sweet spot and spatial resolution, but computational speed decreases. 3^{rd} order ambisonics is regarded by most audio professionals as good enough for accurate recreation (Artega, 2017).

Ambisonics can either be used with synthetically created 3D sound productions or a recording of a 3D sound field. To record this, one would need a sound field microphone that uses four (tetrahedral microphone) or more capsules to capture the sound and the directional data.

The four channels that ambisonics encodes the sound signals into represent the pressure and velocity at origin. The following equations show how each channel is divided.

$$W(t) = \frac{s(t)}{\sqrt{2}} \tag{2.1}$$

$$X(t) = s(t)\cos(\phi)\cos(\delta) \tag{2.2}$$

$$Y(t) = s(t)\sin(\phi)\cos(\delta) \tag{2.3}$$

$$Z(t) = s(t)\sin(\delta) \tag{2.4}$$

Some sound field microphones convert directly into the four channels above, which are called B-format. Equation 2.1 is proportional to the pressure field at the origin. Equation 2.2, 2.3, and 2.4 are proportional to acoustic velocity on their respective axis. A-format is the name for a signal not yet encoded. Note that these equations are for 1st order ambisonics only. A linear system of equations is used to solve for these coefficients (Artega, 2017).

After the signal in encoded, it must be decoded and routed to the speaker array situated around the room. Using the coordinates of the speakers in the room (location $\hat{u}_i(\phi_i, \delta_i)$), with n number of speakers and a signal s_i, you can find the speakers signal as:

$$s_i(t) = w_i W(t) + x_i X(t) + y_i Y(t) + z_i Z(t) \tag{2.5}$$

These signals are then routed to each paired speaker and sent out. The method that was used to do this will be covered in the methodology chapter.

With that, the spatial audio portion is complete. Next will be the discussion of VR technology, its limitations and what will be important to the study.

2.5 Virtual Reality

Virtual reality technology has made great advances in the last few decades, with recent booms in consumer availability and plausibility in the last several years. With the increase in hardware technology, VR has been capable of being used by all sorts of individuals for a wide variety of applications, from entertainment to therapy. This progress of hardware has given way to vast amounts of research opportunity too. As the immediate excitement wears off, researchers have been studying applications, 3D user interfaces, interaction devices, and psychological connections with virtual reality.

Wide ranges of applications have blossomed from VR, including psychological therapy (Rothbaum et al., 1999; Parsons & Rizzo, 2008). This has brought about a lot of attention to the many applications of VR. Another powerful use of VR is with training and simulations without risk of failure, such as medical surgeries (Seymour et al., 2002). With this wide variety of applications, the design and constraints are specific to each use case. Different amounts of fidelity are required for some, and others require certain amounts of immersion or realism to be effective. Bailey, Bailenson, Won, Flora, and Armel (2012) conducted a study of how presence, a measure of how much someone feels like they are in the virtual environment, can help influence behavior change, but hinder memory.

Two major components of designing a VR application are levels of immersion and presence. Presence, as discussed prior, is a subjective metric of actively feeling that you are in the environment you are exposed to. This could be required for particular uses, as it is highly associated with emotional response (Diemer, Alpers, Peperkor, Shiban, & Mühlberger, 2015). The second component is fidelity, sometimes known as immersion. This can be further broken down into display fidelity, which is the quality level of things you perceive, such as graphics or audio in a virtual environment, and interaction fidelity, which is about how realistic and effectively you can interact with the environment. We will be focused on the former, which includes the audio reproduction methods. Levels of fidelity can influence both effectiveness of VR environment and levels of presence experienced by the user.

Zimmons and Panter (2003) created a study which measured performance of a task and associated presence through physiological and self-reported measures. This showed relationships between how well the participant did and the presence that was self-reported and that correlated with heart rate. It should be noted that the task involved heights, which could evoke higher levels of heart rate and emotion. Bowman and Mcmahan (2007) discuss how there may only be a certain level of display fidelity required for doing certain tasks with high performance. This explains that for certain uses, high display fidelity is not required.

Hendrix and Barfield (1995) did a study measuring presence with questionnaires with varying levels of display fidelity. Adding components such as stereoscopic cues or spatialized audio, the researchers allowed participants to navigate a virtual environment and report the presence. When these increases of fidelity were added, the user reported more presence. The largest problem with this study is there was only the addition of spatial audio or none, which didn’t see the relationship over a more gradual change.

AR and VR studies can be helpful to one another due to their similar research needs and application designs. AR, instead of completely putting a user in a new synthetic environment, just augments or changes the current environment to enhance experience. This is usually done with a transparent visor or glasses which have images projected on the screen and ways to interact with them.

A study done by Kytö, Kusumoto, & Oittinen (2015) aimed to find the significance of the Ventriloquist Effect in AR. They found that the Ventriloquist Effect was upwards of 5° to 15° larger in AR than in a normal environment. This means that users could not tell the difference between where the audio was coming and a visual source if they were less than 30° from each other. This has significant importance when designing AR since you want to have a way to point the user to different points of interest with both visual and audio cues while they also can see the environment around them. VR also can utilize this information for design of virtual environments, however the significance in the ventriloquist effect may differ.

Creating things with VR, like a VR workstation instead of the common cubical in most offices, sounds like it could be beneficial but needs proving. Slater, Linakis, Usoh, Kooper, and Street (1996) created a study with 3D chess, in which they tested users' spatial abilities and task performance with different levels of display fidelity. As fidelity increased, so did performance. Presence increase also seemed to be more self-reported with the increase of fidelity but was not related to performance.

One difficult task is understanding what induces feelings of presence and how to measure it. Since it is mainly subjective, questionnaires have currently been accepted as the main method to measure it. Slater, Usoh, and Steed also created a presence questionnaire, called the Slater-Usoh-Steed (SUS) presence questionnaire, which was tested and found to have a high mean score when compared to another presence questionnaire (Usoh, Arman, and Slater, 2000). This questionnaire is used frequently by VR researchers. By using physiological measures, researches could get a better metric to measure presence (Zimmons et al., 2003). However, depending on the environment, something like heart rate would not be accurate. There may be better ways of measuring presence, which will be considered much later in this thesis.

Due to knowledge gaps in display fidelity and presence relationships, particularly with regard to gradual changes in audio fidelity, this study aims to shed more light on the effects and requirements of audio fidelity for virtual reality applications.

2.6 Literature Summary

The three main topics covered in this section were human perception of sound and visuals, spatial audio algorithms, and virtual reality applications in response to the two prior topics. These three overlap for virtual reality, since VR requires a large amount of understanding of how humans perceive their environments in order to create an effective one. Hearing being one of the largest relied-on senses next to vision, it constitutes an understanding of how to best recreate accurate sound fields and how they interact with visual aspects.

Psychoacoustics covered the two main localization cues, ILD and ITD. These two factors determine how humans use them to localize sound around them, when they are effective and when they fail. The reliance of visual cues skewing auditory cues can both be problematic but also useful when trying to recreate accurate environments under physical constraints. Periphery and bimodal perception are two interesting and complicated tasks that may need to be tackled separately after the initial study.

Recreating a 3D sound field is complex and has many ways to achieve such a task. Within a virtual environment, creating these sound fields can improve the effectiveness of a VR application by increasing the user’s understanding and task performance. The method of ambisonics was explained, including why it was used in the study and how it works. The robust nature of ambisonics, the available tools with the spatial algorithm, and the acceptance of the VR application are the main reasons for choosing so.

Display fidelity and presence are two important factors to consider when creating virtual environments. Display fidelity seems to be important in affecting presence and developing effective VEs, but audio aspects have been neglected when it comes to VR. The coming study aims to help fill that knowledge gap.

Methodology

3.1 Body of Participants

The experiment had 18 participants, who were all deemed healthy enough to take part in the study. The volunteers were from 19 to 35 years old, with 1 being female and 17 being male. They were all either students, graduate students, post doc, or had just completed their degree. All 18 had either normal vision (able to see without glasses) or corrected to normal vision (wore glasses or contacts to correct vision). They all self-selected themselves to participate after reading advertisement of the study through emails or flyers, which noted they would be exposed to a virtual environment which could promote sickness or discomfort and in which they would be required to move their arms and hands around. During the study, they were also checked for any hearing loss. If they didn't pass, they would be dismissed; however, none of the volunteers failed any of the requirements. Each participant was also asked how tired they were, how much VR experience they had, and also how much video game experience they had. I hypothesized these factors could be correlated with results, such as experience in VR aiding in performance. After the experiment, participants had to fill out a questionnaire about how the experience was. Some of these questions reflects self-reported presence spoken about in section 2.4, allowing for analysis after the experiment.

3.2 Lab Setup

The experiment, aside from the hearing test and signing of the consent form, took place in Virginia Tech's ASPIRe Lab. This lab was built and supplied by the primary investigator of this research, Dr. Michael Roan, in conjunction with Virginia Tech. The lab is built for the purpose of immersive reality, combining a dense field of surrounding high quality loudspeakers with computers able to stream data to all speakers and run virtual reality software. This becomes a small version of The Cube, which is a similar facility in the Moss Arts Center at Virginia Tech. Later, this space will be discussed due to a model of it being used for the virtual environment within the experiment.

The ASPIRe Lab is similar to The Cube due to its densely packed 44 JBL LSR305 loudspeakers located around the room. Twenty-one speakers are situated mainly in a ring at seated height, while the rest are situated on a higher tier near the ceiling, with a few on the ceiling and floor. Figure 3.1 shows a configuration of the room and what it looks like normally.

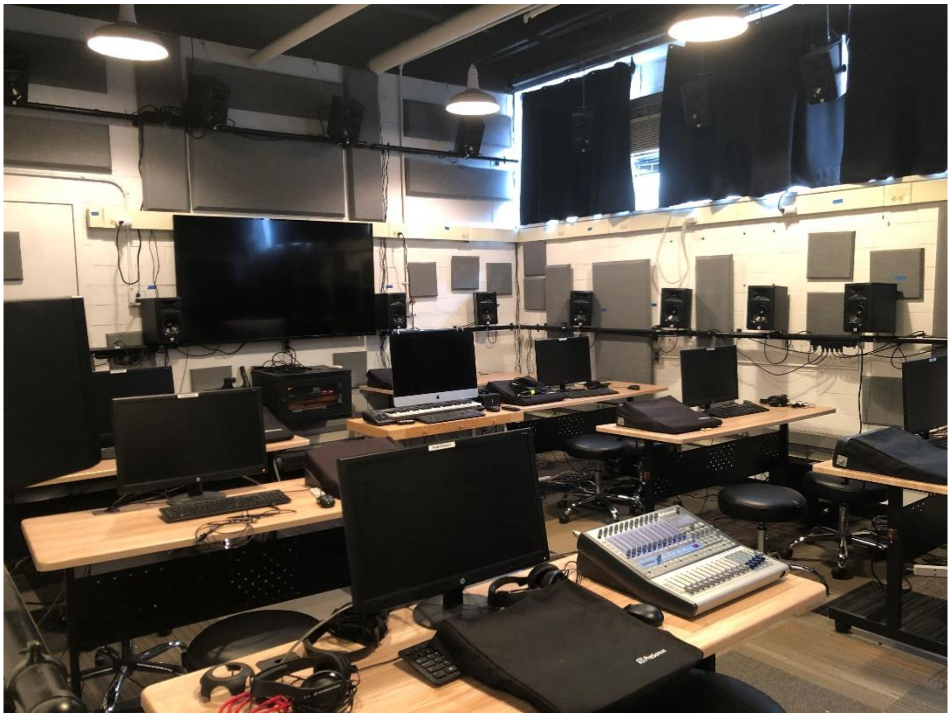


Figure 3.1: The ASPIRe Lab at Virginia Tech where the study was performed.

Each computer is outfitted with a GeForce GTX 1060 graphics card to enable the Oculus Rift CV1 to run on. The Oculus Rift is a head-worn display for experiencing virtual reality, with two OLED panels for each eye. Each display has the resolution of 1080x1200 and a global refresh rate of 90 Hz. The field of view can range due to the interpupillary distance lens adjustment knob, but can be up to 110 degrees. Prescription glasses can be worn with the Rift, which allowed for users during the study to maintain their corrected vision. The HWD is about a pound in weight and has onboard 360 sound, which was moved out of the way for these experiments. These specifications allow for an adequate experience for the users to complete the experiment comfortably and accurately to the best of their ability. Figure 3.2 shows the Oculus Rift and Figure 3.3 displays the Oculus Touch right controller.

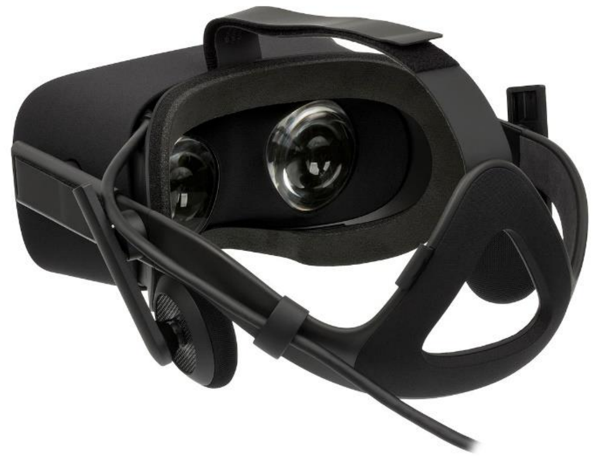


Figure 3.2: Oculus Rift Head-Worn Display. This HWD was used to view VR environment for the experiment. Source: [public domain] Author: Evan-Amos, Wikimedia.org

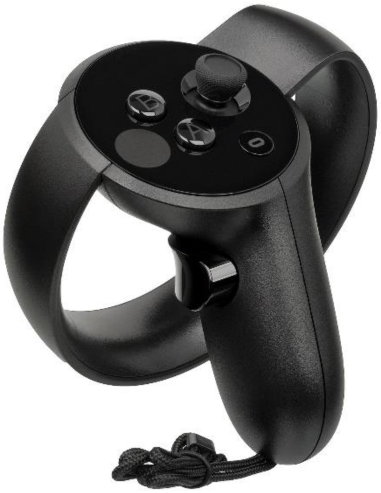
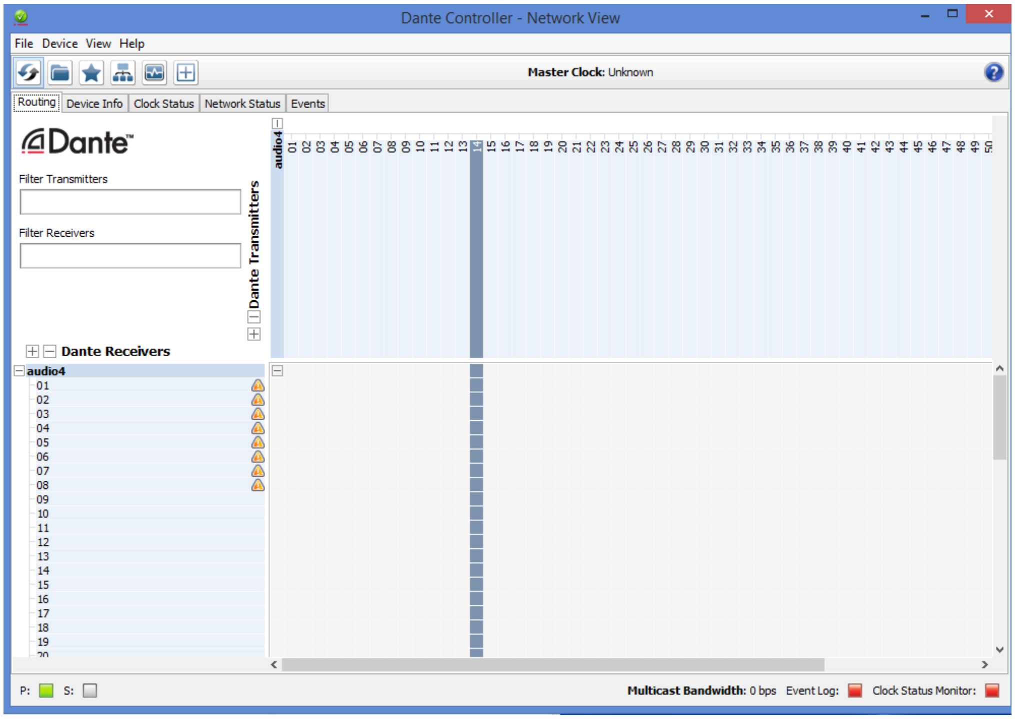


Figure 3.3: Oculus Touch Right Controller. The controller was used during the experiment for interacting with the VR environment. Source: [public domain] Author: Evan-Amos, Wikimedia.org

These desktops, which were used during the experiment, were all linked by Ethernet with Dante^{TM} Virtual Soundcard to an Atterotech unD32. This Dante^{TM} Break Out interface is capable of outputting 32 balanced analog outputs. These outputs are assigned using Dante^{TM} Virtual Sound Card (Dante^{TM} Controller) or directly on the box, featured in Figure 3.3 below. This takes the signals sent out of Ethernet cables and routes them to the JLB Loudspeakers throughout the room. An example of these speakers is featured in figure 3.5.



Master Clock: Unknown

Multicast Bandwidth: 0 bps

Figure 3.4: Dante Controller User Interface used to assign the input channels to output speakers

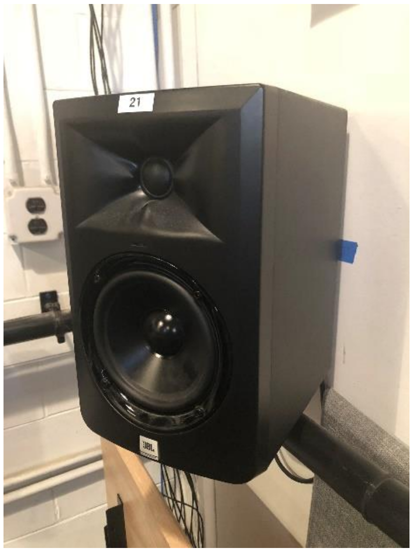


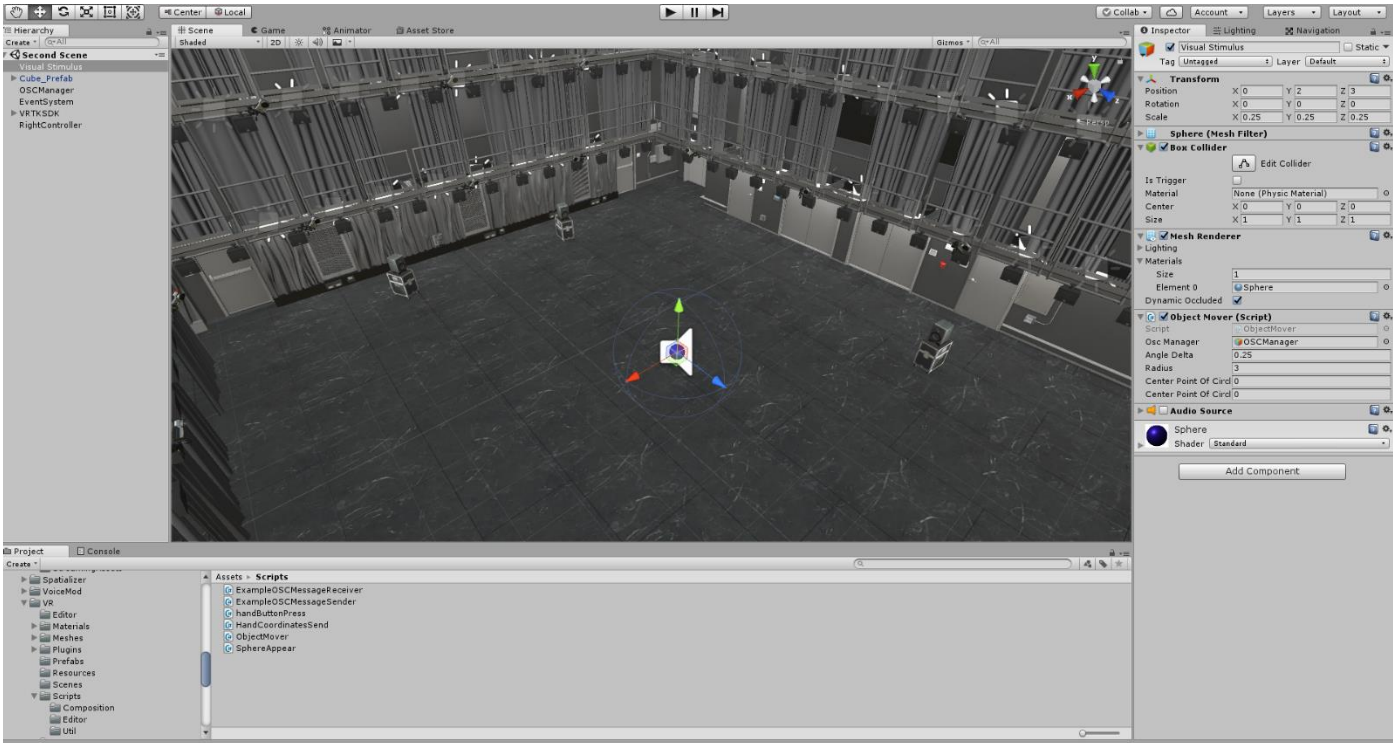
Figure 3.5: JLB Speaker used for recreating sound field in ASPIRe Lab

3.3 Virtual Environment Development

To develop the 3D virtual environment, Unity game engine was used. This engine had the tools to import the model of The Cube, which was created by a Masters of Fine Arts student Lucas Freeman at Virginia Tech. The cube is a four-story, 50x40-feet, acoustically treated black box theater outfitted with capabilities for large immersive environments, research, audio and visual art, and is outfitted with 124 loudspeakers. The Cube model was used due to its similarities in capability to reproduce sound, its large scale which lent more freedom in design of the experiment, and familiarity to students and faculty participating. Users being placed inside a real place that they are familiar with can allow for realism, which I speculate can aid in understanding of the virtual environment.

Unity enables seamless Oculus Rift integration, allowing for accurate and simple development for the Rifts used. The mapping was taken from the Oculus sensors onboard the Rift and the Oculus Cameras placed on either side of the user during the experiment. These tracked the user’s head orientation and movement, although the participants were not supposed to move from the spot in the center of the lab. Also, the right hand of the user was tracked by an Oculus Touch controller, which was used to point at objects in the room or virtual environment, with a laser pointer when the trigger was depressed.

Inside the virtual environment, The Cube was almost completely empty. The user was placed in the middle of the room and could look around freely with the Rift. They could also move their hand around and see the Oculus Touch controller in the exact translational and rotational space that it was in in real life, giving the user an accurate and connected experience. Figure 3.6 shows an image of the developing environment and the layout of virtual environment. The user appeared to be directly in the middle of the room, where an object is pictured.



Second Scene
- Visual Stimulus
- Cube_Prefab
- OSCManager
- EventSystem
- VRTKSDK
- RightController

☑ Visual Stimulus □ Static
Tag Untagged Layer Default

Transform
Position X 0 Y 2 Z 3
Rotation X 0 Y 0 Z 0
Scale X 0.25 Y 0.25 Z 0.25

Sphere (Mesh Filter)

☑ Box Collider
Is Trigger □
Material None (Physic Material)
Center X 0 Y 0 Z 0
Size X 1 Y 1 Z 1

☑ Mesh Renderer
Lighting
Materials
Size 1
Element 0 Sphere
Dynamic Occluded ☑

☑ Object Mover (Script)
Script ObjectMover
Osc Manager OSCManager
Angle Delta 0.25
Radius 3
Center Point Of Circl 0
Center Point Of Circl 0

□ Audio Source

Sphere
Shader Standard

- Spatializer
- VoiceMod
- VR
  - Editor
  - Materials
  - Meshes
  - Plugins
  - Prefabs
  - Resources
  - Scenes
  - Scripts
    - Composition
    - Editor
    - Util

Assets ▸ Scripts
- ExampleOSCMessageReceiver
- ExampleOSCMessageSender
- handButtonPress
- HandCoordinatesSend
- ObjectMover
- SphereAppear

Figure 3.6: The development environment used, picturing The Cube that the user experienced.

When the trigger was depressed, a blue sphere, contrasting the dark grey colors in the cube, appeared. This was at a distance of 3 meters away from the user at all times, 0.25 meters in diameter, and moved in a hemispherical pattern, 180 degrees around the user. It began right in front of the user at 0 degrees, and then moved right until it reached 90 degrees, and then moved the other way until it reached -90 degrees. It was at a constant speed of half a frame per second, since Unity’s C# scripting uses an Update call each frame. The engine ran at 60 frames per second, allowing the sphere to move 30 degrees a second. This took 6 seconds for it to go a full cycle, and this is a speed that could be easily tracked by any healthy adult. When the trigger was let go on the Touch controller, the sphere disappeared and stopped, and so did the laser pointer. The trigger could be depressed again to continue the cycling. The scripts in the engine were used to track the head movement and controller, create the laser, move the sphere, log and create action with button presses, and send the data to another program. These scripts are included in Appendix A.

The coordinates of the sphere, and the angle at which the touch controller was pointing, were sent from Unity using scripts that took advantage of OSC, or Open Sound Control. OSC is a protocol that is usually used for audio in artistic performances that can be used with computers. These scripts created bundles of information and sent a message each frame using IP addresses and IP ports. It utilizes User Datagram Protocol (UDP) to allow Unity to send these messages to whatever IP of another computer on a similar network. This was chosen due to its quick and simple communication method that doesn’t have retransmission delays, which is good for streaming this data real-time. These transfer times are faster and more efficient than Transmission Control Protocol (Coonjah, Catherine, and Soyjaudah, 2015).

3.4 Max Data Pipeline

These messages then were collected by another computer on the same network which was used to control the audio in the room. These data packages were handled and all data being saved for the experiment was done in MaxMSP. Max is a software which allows for visual programming for music and other multimedia. This was programmed to do three main things: take in the data exported from Unity and convert it into spatial audio pink noise, create a way to shift the audio away from the coordinates in a controlled fashion, and record all data into files.

This was also used for the baseline test given to each participant before the main trials to get an understanding of how each participant could pick out where audio was coming from around the room. For this it used a random number generator to pick out the 11 speakers within 90 degrees to the left and right of the participant on the height of their seated level. Once it picked a speaker, it sent pink noise to the speaker. Then it could randomly pick another, without repeating, until all speakers were chosen once.

The program used a UDP receive command to take the data from Unity and route each numeric value to where it needed to be in the data pipeline. The translational coordinates of the sphere were taken and sent into an Ambisonics Tool created by Zurich University of the Arts. This tool first used an "ambimonitor," which is a GUI object to track and output the spatialized coordinates into polar coordinates. This data was then interfaced with a linear timed ramp tool to generate a slowly increasing offset. This allowed for the audio source to drift away at controlled speeds and direction. A second ambimonitor was used to track this data as well. This data was then encoded and decoded using the speaker locations in the room with ambisonic tools. Afterwards, the decoded signals were routed to the speakers in the room. This output whatever signal was required, in this case pink noise. Pink noise has equal energy in every octave as opposed to white noise having equal energy over each frequency. This can be more suitable for human listeners, because it is perceived not as high pitch as white noise and therefore not as bothersome.

Then the time data, the visual and audio source coordinates and the difference between the two's angle, and the angle of the user's pointer were all logged. This only took data when the trigger was depressed by the user, and the data collected was cleared out between trials, separated, and printed into text files for later import for analysis. Below, figure 3.7 is a screenshot of the entire data pipeline in Max.

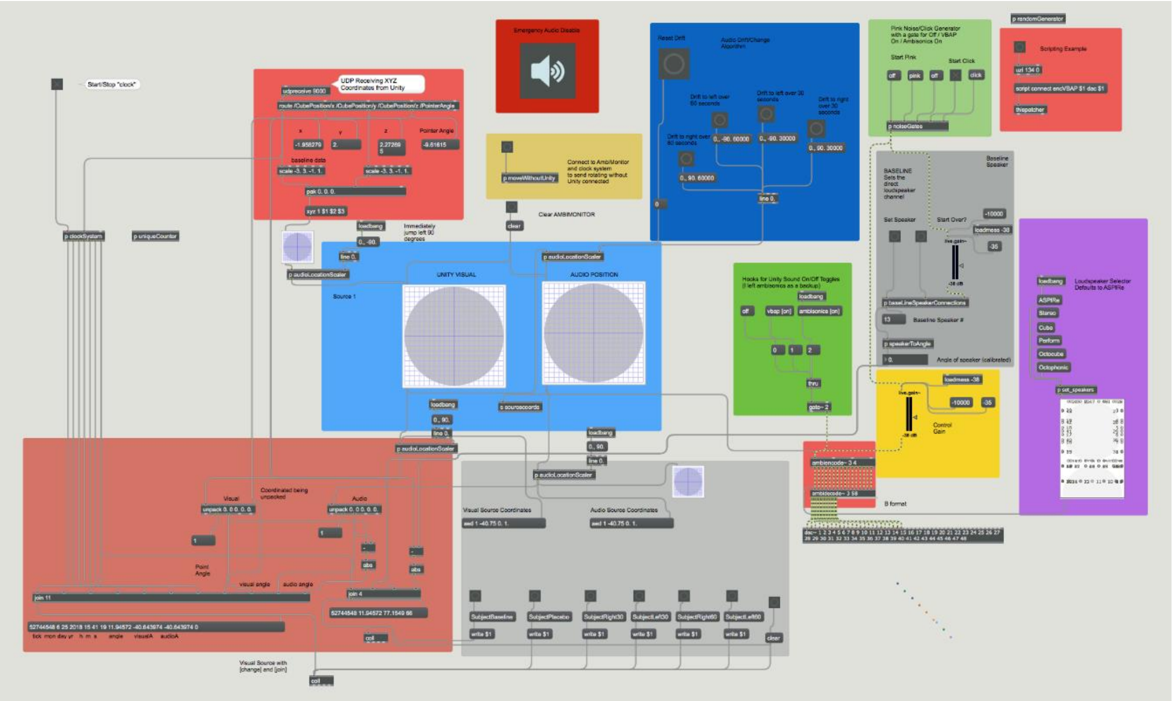


Figure 3.7: The entire data pipeline programmed in Max.

As you can see in the figure, the data flow is complicated for use. A simpler user interface was created in the program’s “presentation mode” for use during experiments, featured in figure 3.8. It contains:

- monitors to check where the sources are visually and aurally and all the data,
- buttons to trigger events for the trials, baseline tests, turning off the audio,
- sliders to change the dB levels of the audio with limiters to prevent anything from being too loud or quiet for users, and
- buttons to save and clear data being logged.

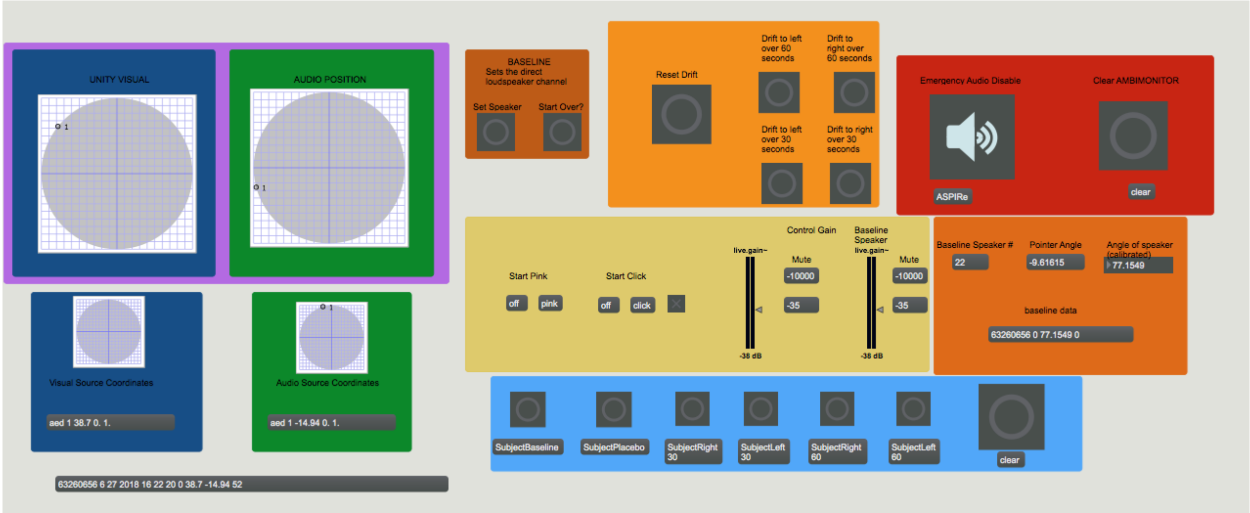


Figure 3.8: The user interface used during experiments to control, monitor, and store data.

3.5 Experimental Design

In order to find out how well humans without any major hearing loss could identify when a visual component is out of sync with an aural component in a virtual environment, a study with human participants was necessary. It required a way for participants to focus on a task, in which they had to notify as soon as a sound source was perceived to be originating from somewhere other than the visual stimulus they were focusing on. The focus is important, since performance is typically skewed by lack of focus. Therefore, the experiment needed a way for the user to interact with the virtual environment, enabling the user to focus on said task. The participant needed to be able to both be included in the virtual environment to have a stronger sense of presence of the task and be tracked with some sort of input device (Usoh et al., 1999).

The chosen virtual environment was a real place in the Moss Arts Center at Virginia Tech called The Cube, which has the capacity to do this test. The access to the facility was limited, and the experiment could be replicated in the Virginia Tech ASPIRe lab at smaller scale with similar capability. This chosen environment was picked due to it being a real place, to create a more realistic environment. Since it is capable of the same things, and since Virginia Tech faculty and students mainly know about The Cube, it will be a believable connection as opposed to a completely fabricated environment.

The participant needed to keep their head aligned with the visual source throughout the study, so that when the audio source deviated from the visual, the user could tell that something had changed. This is due to the HRTFs of the human, acoustic shadowing, and other effects that allow humans to perceive sound directionality, as discussed previously. As mentioned, human perception decreases significantly when dealing with stimulus in the periphery. It can be unduly difficult to distinguish whether sources have separated if they did so in the periphery, so it was decided to allow the user to move their head to track the stimuli. The goal was to utilize a moving visual source so that the user maintained focus.

One issue that this introduces is the necessity for the participant to have the ability to turn their heads in sync with the visual source. Therefore, the user sat on a stool that could rotate about the base, and they were tasked with pointing at the visual source while only pivoting via the stool.

We tracked the hand and the pointed direction through the use of a tracked input device. This displayed where the user was pointing, showed a visual representation of the controller to allow the user to feel confident where their hand and controller were in space, and allowed me to determine where the user was pointing and when they felt the source separated.

The visual source moved with a paired audio source, and the user had to track the visual with their head. While following along that path, the virtual audio source slowly drifted in the opposite direction, to the left or the right depending on which direction the sphere was moving at that point in time in relation to the user. No elevation was introduced since that would have added unnecessary complications to the experiment, although the lab was capable, meaning a similar feature might be added in future studies. The two sources were tracked and the precise coordinates were logged very quickly in order to get an accurate representation of where the visual and audio stimulus was. The user tracked the visual source until they felt like the source was separated, and at that time they performed some sort of input that stopped the audio and visual from moving. This data from the start of the sequence until this point was saved, and the difference between the audio and visual source in degrees was recorded along with the tracking data of the hand orientation and the time. This allowed for analysis of how far to the left and the right, in multiple sequences, the user could tell the difference. Also, it allowed us to check if the user was correctly tracking the visual source to retain focus and keep the body and head aligned.

Prior to the main test just described, the users were required to prove that they could participate in the study. First they needed to have corrected or normal vision. This was important so that the virtual environment was accurately represented. Another thing they needed to have was no significant hearing loss. This is defined as the participant hearing from 0 to 25 dB minimum (World Health Organization, 2018). Anything above is categorized as mild or higher. Therefore, when looking for participants, I sought users who had no knowledge of hearing loss in order to filter out those who knew they had hearing loss. Since age also tends to correlate with hearing loss, students would also be more likely pass a hearing test due to them being younger.

One last requirement of the study was in regard to range of motion of the arms and neck, to ensure that the user was physically able to perform the task. While the testing was done, the participant was monitored for any signs of motion sickness, or any other discomforts. They were allowed to take breaks between segments and were provided with an introduction and practice with all of the tools used during the experiment.

3.6 Procedure

First, the participant arrived to acoustics lab adjacent to the ASPIRe Lab for the hearing test. This is where they each signed the consent form, and a copy for themselves if they required, and asked any questions. Then they were asked about their eyesight or glasses. After this they underwent the hearing test with the experimenter.

For the hearing test, the participant sat opposite of the experimenter with their back turned so that the test would not be easily cheated. Since the experimenter had to manually trigger the sounds and the changes, watching slight movements would have been possible even with the small barrier audiometer in the way. Turning the participant around ensured that they wouldn't be able to see any movements. They placed the headphones on, with the designated cup on each ear. For reference, Figure 3.9 below shows the Welch Allyn AM 232 Manual Audiometer which was used during the experiment.

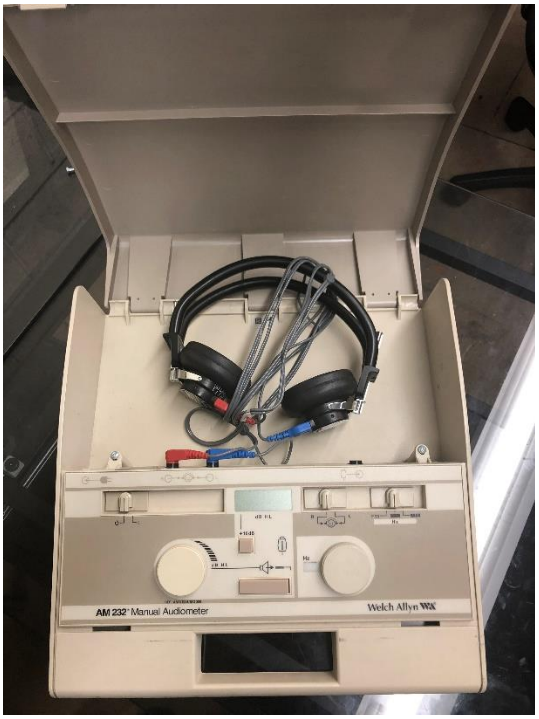
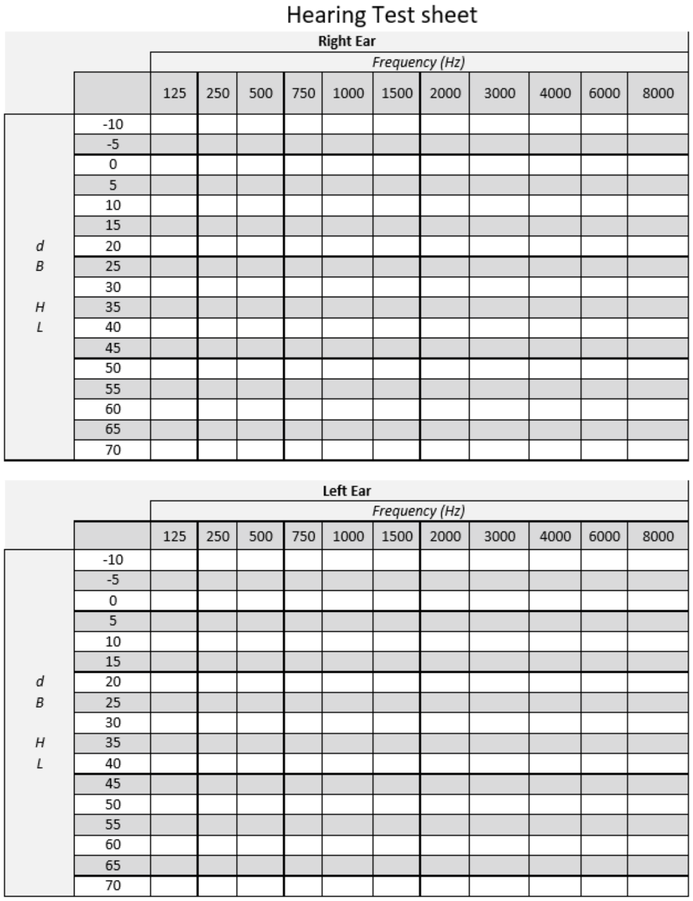


Figure 3.9: Allyn AM 232 Manual Audiometer, which has a set of headphones and a control panel for hearing tests.

The participant first completed a practice round to ensure that they understood the concept and the task required of them. A 125 Hz, 1000 Hz, and 6000 Hz pure tone was played at 50 dB and they were to raise their hand for the corresponding ear they heard it in.

Once the practice was over, the experimenter went from 70 dB and downward in each frequency. The audiometer used can hit all the frequencies and the dB in figure 3.10 below. If the user could hear, the next dB was half of that dB. If they couldn't hear it, it was raised by 1.5 times. This was done until each frequency hearing threshold of the user is found for each ear. The test required participants to hear each frequency at 25 dB or less to continue the study. If the participant could not, they would have been dismissed to try to minimize bad data collection for the experiment.



Hearing Test sheet

Right Ear

		Frequency (Hz)										
		125	250	500	750	1000	1500	2000	3000	4000	6000	8000
	-10											
	-5											
	0											
	5											
	10											
	15											
d	20											
B	25											
	30											
H	35											
L	40											
	45											
	50											
	55											
	60											
	65											
	70											

Left Ear

		Frequency (Hz)										
		125	250	500	750	1000	1500	2000	3000	4000	6000	8000
	-10											
	-5											
	0											
	5											
	10											
	15											
d	20											
B	25											
	30											
H	35											
L	40											
	45											
	50											
	55											
	60											
	65											
	70											

Figure 3.10: Hearing test chart used for checking if any participant had significant hearing loss.

Afterward, participants filled out the pre-questionnaire. This data was used to supplement the understanding of the user demographic as commonly used in IEEE and other journals when doing virtual reality studies. This allows us to know things like how familiar someone is with virtual reality and how this could alter our results. An example questionnaire can be found in Appendix C.

Next, the participant and experimenter made their way to the ASPIRe lab, where the rest of the session took place. The ASPIRe Lab was set up to specifically perform this experiment, with a stool in the center of the room at the sweet spot, which was spoken about in section 2.4. The participant sat on the stool and was introduced to the room with the equipment aforementioned in the lab setup section. They were explained how each works and how their interaction would impact the virtual environment. If they had no prior experience with virtual reality, they were introduced to a simple environment and could move their hand and head around within the space.

The first test administered for each participant was the baseline test. This test channeled pink noise to the speakers situated in 180 degrees around the user. These were the ones used in the baseline since they were the only ones needed for the disparity test. This test, using Max MSP, involved routing this signal to these 11 speakers randomly, and not more than once. The Unity program on an adjacent computer was started beforehand so that the angle of where the user pointed with the tracked Oculus Touch Controller was recorded live. This is due to Unity and Max computers being different, and the computer running Unity dictates all the Oculus accessories. As long as the user stayed within the bounds, the Oculus Sensors could capture the angle at which the user was pointing. The bounding box was within arm's reach of the stool, and around the periphery equipment in the room. This is to prevent participants from hitting anything in the real environment while in the virtual environment.

The experimenter clicked a toggle, which each time randomly routed the signal, at which time the user pointed to the speaker they believed the noise was coming from. They were allowed to take as much time as needed without leaving the sweet spot, and they were asked to turn their head toward the noise to aid in locating it. The pointing was to be done with an outstretched arm, due to calibration being performed like this. After pointing, they were to click the trigger on the Touch Controller. This sent data of the angle of the speaker and the angle in which they were pointing for as long as the trigger was depressed at about 60 Hz. This data was averaged to account for small adjustments in the arm and the human imperfection in holding an outstretched hand.

The major point of this baseline test was to be able to tell how well each participant could locate the source of the sound. These difference were later averaged and considered when analyzing the data. If a user with very poor sound localization did poorly during the disparity test afterward, it could be attributed it to this baseline check. As reference, even when the experimenter tested themselves with the knowledge of which speaker was making the noise, there was still disparity of up to 10 degrees off. This could be contributed to the user not exactly holding the controller forward, not being able to point perfectly and still, and the Oculus Unity system not perfectly being aligned with the room due to human error. This is unimportant since the angle pointing to the wrong speaker would be much farther off, and I would check physically to make note if any participants pointed at the wrong speaker.

After the data was collected from the baseline test, the user took the Oculus Rift HWD and put it on. They were instructed to depress the trigger once they were ready, which caused the visual source or sphere to begin in front of them and start revolving around them in a semi-circle with pink noise perfectly matching. Then they were to trace the sphere with the pointer in the virtual environment and follow the sphere by pivoting around the stool. Once the user felt that the sound no longer was coming from the sphere, they let go of the trigger. This caused the sphere and noise to stop, and the data stopped being collected. This data was logged and the next test began.

The user was firs guided through a practice, so that they had a chance to see what was coming and wouldn’t have such a drastically poor performance the first time, skewing the data. Afterward, 5 tests were performed. One was a placebo in nature, keeping the audio and visual source together. The other four drifted either left or right over a 30 or 60 second period. By the end of the period, the audio was situated 90 degrees in that drifted direction from the sphere, which is quite noticeable especially if they performed as expected on the baseline. All 5 were recorded, with hand movement tracked, as well as the angle difference between the aural and visual stimulus. An exception to this was if the participant chose the incorrect direction (choosing left when the audio drifted to the right of the sphere, or vice versa). Choosing the wrong direction indicated that the participant could have been guessing or not understanding the task. Therefore, if that occurred, those trials were repeated at random until the user chose correctly to eliminate skewed data.

In between each of these tests the participant had a chance to take a break, taking off the HWD and putting down the controller. This was to prevent loss of concentration, give their arms a break, prevent eye strain, and provide the experimenter a chance to check the user for any symptoms of sickness. The questions that were prepared in advance, in the event that a participant needed a break, are included in Appendix C.

After the entirety of the testing, the participants were checked on once more and were given the post-questionnaire. This can be found in Appendix C. This was given to check how the user felt about the environment, facilitating a method of assessing how much presence the user experienced. These questions were based on the SUS Questionnaire previously mentioned.

After this questionnaire was completed and the participant had all their questions answered, they were dismissed. All data were organized and all sheets retained in the appropriate locations specified in the IRB protocol.

3.7 Methodology Validation

A study conducted in 2017 by Ballestero, Robinson, and Dance used very similar methodology. Their goal was to recreate VR environments that would be accurate both visually and acoustically. For example, a user could put on a HWD and see a virtual Roman Coliseum, and when music was played it would be acoustically accurate to how it would sound if the user was actually in such an environment.

This study used most methods outlined in the methodology: using Unity and the Oculus Rift for replicating the VE, sending data from Unity to Max Via OSC and UDP, and using Max to reproduce the sound field using ambisonics. One difference was they used Max to develop HRTF models before sending the audio data to the speaker array. They were successful in doing their proposed goal. This study was not found prior to experimental design, however it further validates the tools and methods used during this experiment.

Results and Discussion

During the experiment, data was collected three ways. One minor way was through observation of the participants during the study. This was to attempt to keep consistent testing and note anything abnormal or possibly useful for later. The main method was the quantitative data of the trials and baseline tests. This was all stored and analyzed from the live data recorded via Max during the experiment. The final way was the post-questionnaire, which allowed for self-reporting of difficulty and presence. This could be used to assess whether the data taken during the experiment were influenced by participant's perceived difficulty, and if presence had any effect on the test. The discussion following each of the results will attempt to find trends, draw conclusions, and explain the data collected during the experiment.

4.1 Observations

The participants first had to fill out questionnaires of their general background and knowledge of games and VR. These were used to see if any trends arose, such as experience aiding in performance. This was also used alongside the hearing tests to see if they could participate in the study. This data alongside visual and audible observations gave some light to understanding what could factor into human hearing in virtual environments.

One thing was that there were a total of four wrong answers across all 72 non-placebo trials across all 18 participants. This means that they incorrectly chose the audio drift direction. Two of the four were from one participant. They most likely were rushing, favoring speed over accuracy. Note that there were no implications that the user to go as fast as possible; they were simply instructed to let go when they felt the sound drifted from the sphere. The participants scored largely accurately, approximately 95%, which meant only 5% of trials had to be discarded incorrect drift choice and performed again. However, 50% of answers were incorrect answers when it came to the placebo trials. This means a participant signaled that the stimuli separated and that it either drifted left or right, when truly the stimuli had not done that.

The users who tended to have fast detections of the audio separation (meaning that the angle at which the stimuli were separated when the participants observed there was a separation was very small) were prone to error. The wrong answers were all given within 10 degrees of separation, which was pretty low compared to the mean.

There was a large increase in participant performance as they practiced the trials. This was mostly visible between the first and last practice, which could be upwards of the stimuli being 30 degrees closer together when drift was observed in the last practice versus when it was observed in the first. Unfortunately, the tasks were randomized between studies, and the number of practices or practice data was discarded. Therefore no thorough analysis can be done on performance increase based on training. However, it was quite obvious that the first couple of practices were much worse than those that followed. The participants always increased between the first and second practice, and some continued to improve throughout the trials. This could influence the data and could potentially point to localization being a learned skill that can be improved.

Once the user felt comfortable with the equipment, felt they understood the task, and was ready to begin, they could go onto the real trials. This led to inconsistent practice rounds for participants, although it was done intentionally in order to "level the playing field" with those who had no VR experience and those who were somewhat familiar. This was done in order to avoid outliers from those who first performances were their first interaction with VR environments.

4.2 Baseline Tests

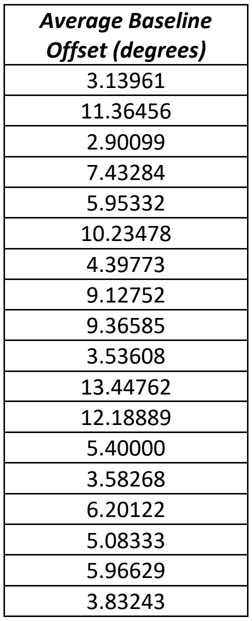
The two tests were the hearing test and the baseline localization test. The hearing test wasn't focused on in analysis and was only required to make sure participants were qualified to complete the study. Each participant successfully completed the hearing test with no signs of hearing loss, at least hearing down to 25 dB and numerous frequencies. One interesting note was that the participants could hear better at all frequencies aside from

125 Hz to 500 Hz. This could be due to the testing equipment not being able to reproduce lower frequencies as well as higher.

The baseline localization test, as discussed in chapter 3, asked participants to point to which speaker the pink noise was coming from. Due to imperfections in calibration between the actual speaker locations and the Oculus sensors' positioning in the room, there was a natural offset. This, in combination with different arm lengths, pointing methods, and shaking muscles, caused a decent amount of natural offset for the baseline test.

The data gathered from pointing to each of the 11 speakers was averaged to one number per participant in table 4.1 below. These 11 speakers were used for the disparity trials after the baseline test. The average between all participants was 6.842 degrees off from the speaker.

Table 4.1: Average Distance in Degrees between Direction Pointed and Speaker



Average Baseline Offset (degrees)
3.13961
11.36456
2.90099
7.43284
5.95332
10.23478
4.39773
9.12752
9.36585
3.53608
13.44762
12.18889
5.40000
3.58268
6.20122
5.08333
5.96629
3.83243

The correlation, Pearson's r, was calculated between these values and the average performance during the following trials. There is not much productive analysis to be done about the baseline test, aside from that there was almost no correlation between how well the participants did on the baseline localization and the disparity test. Pearson's r was

0.136. This is most likely because the baseline was simply used to make sure that participants could reasonably know from where sound was originating.

Since all scored under 15°, and the speakers were at least 15° apart, it felt appropriate that all had sufficient or normal localization ability. If the angle was 30° or higher it could indicate that the participant was pointing to the wrong speaker. This large of a variance was caused by human differences (length of arm, how they held or angled the touch controller) and imperfect calibration, aside from the distance of the speakers apart from one another, which further justifies valid performances across the participants.

From observations, even though all participants averaged at most 15° off, each participant was always pointing to the correct speaker from which pink noise was coming, which can act as a second validation.

4.3 Disparity Trials: Overall

The major goal of the experiment was to determine how well humans could perceive auditory and visual changes in a virtual environment. More specifically, I wanted to see how well someone could localize and determine that a sound was not coming from a visual component, while performing a task in a VE wearing an HWD, using Ambisonics.

The main measure of this would be the mean angle between the visual and auditory component at which participants noticed the disparity. Multiple trials were used in order to collect more data and to see if there was any difference that would arise from direction or speed of drift of the virtual audio source.

The mean score for all 18 participants was 33.663° apart. This was much larger than what was stated by Zwicker et al. at 5°, as mentioned in chapter 2. Although 5° is the minimum possible, there were few who scored that little of a difference. This leads me to believe, from 72 trials, that VR environments make it much more difficult than in reality to localize sound when a visual source is paired. Kytö et al. concluded in AR that there needs to be

30° azimuth between a visual component and a separate audio component for someone to believe that they are not associated. This is further backed up in VR by the 33.663° average within this test.

Table 4.2 shows the general statistics for the four main trial categories. Figure 4.1 shows the data between trials, and figure 4.2 shows the distribution of the average scores from each trial. This is to give general visualizations of the data collected.

Table 4.2: General Statistics of All Four Drift Categories

Groups	***Count***	***Sum***	***Average***	***Variance***	***SD***
Left 30	18	668.1	37.11667	302.3598	17.38849687
Left 60	18	595.89	33.105	337.7609	18.37827306
Right 30	18	623.4	34.63333	290.2051	17.03540811
Right 60	18	536.34	29.79667	170.4943	13.05734501

There were 4 scores under 10°, and 11 between 10° and 20°. This is a very small number of high performance scores; two of the scores under 10° were from a participant who thought there was drift when there wasn’t (placebo) and got two of the four wrong directions on drift. This leads me to believe that they were possibly guessing or pressuring themselves to be quick. Without those two scores, the average jumps to 34.5°.

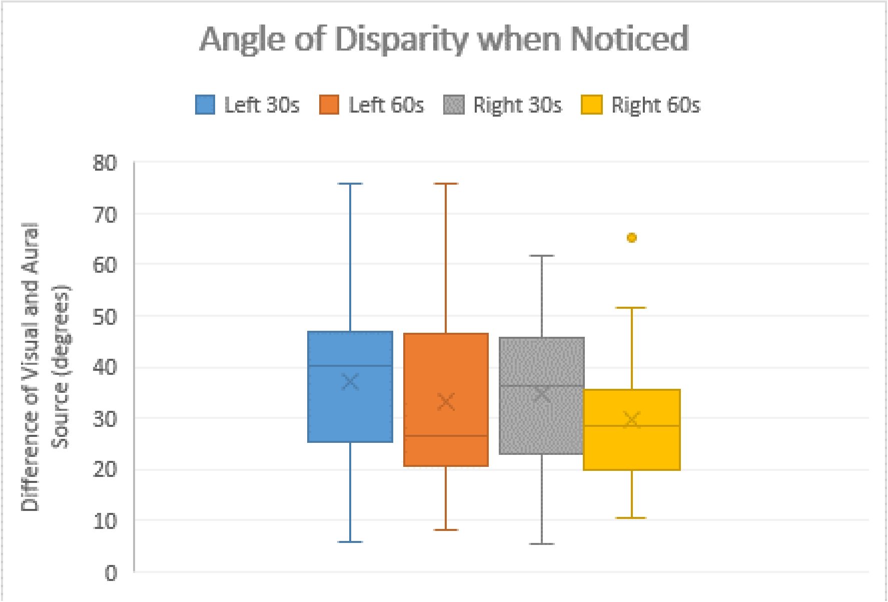


Figure 4.1: The four trial categories are displayed as box-and-whisker plots. All Four categories have similar means.

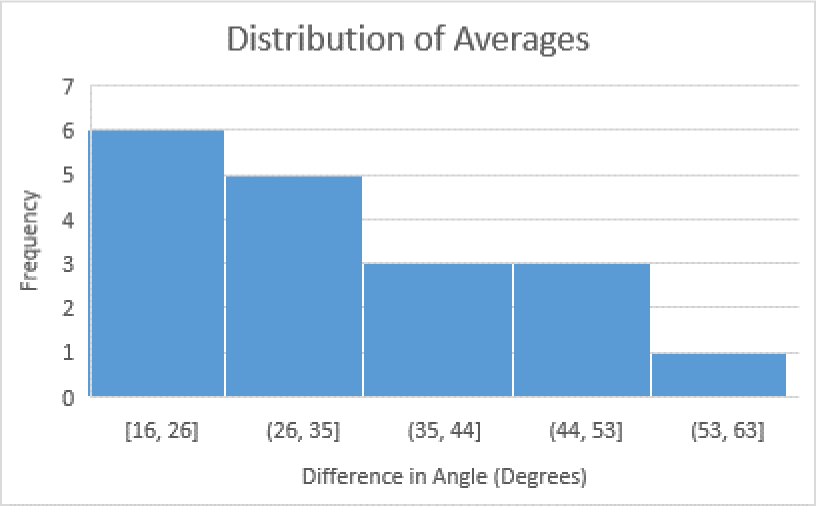


Figure 4.2: The distribution of average scores within a histogram. The data was left-skewed, most scores below 35°.

The participants' individual averages ranged from 16.44° (by the participant with the possible guessing) to 62.6°. This is a fairly large range, and it seems that there may be more testing required to get a more accurate mean. However, looking at Figure 4.2 shows that the distributions of average scores were all mostly from 16° to 26° (6 average scores) and

26° to 35° (5 average scores). This appears to show that most participants had decent localization in VR in comparison to AR (up to 30°), but not nearly like a normal environment (minimum of 5°). This would need to be validated with more participants, as the average is still above 30°.

4.4 Disparity Test: Between Categories

When comparing the data, an important aspect to check is whether there are any trends in left versus right drift or 60-second drift versus 30-second drift. For this analysis, I combined all data pertaining to left versus right drift and compared these averages. Then I did the same for 60- verses 30-second drift. These categories were explained in chapter 3. Figure 4.2 shows a chart comparing the averages of these various drift categories.

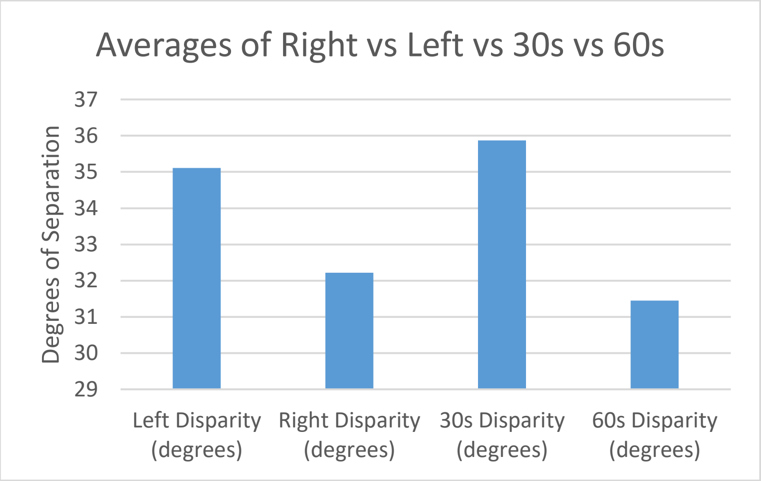


Figure 4.3: The averages between the four categories shown in a bar chart.

The four main categories of trials were left drifting over 30 seconds, left drifting over 60 seconds, or right drifting over 30 or 60 seconds. One reason for doing these different tests was so there could be more trials with the same participants without repeating the same sequence. Without the variety of trial categories, the sequence could have been memorized by the participant and could have led to unreliable answers based more on training than actual perception. The other reason is to compare these categories to see if there are any significant differences between them.

Figure 4.2 shows the averages between the categories. The degrees of separation difference between the 30- and 60-second drifts was approximately 4°. The difference between the left and right categories was even smaller.

In order to see if there was any statistical difference, a single factor ANOVA was performed. Table 4.3 shows the conclusions of this data. The p-value was 0.6087, which is much above the standard 0.05. This led me to not reject the null hypothesis, meaning that the means of all categories were about equal. So even though the figure above shows some difference, due to the small difference and small sample size, the averages are approximately equal. There wasn't any evidence to show that there should be a difference between the categories prior, so this lines up with expectations.

Table 4.3: ANOVA Test for Drift Categories

Source of Variation	***SS***	***df***	***MS***	***F***	***P-value***	***F crit***
Between Groups	506.3266	3	168.775	0.6132	0.608730183	2.7395
Within Groups	18713.94	68	275.205			
Total	19220.268	71				

The student's paired two-sided t-test was performed between the different drift categories to see if there was any possibility that one category was statistically different than another. This could hint at trends between the categories and if one was easier to localize than another. This test was done twice, between the drift left and right, and the 30- and 60-second drift trials. Below shows a table of those two tests.

Table 4.4: T-Test P-values between Drift Categories

T-Test Left vs Right	***T-Test 30s vs 60st***
0.34781375	0.103392637

Between left and right drifts, the p-value was 0.348. Therefore, the means of the left and right drift categories did not differ significantly. This was mainly suspected between the left and right categories with a p-value of 0.348, since there is no current evidence why a normal, healthy human being would favor one direction over the other.

The p-value between the categories for the time it took for a full 90° drift was 0.103. The drift time is a function of the speed, which I speculated could present a change between performances of these two categories. The p-value for this test was much closer to the null hypothesis rejection, but still it is not likely that there is any true difference in these two categories. This follows what was expected, since the drift speeds were 1.5°/s for 60-second drift time and 3.0°/s for 30-second drift time. These speeds are relatively slow for humans to track, compared to the revolution speed of the sphere which was 30°/s.

4.5 Disparity Test: Prior Experience

The few questions that were given prior to the experiment aimed to check for possible factors that could cause changes in performance. Things like how tired the participant was or how much experience the participants had with VR was speculated to influence performance. These questions were featured on the questionnaire, which can be found in Appendix C. A correlation matrix, using Pearson’s r, was created to see if performance average in the disparity trials was correlated with any of these questions. Table 4.5 shows this below.

Table 4.5: Correlations between Prior Experience and Average Score

	Average for Participants
Tiredness	0.095354204
VR Experience	-0.224278469
Video Game Experience	0.467437401

As shown, none of the questions showed any real correlation. Looking at the correlation matrix, there were very low correlation values. The largest and least understandable result was video game experience correlated 0.467 with the performance. The correlation was positive, which meant that the more experience they had the worse they did, since higher degrees of separation of when participants noticed would mean they noticed later and were therefore less accurate. These trends don't seem to confirm any speculation, and possibly just require a larger participant base to confirm them. Video game experience was the closest to showing correlation, but experience being correlated with poor performance doesn't allow for any conclusions to be drawn.

4.6 Presence Questionnaire

Finally, questions regarding presence and difficulty of the task were asked of the participants after the tasks were all completed. These post-study responses were requested in order to see if presence aided in completing the task, and to see if perceived difficulty influenced results – for example, if difficulty correlated with poor performance. Table 4.6 displays the correlation matrix using Pearson's r results. The asterisks signal that the question relates to self-reported presence.

Table 4.6: Correlation Matrix between Questionnaire and Average Performance

	Participants' Averages	***Q1***	***Q2****	***Q3****	***Q4****	***Q5***
Participants' Averages	1					
Q1	0.10009	1				
Q2*	-0.28016	-0.06362	1			
Q3*	**-0.34728**	-0.05381	**0.78106**	1		
Q4*	0.05553	0.08844	0.00932	0.29631	1	
Q5	0.15411	**-0.59591**	-0.10394	-0.08338	0.17608	1

The questionnaire was also compared to itself for a factor analysis, in order to see if questions asking similar things received similar responses. As designed, questions 1 and 5 were about the difficulty, and 2 through 4 were self-reported presence. These questions were answered on a scale of 1 to 5. The presence questions in the post-study questionnaire, although mostly correlated with each other, did not shed any light on if they had an effect on performance during the disparity task. There was only one large correlation between question 2 and 3, which makes sense since both are related to presence. Although, question 4 wasn't correlated with 2 or 3 as much as it should have been, given the correlation of the two other similar questions.

Questions 1 and 5 were also strongly negatively correlated with each other, which was important since both were on difficulty (but asking the opposite question). Unfortunately, there was very little correlation found between the average score of participants and any of the questions asked. This shows that participants, regardless of perceived presence or difficulty, performed about the same.

There was a -0.347 correlation between one of the presence questions and performance, meaning there was a slight trend between how present the participants seemed and how quickly they could identify the difference. Although there is possibility that presence can aid in sound localization and not being confused by the ventriloquist effect, there needs to be more testing done to find a solid answer.

Conclusion

5.1 Conclusions

Virtual Reality technology has come a very long way, from the 1960s until now. The technology is now at the point where it can be used for a wide variety of applications, reaching new heights of potential utilization. However, there is much research that is required to harness such technology and use it effectively.

Research efforts have mainly focused on VR display fidelity, interaction, and presence measures. These are all very important to design applications and fully understand how capable VR is. However, there is a large gap of knowledge when it comes to VR and audio, especially with spatial audio outside of a traditional headphone setup. Therein lies the space for the research performed and described in this thesis.

The main goal of this research was to better understand how virtual environments can affect one's ability to localize sound and to determine what the limits are of an average, healthy human to tell the difference between a visual and auditory component. The experiment was designed to see when someone could tell if virtual sound and a virtual object separated. After conducting the study and analyzing its results, the average score throughout all 72 trials and participants was 33.66° degrees of separation before they correctly identified that it drifted and if it drifted left or right.

Comparing this to Zwicker et al. (2013), who stated that sound localization over the azimuth was 5°, this is much larger. This could largely be due to the ventriloquist effect—Kytö et al. (2015) suggested that within AR, the visual and audio components had to be 30° apart. It seems that AR and VR localization is greatly influenced by the ventriloquist effect, which can both be a design constraint and tool for developers. For VR applications, the environment can use a lower audio fidelity when matching an audio source to a visual source because of this large ventriloquist effect. However, designing an environment with multiple bimodal stimuli at once, it is important to keep them separated avoiding confusion.

This validates Kytö’s work et al., while also showing the parallels between AR and VR environments.

As far as how fast a sound drifts from its visual component or which way it drifts, the human performance of identifying the separation does not differ. Neither does self-reported presence and tiredness, task difficulty, nor experience with VR or video games. However, these may need more investigation and an increased sample size to conclude for sure if there is no influence from these factors.

5.2 Experimental Limitations

The experiment at its conclusion provided valuable insight into how it was limited and how it could be improved. This became evident throughout testing multiple participants, feedback from participants, and analyzing the data.

The first change that I would make is using a gaze pointer instead of a hand pointer. The hand pointer was chosen so that the user could do the baseline test free of the HWD and focus on the task. However, during the actual task completion, it could have been more natural and consistent to use a gaze pointer. A gaze pointer uses the HWD orientation as a laser pointer and tracks where the user looks via head rotation and tilt. This could have kept focus while also ensuring consistent HRTFs by forcing participants to follow with their head. They would also have feedback from the pointer whether they were looking directly at the sphere or not.

Another change is to add a second task. Prior to doing the tasks with the sphere, have them follow the noise with the pointer alone in the environment. This data could be used to see if being in a VE influences localization without the ventriloquist effect impacts.

Adding more to the baseline test, such as putting virtual sources in between speakers, could also be beneficial. This could be more rigorous on participants’ localization ability and further validate the use of ambisonics.

Some logistical notes would be to keep track of incorrect answers and the data along with it. This could aid in figuring out speed-accuracy issues during experiments. Also, if the trials were not randomized for each participant, there could be some analysis of how sound localization could be learned through practice. This would be more effective than just observation.

5.3 Future Work

As this research comes to a close, there are many research objectives that can be derived from this work. The first is simply building onto this experiment, such as altering the aspects discussed above and increasing participant numbers. VR studies tend to have a low amount of participants which is hard to gather strong statistics when dealing with things like behavior. This could provide validation of this experiment and increase our understanding of audio fidelity in virtual environments.

An additional study that could be very useful is similar to this, but instead of using broadband frequency like pink noise, it could employ human speech. Using speech intelligibility and a face on the sphere, one can match mouth movement to the audio cues and see if this has any effect on human hearing and localization. This could provide insight into the ventriloquist effect with human speech and be useful for telepresence VR applications, such as virtual meetings or social gatherings.

Finally, a large body of work has been devoted to presence and finding a quantitative measure instead of the qualitative measure that is used now through questionnaires. I believe that instead of measuring a physiological sense like heart rate, brain waves and event-related potentials (ERPs) could be used as a measure for presence. Using a machine like an electroencephalography (EEG) can be used to analyze the brain's electrical activity. Kober and Neuper (2012) stated that this is possible, and looking for mismatch negativity and other ERPs such as slow waves could provide insight into such measures of presence.

Using the current or similar experimental setup, and adding an EEG to observe the brain while it realizes the sound and sight disparity, could highlight certain brain waves or signals that are linked to presence. When the brain realizes something and allocates resources to the attention of this physical change, there will be a signal that may be measurable. This could be insightful as it could be used as another measure of presence, further boosting our understanding and development of VR technology for the future.

References

Anderson, P. L., Price, M., Edwards, S. M., Obasaju, M. A., Schmertz, S. K., Zimand, E., & Calamaras, M. R. (2013). Virtual reality exposure therapy for social anxiety disorder: A randomized controlled trial. *Journal of consulting and clinical psychology*, *81*(5), 751.

Bailey, J., Bailenson, J. N., Won, A. S., Flora, J., & Armel, K. C. (2012, October). Presence and memory: immersive virtual reality effects on cued recall. In *Proceedings of the International Society for Presence Research Annual Conference* (pp. 24-26).

Ballestero, E., Robinson, P., & Dance, S. (2017). Head-tracked auralisations for a dynamic audio experience in virtual reality sceneries.

Bolia, R. S., D'Angelo, W. R., & McKinley, R. L. (1999). Aurally aided visual search in three-dimensional space. *Human factors*, *41*(4), 664-669.

Bowman, D. A., & McMahan, R.P. (2007).Virtual reality: how much immersion is enough?. Computer, 40(7).

Brooks, F. P. (1999). What's real about virtual reality? IEEE Computer Graphics and Applications, 19(6), 16-27. DOI: 10.1109/38.799723

Calvert, G. A., Hansen, P. C., Iversen, S. D., & Brammer, M. J. (2001). Detection of audio-visual integration sites in humans by application of electrophysiological criteria to the BOLD effect. *Neuroimage*, *14*(2), 427-438.

Charbonneau, G., Véronneau, M., Boudrias-Fournier, C., Lepore, F., & Collignon, O. (2013). The ventriloquist in periphery: impact of eccentricity-related reliability on audio-visual localization. *Journal of vision*, *13*(12), 20-20.

Coonjah, I., Catherine, P. C., & Soyjaudah, K. M. S. (2015, December). Experimental performance comparison between TCP vs UDP tunnel using OpenVPN. In Computing, Communication and Security (ICCCS), 2015 International Conference on (pp. 1-5). IEEE.

Culbertson, C., Nicolas, S., Zaharovits, I., London, E. D., Richard De La Garza, I. I., Brody, A. L., & Newton, T. F. (2010). Methamphetamine craving induced in an online virtual reality environment. *Pharmacology Biochemistry and Behavior*, *96*(4), 454-460.

Diemer, J., Alpers, G. W., Peperkorn, H. M., Shiban, Y., & Mühlberger, A. (2015). The impact of perception and presence on emotional reactions: a review of research in virtual reality.

Evan-Amos. (2018, July 14). https://commons.wikimedia.org/wiki/File:Oculus-Rift-CV1-Headset-Back.jpg, (accessed April. 29, 2018). Public Domain

Evan-Amos. (2018, July 14). https://commons.wikimedia.org/wiki/File:Oculus-Rift-Touch-Controller-Right.jpg, (accessed April. 29, 2018). Public Domain

Hoffman, H.G., “Virtual Reality Therapy”, Scientific American, August 1, 2004.

Kato, M., Uematsu, H., Kashino, M., & Hirahara, T. (2003). The effect of head motion on the accuracy of sound localization. *Acoustical science and technology*, *24*(5), 315-317.

Kober, S. E., & Neuper, C. (2012). Using auditory event-related EEG potentials to assess presence in virtual reality. *International Journal of Human-Computer Studies*, *70*(9), 577-587.

Kytö, M., Kusumoto, K., & Oittinen, P. (2015, September). The ventriloquist effect in augmented reality. In *Mixed and Augmented Reality (ISMAR), 2015 IEEE International Symposium on* (pp. 49-53). IEEE.

McAnally, K. I., & Martin, R. L. (2014). Sound localization with head movement: implications for 3-d audio displays. *Frontiers in neuroscience*, *8*, 210.

Orr, D. B., Friedman, H. L., & Williams, J. C. (1965). Trainability of listening comprehension of speeded discourse. *Journal of Educational Psychology*, *56*(3), 148.

Parsons, T. D., & Rizzo, A. A. (2008). Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis. *Journal of behavior therapy and experimental psychiatry*, *39*(3), 250-261.

Rothbaum, B. O., Hodges, L., Alarcon, R., Ready, D., Shahar, F., Graap, K., ... & Baltzell, D. (1999). Virtual reality exposure therapy for PTSD Vietnam veterans: A case study. *Journal of Traumatic Stress: Official Publication of The International Society for Traumatic Stress Studies*, *12*(2), 263-271.

Seymour, Neal E., et al. "Virtual reality training improves operating room performance: results of a randomized, double-blinded study." *Annals of surgery* 236.4 (2002): 458.

Slater, M., Linakis, V., Usoh, M., Kooper, R., & Street, G. (1996, July). Immersion, presence, and performance in virtual environments: An experiment with tri-dimensional chess. In *ACM virtual reality software and technology (VRST)* (Vol. 163, p. 72). New York, NY: ACM Press.

Su, T. I. K., & Recanzone, G. H. (2001). Differential effect of near-threshold stimulus intensities on sound localization performance in azimuth and elevation in normal human subjects. *Journal of the Association for Research in Otolaryngology*, *2*(3), 246-256.

Usoh, M., Arthur, K., Whitton, M. C., Bastos, R., Steed, A., Slater, M., & Brooks Jr, F. P. (1999, July). Walking> walking-in-place> flying, in virtual environments. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (pp. 359-364). ACM Press/Addison-Wesley Publishing Co..

Usoh, M., Catena, E., Arman, S., & Slater, M. (2000). Using presence questionnaires in reality. *Presence: Teleoperators & Virtual Environments*, *9*(5), 497-503.

Wang, D., & Brown, G. J. (2006). Computational auditory scene analysis: Principles, algorithms, and applications. Wiley-IEEE press.

World Health Organization. (2018, March 15). Deafness and hearing loss. Retrieved April 18, 2018, from http://www.who.int/en/news-room/fact-sheets/detail/deafness-and-hearing-loss

Zimmons, P., & Panter, A. (2003, March). The influence of rendering quality on presence and task performance in a virtual environment. In *Virtual Reality, 2003. Proceedings. IEEE* (pp. 293-294). IEEE.

Zwicker, E., & Fastl, H. (2013). *Psychoacoustics: Facts and models* (Vol. 22). Springer Science & Business Media.

Appendix A: Unity Scripts in C#

OSCMessageSender: Takes messages from other scripts and sends via UDP to Max

```
using UnityEngine;
using System.Collections;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Threading;
using System.Net;
using System.Net.Sockets;
using System.Diagnostics;
using Bespoke.Common;
using Bespoke.Common.Osc;

// uses the Bespoke OSC implementation, so that must be present in your Unity
project for this to function correctly

public class ExampleOSCMessageSender : MonoBehaviour
{

    public static int localPort = 10025; // this could be any port number, 10025
is randomly chosen

    public string externalIP = "127.0.0.1"; // the IP of the machine to which you
want to send messages
    public int externalPort = 9000; // the port on which the receiving machine
will be listening for messages

    private IPEndPoint localEndPoint;
    private IPEndPoint externalEndPoint;

    private List<OscMessage> messagesThisFrame = new List<OscMessage>(); // a list
of all the messages you Append during a frame, to be bundled together and sent at
the end of the frame

    void Start()
    {
        // initialize EndPoints
        localEndPoint = new IPEndPoint(IPAddress.Loopback, localPort);
        externalEndPoint = new IPEndPoint(IPAddress.Parse(externalIP),
externalPort);
    }

    void Update()
    {

        // call SendBundle to send all the messages together in a bundle
        SendBundle();
    }

    public void AppendMessage(string address, List<object> values)
    {
        OscMessage messageToSend = new OscMessage(localEndPoint, address);
        messageToSend.ClearData(); // do i need this?
        foreach (object message in values)
        {
            messageToSend.Append(message);
        }
        messagesThisFrame.Add(messageToSend);
    }

    //  sends the messages stored in messagesThisFrame as a bundle, then clears
messagesThisFrame
    private void SendBundle()
    {
        OscBundle frameBundle = new OscBundle(localEndPoint);
        foreach (OscMessage message in messagesThisFrame)
        {
            frameBundle.Append(message);
        }

        frameBundle.Send(externalEndPoint);

        messagesThisFrame.Clear();
    }
}
```

handButtonPress: activate laser pointer when trigger pressed

```
using System.Collections;
using System.Collections.Generic;
using UnityEngine;

public class handButtonPress : MonoBehaviour {


    private LineRenderer objectLineRenderer;
       // Use this for initialization
       void Start () {
        objectLineRenderer = GetComponent<LineRenderer>();
    }

       // Update is called once per frame
       void Update () {
              if(OVRInput.Axis1D.SecondaryIndexTrigger > 0)
        {
            objectLineRenderer.enabled = true;
        }

       }
}
```

HandCoordinatesSend: send hand orientation from touch controller to OSC script

```
using System.Collections;
using System.Collections.Generic;
using UnityEngine;

public class HandCoordinatesSend : MonoBehaviour
{

    public GameObject oscManager;
    private float rotation;

    // Use this for initialization
    void Start()
    {

    }

    // Update is called once per frame
    void Update()
    {



        rotation = transform.rotation.eulerAngles.y;

        if (rotation > 180)
        {
            rotation -= 360;
        }


        List<object> OSCPositionMessageA = new List<object>();



        string messageAddressA = "/PointerAngle";


        OSCPositionMessageA.Add(rotation);



oscManager.GetComponent<ExampleOSCMessageSender>().AppendMessage(messageAddressA,
OSCPositionMessageA);

        Debug.Log("Angle Pointer = " + rotation);
    }


}
```

ObjectMover: controlled and moved sphere around user, and send coordinates to OSC script

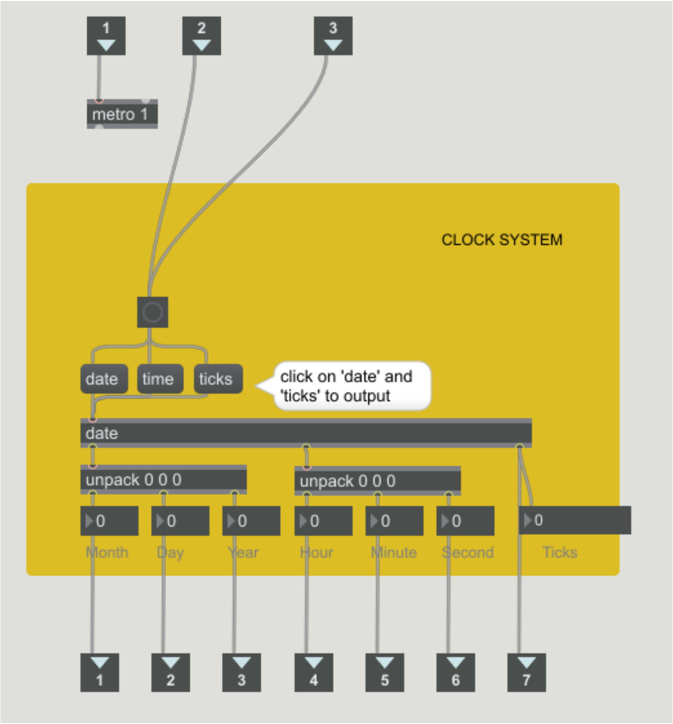
```
using System.Collections;
using System.Collections.Generic;
using UnityEngine;

public class ObjectMover : MonoBehaviour
{

    public GameObject oscManager;


    float angle = 90; // in degrees
    public float angleDelta = 0.5f; //number of degrees to move per frame ... 1 is
too fast
    public float radius = 3; //number of meters from center point of the circlle
    public float centerPointOfCircleXCoordinate = 0;
    public float centerPointOfCircleZCoordinate = 0;
    private bool flip;
    private MeshRenderer mesh;

    // Use this for initialization
    void Start()
    {

    }



    // Update is called once per frame
    void Update()
    {

        if (OVRInput.Get(OVRInput.Axis1D.PrimaryIndexTrigger,
OVRInput.Controller.RTouch) > 0)
        {

            float xCoord = centerPointOfCircleXCoordinate + radius *
Mathf.Cos(Mathf.Deg2Rad * angle);
            float zCoord = centerPointOfCircleZCoordinate + radius *
Mathf.Sin(Mathf.Deg2Rad * angle); // xCoord and yCoord or x and z or whatever
based on what axes you want it to move)
            angle += angleDelta; //increase the angle every frame to make the
sphere move around the semicircle



            // if angle has hit either side of the semicircle, then reverse the
direction
            if (angle >= 150 || angle <= 30)
            {
                angleDelta *= -1;
            }
            transform.position = new Vector3(xCoord, 2, zCoord);

        }

        List<object> OSCPositionMessagex = new List<object>();
        List<object> OSCPositionMessagey = new List<object>();
        List<object> OSCPositionMessagez = new List<object>();



        string messageAddressX = "/CubePosition/x";
        string messageAddressY = "/CubePosition/y";
        string messageAddressZ = "/CubePosition/z";


        OSCPositionMessagex.Add(transform.position.x);
        OSCPositionMessagey.Add(transform.position.y);
        OSCPositionMessagez.Add(transform.position.z);



oscManager.GetComponent<ExampleOSCMessageSender>().AppendMessage(messageAddressX,
OSCPositionMessagex);

oscManager.GetComponent<ExampleOSCMessageSender>().AppendMessage(messageAddressY,
OSCPositionMessagey);

oscManager.GetComponent<ExampleOSCMessageSender>().AppendMessage(messageAddressZ,
OSCPositionMessagez);

    }
}
```

SphereAppear: Triggered Sphere to appear and move when trigger pressed

```
using System.Collections;
using System.Collections.Generic;
using UnityEngine;

public class SphereAppear : MonoBehaviour {

    public GameObject Sphere;
       // Use this for initialization
       void Start () {

       }

       // Update is called once per frame
       void Update () {
        //OVRInput.Get(OVRInput.Button.One) ||
        if ( OVRInput.Get(OVRInput.Axis1D.PrimaryIndexTrigger, 
OVRInput.Controller.RTouch) > 0)
        {
            Sphere.SetActive(true);

        }
        else

        {
            Sphere.SetActive(false);
        }
    }
}
```

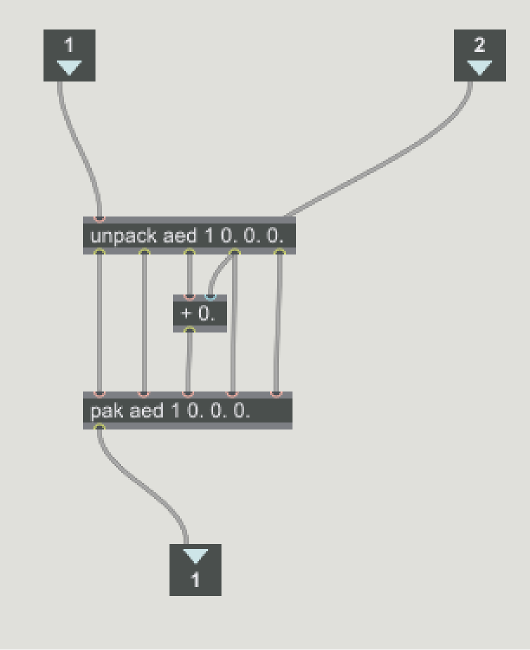
Appendix B: MaxMSP Sub-Patches

Clock System

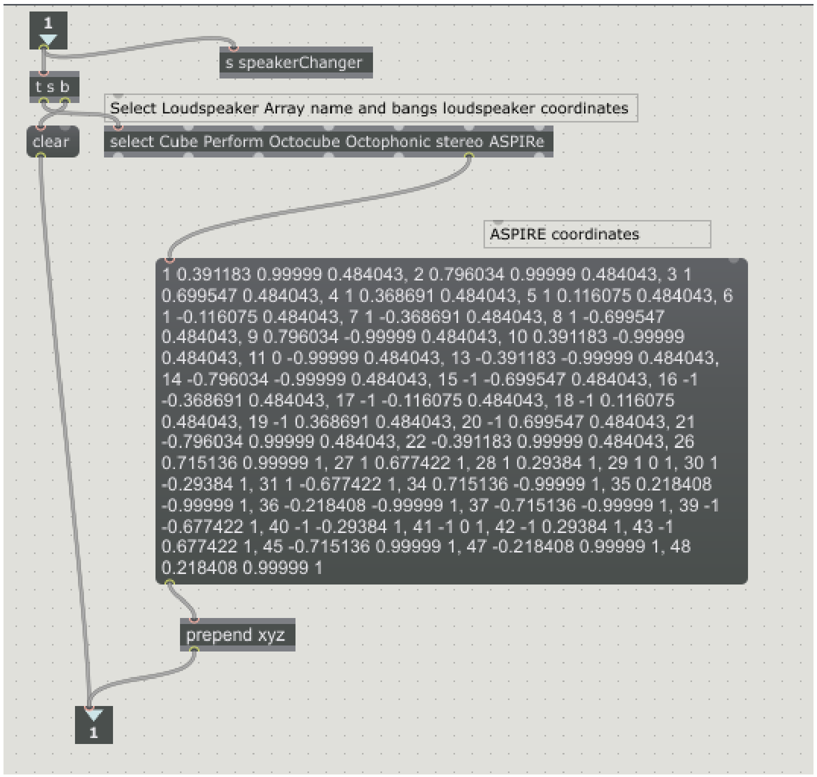


click on 'date' and 'ticks' to output

AudioLocationScaler

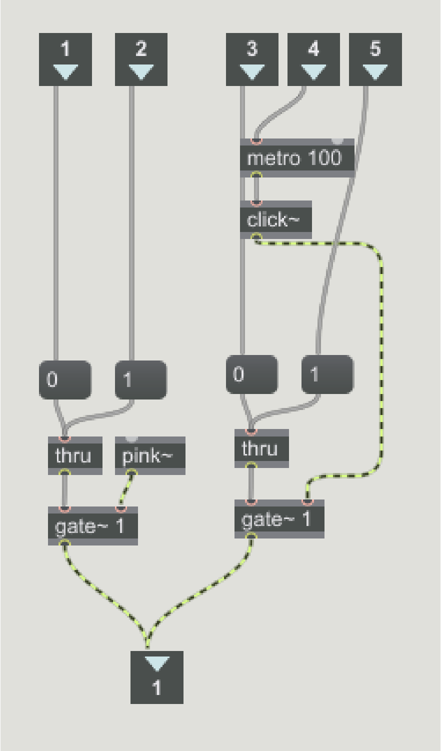


SetSpeakers

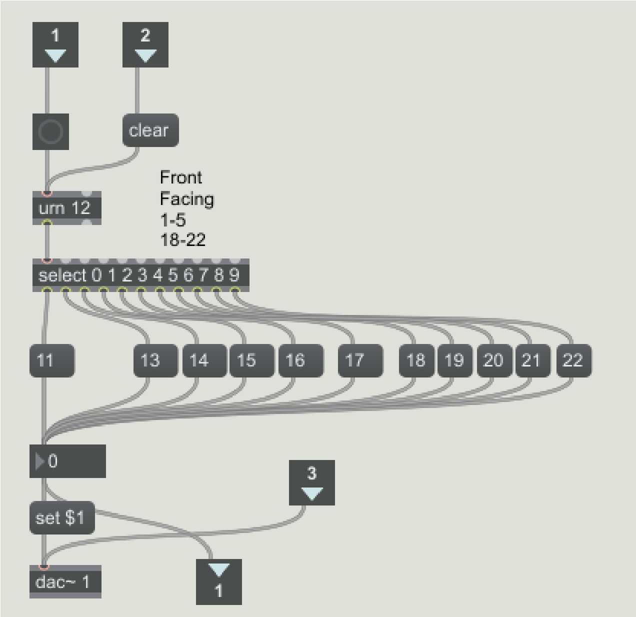

1
s speakerChanger
t s b
Select Loudspeaker Array name and bangs loudspeaker coordinates
clear
select Cube Perform Octocube Octophonic stereo ASPIRe
ASPIRE coordinates
prepend xyz
1

```
1 0.391183 0.99999 0.484043, 2 0.796034 0.99999 0.484043, 3 1
0.699547 0.484043, 4 1 0.368691 0.484043, 5 1 0.116075 0.484043, 6
1 -0.116075 0.484043, 7 1 -0.368691 0.484043, 8 1 -0.699547
0.484043, 9 0.796034 -0.99999 0.484043, 10 0.391183 -0.99999
0.484043, 11 0 -0.99999 0.484043, 13 -0.391183 -0.99999 0.484043,
14 -0.796034 -0.99999 0.484043, 15 -1 -0.699547 0.484043, 16 -1
-0.368691 0.484043, 17 -1 -0.116075 0.484043, 18 -1 0.116075
0.484043, 19 -1 0.368691 0.484043, 20 -1 0.699547 0.484043, 21
-0.796034 0.99999 0.484043, 22 -0.391183 0.99999 0.484043, 26
0.715136 0.99999 1, 27 1 0.677422 1, 28 1 0.29384 1, 29 1 0 1, 30 1
-0.29384 1, 31 1 -0.677422 1, 34 0.715136 -0.99999 1, 35 0.218408
-0.99999 1, 36 -0.218408 -0.99999 1, 37 -0.715136 -0.99999 1, 39 -1
-0.677422 1, 40 -1 -0.29384 1, 41 -1 0 1, 42 -1 0.29384 1, 43 -1
0.677422 1, 45 -0.715136 0.99999 1, 47 -0.218408 0.99999 1, 48
0.218408 0.99999 1
```

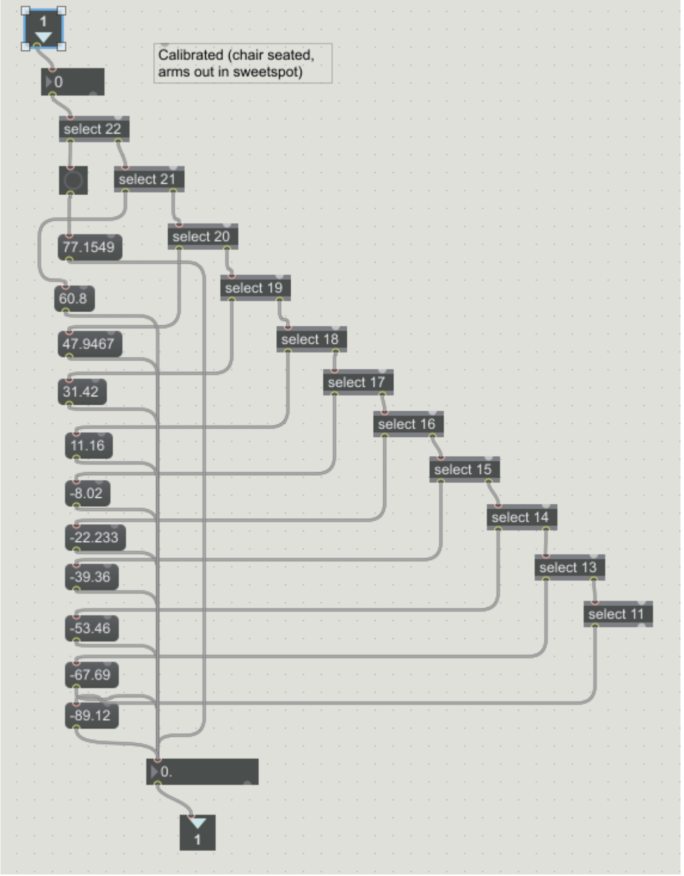
NoiseGates



BaselineRandomizer



SpeakerToAngle



Appendix C: Questionnaires

Participant #:
Date:

Background Questionnaire

Please help us to categorize our user population by completing the following items.

Gender (circle one): Male Female Other

Age: _____________

Occupation (if student, indicate graduate or undergraduate):

____

Major / Area of specialization (if student): _________________________________

Do you have normal or corrected to normal vision? Yes No

Rate how tired you are today: (circle one)

•----------------------•----------------------•----------------------•

very tired somewhat tired a little tired not tired at all

Rate your experience with Virtual Reality (VR): (circle one)

•----------------------•----------------------•----------------------•

beginner amateur intermediate advanced

Rate your experience with video games: (circle one)

•----------------------•----------------------•----------------------•

Never Sometimes Often Everyday

Participant #: ______________________

Date (mm/dd/yyyy): ________________

Audio and Visual Disparity in Virtual Reality

Post-Session Questionnaire

Please complete the following questions.

(1) How accurate do you feel about the task you just tried?
With 1 being not accurate estimate at all and 5 being very accurate

1 2 3 4 5

•-----------------------•-----------------------•-----------------------•-----------------------•

Not accurate at all Very accurate

(2) Please rate your sense of being in the virtual environment.
With 5 representing your normal experience of being in a place

1 2 3 4 5

•-----------------------•-----------------------•-----------------------•-----------------------•

(3) To what extent were there times during the experience when the virtual environment was the reality for you?

1 2 3 4 5

•-----------------------•-----------------------•-----------------------•-----------------------•

Artificial Reality

(4) During the time of your experience, did you often think to yourself that you were actually in the virtual environment

1 2 3 4 5

•-----------------------•-----------------------•-----------------------•-----------------------•

(5) How difficult did you think the task was?

1 2 3 4 5

•-----------------------•-----------------------•-----------------------•-----------------------•

Easy Very Difficult

Any other comments?

Questions addressed to participants during breaks: to check for motion sickness, nausea, or any other problems:

“I will now take your controller and then you can take off the headset.”

“Do you feel sickness, nausea, or any head pain?”

“Are your arms fatigued or are you experiencing any other negative sensations?”

“Take as much time as needed and let me know when you are ready for the next round.”

“Remember, you can stop participation at any time.”

“After you put the headset back on, I will hand you the controller and you can let me know when you are ready”