5-4-2019

Physical Controllers vs. Hand-And-Gesture Tracking: Evaluation of Control Schemes for VR Audio Mixing

Justin Bennington
justin@somewhere.systems

Physical Controllers vs. Hand-And-Gesture Tracking: Evaluation of Control Schemes for VR Audio Mixing

Master's thesis presented to the faculty of the
Audio Engineering Graduate Program
of
The Mike Curb College *of* Entertainment *&* Music Business
Belmont University, Nashville TN

In partial fulfillment of the requirements for the degree

Master of Science
with a major in
Audio Engineering

Justin Bennington

May 4, 2019

Advisors

Wesley A. Bulla
Doyuen Ko
Eric Tarr

ABSTRACT

Alternative control schemes for affecting the characteristics of audio signals have been designed and evaluated within the audio research community. The medium of virtual reality (VR) presents a unique method of sound source visualization using a headset which displays a virtual environment to the user, allowing users to directly control sound sources with minimal intermediary interference with a variety of different controllers. In order to provide insight into the design and evaluation of VR systems for audio mixing, the differences in subject preference between physical controllers and hand-and-gesture detection controls were investigated. A VR audio mixing interface was iteratively developed in order to facilitate a subject evaluation of some of the differences between these two control schemes. Ten subjects, recruited from a population of audio engineering technology undergraduate students, graduate students, and instructors, participated in a subjective audio mixing task. The results found that physical controllers outperformed the hand-and-gesture controls in each individual mean score of subject-perceived accuracy, efficiency, and satisfaction, with mixed statistical significance. No significant difference in task completion time for either control scheme was found. Additionally, the test participants largely preferred the physical controllers over the hand-and-gesture control scheme. There were no significant differences in the ability to make adjustments in general when comparing groups of more experienced and less experienced users. This study may provide useful contributing research to the wider field of audio engineering by providing insight into the design and evaluation of alternative audio mixing interfaces and further demonstrate the value of using VR to visualize and control sound sources in an articulated and convincing digital environment suitable for audio mixing tasks.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DEFINITIONS OF TERMS

Stage Metaphor: A system where the gain and stereophonic position parameters of each audio source are represented as an object in 2- or 3-dimensional space, positioned relative to the listener. Originally proposed by David Gibson as a "virtual mixer".

Channel-Strip Metaphor: A common design in audio mixing hardware and software, where the gain of a sound source is controlled by moving a sliding control to increase or decrease the level, and round knobs to determine the source's stereophonic position.

Virtual Reality (VR): Technology which uses a head-mounted display (HMD) to allow a user to view a fully-immersive stereoscopic image of a computer-generated three-dimensional world as well as interact with virtual objects using various control systems. Examples include the Oculus Rift™ and HTC VIVE™.

HMD: A head-mounted display which users wear to experience VR applications. It is comprised of sensors for tracking location and movement, and two stereoscopic angled displays to create a depth illusion.

Leap Motion Orion: A device and software which facilitates the recognition of hand gestures, actions such as grasping, and finger movements by way of an optical sensor, which can either be placed on a flat surface or attached to the front of a head-mounted display.

HTC VIVE™: A virtual reality headset created by HTC® and Valve Corporation®.

Unity: A game development engine developed by Unity Technologies® primarily used to create two-dimensional and 3D video games and simulations for various platforms.

DAW: An abbreviation for "Digital Audio Workstation". Software which is used to record, edit, and process audio files.

Signal: A representation of sound, either represented as a measurement of electrical voltage for analog signals, and a series of binary numbers for digital signals.

Track: An audio signal communications channel in a storage device or mixing console.

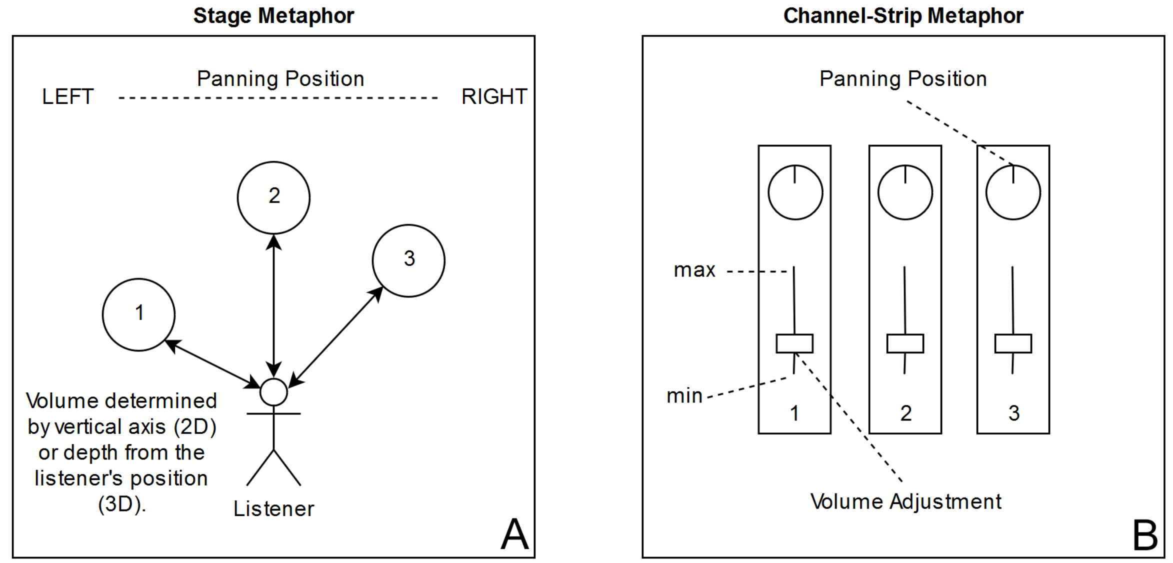
Channel: A single stream of recorded sound with a location in a sound field ("left front loudspeaker").

Virtual Instrument: A computer program or plug-in which generates and/or processes digital audio, most commonly for music.

1. INTRODUCTION

A variety of different control schemes are used for affecting the characteristics of audio signals. Signal adjustments in general afford engineers and musicians the ability to affect the sonic characteristics of each individual audio signal. Some of these characteristics include gain, timbre, and stereophonic position. Audio engineers have traditionally accomplished this by using physical and persistent buttons, knobs, and sliders. Since the 1920s, most audio mixing interfaces have followed the signal-flow metaphor, where the characteristics of a signal are adjusted in between the signal's input and the output of the adjusted signal. This metaphor has stood as one of the most persistent design paradigms in the age of recording technology [1].

Alternatives and changes to the persistent signal-flow metaphor within audio engineering have been periodically proposed over the history of recording. In order to make multiple adjustments at once using a single hand, primarily in order to reduce the amount of assistance needed when mixing live to disk or tape and later with multitrack, Tom Dowd replaced the Bakelite rotary knobs on the recording console at Atlantic Studios in New York with wire slide potentiometers [2]. With the advent of the digital audio workstation, attempting to improve the design of mixing interfaces by developing software which transcends the signal-flow metaphor has led to the exploration of alternative metaphors. One example is the stage metaphor, where the gain adjustment, stereophonic position, and other parameters of each audio source or track are represented as an object in 2- or 3-dimensional space positioned relative to the listener on a virtual "stage" [3]. A diagram comparing the stage and channel-strip metaphors' methods of controlling audio parameters is shown in Figure 1.



The location of objectsrelative to the listener determine the volume and panning of individual sound sources.

Faders control the volume of each channel, and their corresponding knobs control the panning position.

Figure 1. A diagram comparing the stage (A) and channel-strip (B) metaphors.

The effectiveness of some alternative control methods has previously been investigated by comparing them to popular or traditional methods of audio mixing. In a past study, when subjects were tasked with matching the gain balance and panning parameters of a reference mix, the stage metaphor had little difference in performance when compared to the channel-strip metaphor [4]. In one study, the stage metaphor out-performed the channel-strip metaphor when subjects were asked to identify visual information, not only in terms of correctly completed visual search and aural activity tasks, but additionally in overall subject preference [5]. More recently, interactive control schemes such as gesture tracking and motion control allow users to change the same characteristics with different hardware and software. Prior work has suggested that both gestural and traditional (mouse-and-keyboard) control schemes for audio software tend to suffer when they are not designed to be practical, responsive, intuitive, or able to control multiple parameters at once [6, 7].

The use of virtual reality (VR) systems for audio mixing offers another example of alternatives to adjusting characteristics of audio signals within a DAW, while providing both sound source visualization and control. VR is a medium which integrates a compelling illusion of a three-dimensional world with the control of interactive computer graphics. By wearing a head-mounted display (HMD), users are presented with an interactive environment which provides many of the relative spatial and positional cues present in the real world [8]. Drawing from prior designs of both stage metaphor and channel-strip metaphor audio mixing systems, VR has previously demonstrated its ability to visualize audio sources and give engineers the ability to directly control parameters related to them [9, 10]. By combining the stage metaphor with an immersive VR environment, users have the potential to directly interact with audio signals as if they are presented in front of them, represented in the form of three-dimensional objects.

Accessibility is another advantage to investigating the development and evaluation of new audio mixing tools for VR. Consumer VR systems like the HTC VIVE™, commonly used for video games, potentially serve as a new and affordable option for audio mixing interface control. Software designed for systems such as these have plenty of usage in both audio research and the consumer market. One example is DearVR™, a VR-based digital audio workstation [11]. Design schemes of controls which are integrated with other software within the VR medium, including the Leap Motion controller have also been evaluated in research related to gestural controls for audio mixing [12].

VR has been described by some audio researchers as a necessary medium for audio engineers to develop useful tools for, especially within the context of mixing audio for applications native to the medium [13]. Examples of systems which have combined gestural controls for audio and VR systems for audio are scarce. Investigating the differences in preference between various control schemes native to VR may allow for better user experience (UX) and offer information to aid the design of future audio mixing interfaces, especially those which bridge the gap between the physical and virtual world.

1.1 Research Questions and Objectives

The intent of this study was to investigate potential differences in subject preference and time-on-task between two different VR control schemes to perform a simple audio mixing task. The researcher additionally sought to investigate if the experience level of subject groups, when split between more and less experienced users, would show a difference in ratings or time to task completion. The conclusions from this research could influence how VR mixing systems are designed ergonomically.

User preference was measured using a survey of ratings of the subjects' perceived accuracy, efficiency, and satisfaction for both a hand-and-gesture tracking control system and a handheld controller-based system. Both control schemes affected single channels of a musical performance which were visualized within a VR audio mixing environment as three-dimensional objects. The time taken to achieve an "optimal mix balance" was recorded. The users' verbal feedback was recorded in order to provide information related to the user's experience with each control scheme. Each subject's preference of control scheme, or a lack of preference between either scheme was also recorded. The researcher's null hypothesis was as follows: "As measured by time-on task and a satisfaction-oriented survey, there will be no differences respectively in task completion time and subject-reported accuracy, satisfaction ratings or the overall preference between the hand-detection control and the physical controller systems."

2. PRIOR ART

Prior research related to gestural control systems for audio mixing generally involved systems which involved either directly moving sound sources in two or three dimensions (the stage metaphor) or interacting with elements of the channel-strip metaphor to change parameters. Most prior studies used gestural controls in conjunction with two-dimensional display methods such as a computer screen to visualize the changes being made to parameters of the audio mixing system. Other examples used the stage metaphor in combination with VR controllers and a headset to provide proof-of-concept for alternative audio mixing interfaces. It can be argued that gesture-controlled audio mixing interfaces and VR audio mixing interfaces intersect at the stage metaphor, as this has been a recurring theme in both approaches to designing useful alternatives.

Some evidence suggests that gestural controls provide an adequate method for audio mixing tasks, despite possessing much different ergonomics than using a keyboard and mouse. The performance of stage-metaphor based audio mixing interfaces, in comparison to channel-strip metaphor interfaces, has been previously investigated in order to explore if the stage metaphor can serve as an adequate alternative to traditional methods of control.

2.1 Research Comparing the Stage-Metaphor and Channel-Strip Metaphor

In one study, researchers found evidence supporting the claim that stage-metaphor based audio mixing systems and channel-strip based audio mixing systems have comparable performance [4]. Sound sources were visualized to the subject on a computer monitor (primarily for the channel-strip scheme) and a touchscreen tablet computer which was additionally used for controlling source selection within the stage-metaphor scheme. An audio mixing interface was developed to compare the channel-strip and stage-metaphor based control schemes for adjusting volume and panning of a single channel in a stereo mix. The researchers measured how accurately subjects were able to replicate the volume and panning of mixes, and how fast they could accomplish doing so.

Experiments were carried out over two days, with 15 total participants comprised of 7 experts and 8 novices. Each test took 15-20 minutes. Test participants filled in a questionnaire which recorded demographic information such as age and were then presented with the test procedure. The participants practiced the controls 1-3 times, which were a combination of an interface (Launchpad or iPad) and a piece of music (Drums or Guitar), until comfortable with the interface.

At the end of the practice period, the subjects were asked to listen to a prerecorded mix and carefully note the channel for which they were supposed to replicate volume and panning adjustments. After listening for 8 seconds, they turned to the interface, started the audio playback, and adjusted the gain (as volume) and stereophonic position (as panning) of the target signal to match that of the reference. When satisfied, the subjects stopped the audio playback, which registered their completion time and the volume and panning adjustment levels before advancing to the next trial. They were not informed that they were being timed. Each of the 15 participants provided 5 trials for each of the two interfaces, giving a total sample size of 75 for each interface. Scores were calculated as a difference between the reference and the selected values' MIDI signal.

The researchers did not find a significant difference in performance between the two interfaces for the purpose of the mixing task. The only statistically significant difference was found between novice users and expert users' ability to adjust panning in general. The authors stated that even though there was a lack of differences in this evaluation, there was not conclusive proof to argue that no differences exist. Participants were almost always in doubt of preference between which interface they felt performed the best; only a few decided on preference, and the decisions did not indicate a general tendency. The simplicity of the task may have garnered these results. However, the stage metaphor in this case was preferred for its intuitiveness, enjoyability, and its ability to allow users to better visualize spatial elements of a mix.

2.2 Gestural Audio Mixing Controllers in Prior Research

In 2013, a team of researchers designed the WAVE audio mixing interface, a system which used a camera and computer vision software to track movements and gestures for controlling a DAW in order to test subjects on their ability to use a few different control systems to mix eight different instrument tracks. In the study, the authors note that an overwhelming number of sound engineers claimed that mixing music using only a digital audio workstation in place of a recording console/desk made the music worse [14,15]. The authors argued the presence of differences between the algorithms of mixing software and their corresponding physical equivalents in analog mixing desks as being a potential contributing factor. A handful of audio engineers listed in the study claimed the subjective mix quality differences between consoles and digital audio workstations were due to their ergonomics.

The researchers designed a simple audio mixing graphical user interface (GUI) displayed on a projected screen. Adjustments were controlled by a dictionary of gesture controls, processed by a camera which used computer vision software to detect hand movements, gestures and position. Ten professional mixing engineers participated in the evaluation. Each engineer was tasked to mix eight audio tracks with significantly different musical and signal features. Each track contained recordings of a single instrument or of a group of instruments. The genre of the music was either instrumental rock or film soundtracks. The subjects were educated on a familiar system gesture dictionary. The engineers could use five different methods of sound mixing, enumerated on the next page:

1. Mixing using the custom GUI and gestures without parametric visual information displayed,
2. Mixing using the custom GUI and gestures with parametric visual information displayed,
3. Mixing using the custom GUI and a mouse with a keyboard, without parametric visual information displayed,
4. Mixing using the custom GUI and a mouse with a keyboard, and visual information reflecting parametric changes,
5. Mixing using a keyboard, mouse, and MIDI controller for parameter editing.

By using pairs of mixes generated by each method, information was provided to better understand factors such as the precision and accuracy of gestural controls compared to the other control methods. The engineers were asked to subjectively evaluate the recordings by way of a questionnaire.

The authors concluded that mixing audio signals using hand gestures in place of physical controllers such as a keyboard and mouse was not just viable, but intuitive. The intuitiveness, convenience, and precision of parametric editing of the gestural control system were rated on a scale of 1 to 5 by the engineers who participated in the evaluation. Six of the engineers rated the intuitiveness of the system a perfect score of 5, and the lowest score reported was 3. All the engineers greatly approved the possibility of controlling multiple parameters at once. The subjects primarily used this feature for controlling multiple parameters at once, such as the dynamic compression ratio and threshold at the same time. The subjects also showed a tendency to use it in other ways, such as affecting the amount of reverb present in the mix.

Some subjects reported that weariness resulted from insufficient ergonomics (one participant) and the running order of the mixing methods (two participants). The researchers concluded that the ability of engineers to be able to create mixes of equal aesthetic value with the gesture-controlled system as they would be able to create with a physical controller system was substantiated by the results of this study.

Building upon Lech & Kostek's work with the WAVE system, Ratcliffe's research aimed to use the existing framework for multi-parametric audio control existent in the channel strip to advance the mixing interface's design, while addressing both deficiencies in the existing paradigm and nuanced design opportunities. To accomplish this, the researcher utilized Gibson's stage metaphor to create an optical tracking and computer vision assisted mixing system [16].

The author argued that it could be useful to consider possible alternatives which do not model existent physical interfaces. Even though present user interface trends in the design of tools for adjusting audio parameters are based on physical interfaces, Ratcliffe argued a responsibility to "break away from skeuomorphistic design, and provide a more direct sense of control to the user". The Interaction Design Foundation explains skeuomorphism as a term which is used to describe when interface objects mimic their closest real-world counterpart in how they appear and how a user can interact with them [17]. A good example of skeuomorphistic design is the recycle bin icon used for deleting files on most operating systems.

To facilitate the study, Ratcliffe used the Leap Motion controller, an optical tracking device pictured in Figure 2, as well as a tablet with integrated control software to allow users to compare between different mixing methods. These methods were integrated into Ableton Live, a digital audio workstation.

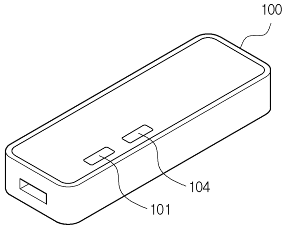


Figure 2. The sensor used in the study, adapted from [18].

Ratcliffe used virtual objects to represent individual sound sources and their position relative to the user to determine adjustment parameters such as gain and stereophonic position. Test participants were required to have experience with a DAW as well as some experience as a musical performer, musical producer, or audio engineer. Nine graduate students participated in the pilot study.

Participants were instructed to focus solely on two elements of the *MotionMix* system: volume level and panning. The individual channels' left and right panning values were controlled by the scaled x-position (left to right), while the volume (depth) of the source in the mix was controlled by z-position (depth) of the objects. The users could change the gain adjustment of a sound source by pushing it further back or pulling it closer in the mix, or move the source left and right to adjust the stereophonic position within a virtual sound stage. Three different interface designs were used by each participant for all tasks in the study: the *MotionMix* interface with no visual feedback, the *MotionMix* interface with the stage metaphor visual representation, and a virtual mixer within the DAW (the channel-strip metaphor). After being given the opportunity to become familiar with the controls and expressing that they were ready for the task, they proceeded to mix an eight-channel session. The duration of the practice period was not recorded.

In the trials, participants were instructed to mix the sources as if they were performing the mixing task for a client, had no time constraint, and to only stop once they were satisfied with the position of each sound source. The trials were timed, from the moment the participants first engaged with the system, to the moment they informed the author that they were satisfied with their mix. After the trials, subjects were instructed to answer questions regarding their preference between the interfaces, and to compare the use of the two variations of the *MotionMix* systems with the DAW's channel-strip metaphor controls, as well as other mixers they had used outside of the test. Several users reported that the system tracked their hands adequately, and most subjects stated they would adopt such a system into their workflows. One subject stated they would not be interested in integrating the system, and one subject left the question blank. 67% of the subjects preferred *MotionMix* with visualization, 22% preferred motion without visualization, and 11% preferred Ableton Live for the task.

Users did not find the *MotionMix* system to perform with significantly less accuracy than other mixers, and additionally did not find the *MotionMix* system with visualization to take substantially longer to complete the task as compares to the digital audio workstation. However, subjects spent more time with the *MotionMix* system without visualization to complete the subject evaluation task. The authors concluded that gestural control of a DAW could have many benefits, and future research should continue to evaluate gestural control systems while ensuring that the design of future studies keep gesture controls simple and easy to use.

2.3 Further Research in Gestural Controls for Audio Mixing Interfaces

Another example of using gestural controls for audio mixing explored the use of the Leap Motion sensor for use in an audio mixing interface, while providing a robust graphical user interface (GUI) to the user [20]. Designed in part to expand upon the functionality of Ratcliffe's *MotionMix* system, the LAMI system developed by Wakefield, Dewey and Gale provided a richer stage-metaphor based visual interface while using the Leap Motion controller. The system contained a variety of different controls based on a dictionary of hand gestures as well as a 3D graphical interface which displayed sound sources, the users' hands, and a variety of parameter adjustments such as EQ, effects, different modes and triggering playback. Subjects engaged in "a defined mixing task" against a benchmark mix in a DAW. The mixes were then judged by three experts on a scale of 0 to 10.

Some errors in the LAMI system at the time of testing detracted from the user experience. For example, an artefact in a smoothing algorithm caused the feature of being able to zero the

Auxiliary Send 2 channel to frustrate users and render the functionality of the feature unusable. The users also did not receive the idea of controlling many parameters at once. At times, the GUI would not follow the gestural actions of the user, which would detract from the user's experience. The conclusion to the study suggested that the multi-mapping of parameter controls to hand movements was incongruent with the test subjects' preference of controlling parameters individually. Many subjects mentioned the LAMI system as "time consuming", "stressful", or "hard to use", however most of the subjects did note the LAMI system as "fun". The study did not provide evidence to suggest the stage metaphor itself was cumbersome, but did provide evidence that gestural controls for audio mixing with many parametric mappings may detract from the users' experience.

2.4 VR as a Medium for Sound Source Visualization

Exploration of the stage metaphor for audio engineering tasks in combination with VR has simultaneously been demonstrated within audio engineering research. The VESPERS system provides an example of an audio system which bridges the concept of the stage metaphor and VR, and has been used for both facilitating listening tests as well as some creative applications. Developed by researchers at the SCENE Laboratory at Stevens Institute of Technology, VESPERS utilizes a 24.2 multichannel speaker array, a VR headset and controllers to represent sound sources as objects in a virtual environment [20]. Utilizing software powered by the Unity development engine, VESPERS serves as an example of 3D control schemes and 3D visualization being used together in a single system in order to provide an immersive method for users to interact with sound, including performing audio mixing tasks.

The design and evaluation new user interfaces which do not closely model traditional schema for audio mixing software is a complex task by nature. Thankfully, some researchers have suggested guidelines for providing the best possible tools for audio mixing tasks [21]. In one such example, the authors argued that audio equipment and software user interfaces which rely on traditional paradigms should be reconsidered in order to develop effective and simpler interfacing options in order to take advantage of the new tools developed for visual representation and interactive interfaces.

In this study, the researchers suggested a list of guidelines for the development of new user interfaces for audio: the determination of user skill level for the audio product, task analysis of the audio process to be performed by the system, creation of paper prototypes for exploration and task analysis before developing prototypes, selection of a metaphor which provides adequate information while simply representing the concept, and design of a user interface that allows the user to directly manipulate the visualization.

The researchers also created a list of requirements for evaluating new user interfaces: informal evaluation by consulting an expert user throughout the design process, simulation of a range of scenarios to refine and develop the prototypes into workable designs, and the evaluation of user interfaces based on different metrics. In this paper, the metrics suggested were efficiency by means of recording normalized task completion time, effectiveness in terms of accuracy in simple tasks completed to the subjects' satisfaction if not objective, and satisfaction level by way of user preference ranking and observing interaction and comments. The researchers argued that using many measures could provide contradictory results, and users can resist change if given interfaces which are radically different from the norm. Additionally, the researchers stated that radical redesigns of existing audio mixing paradigms may elicit poor ratings due to their individuality.

3. METHODS

3.1 Testing Software Design

The program used for the subject evaluation was developed with Unity, a software engine with a native physics engine commonly used for VR applications and video games, building upon prior VR audio mixing demonstrations such as the VESPERS system. A window displaying the user's view was presented on a monitor near the test administrator seated outside of the user's field of motion, in line with the controls for the program as part of Unity's editor window. An example layout of the system as it was placed during the evaluation is pictured below in Figure 3.

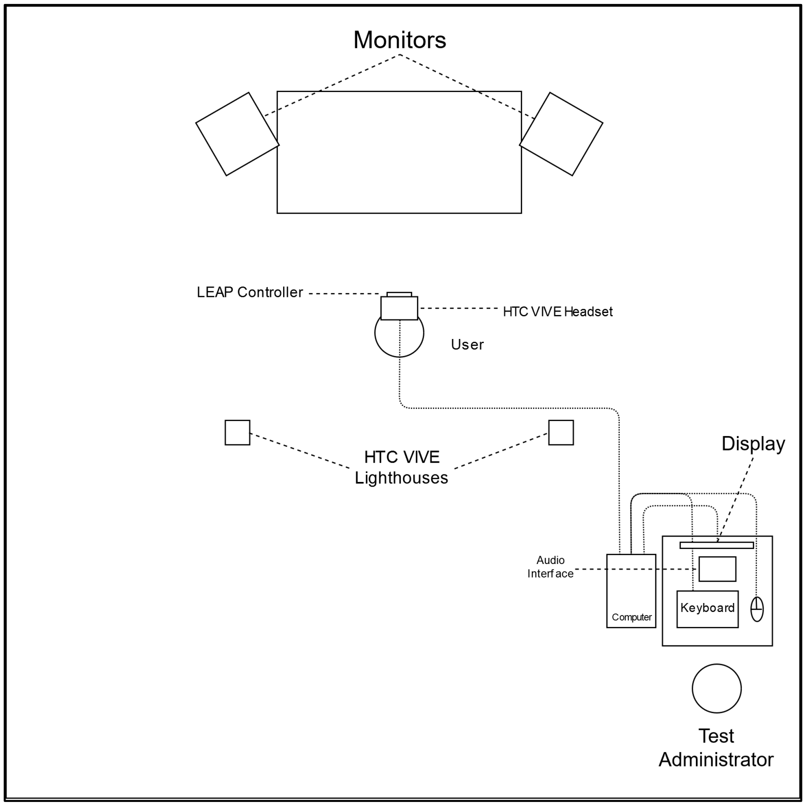


Figure 3. The testing system's physical footprint.

Two experts were consulted throughout the development process. One graduate instructor within the Audio Engineering Technology department at Belmont University participated in several iterations of pilot testing to ensure the software would be able to facilitate the subject evaluation consistently across multiple testing days.

The program was developed to function interchangeably between two control systems. The first system used two handheld controllers which were tracked by infrared-tracking base stations mounted on height-adjustable stands out of reach of the subject. A diagram of these controllers is provided below in Figure 4.

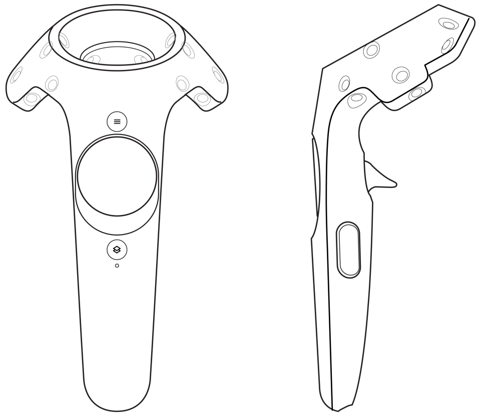
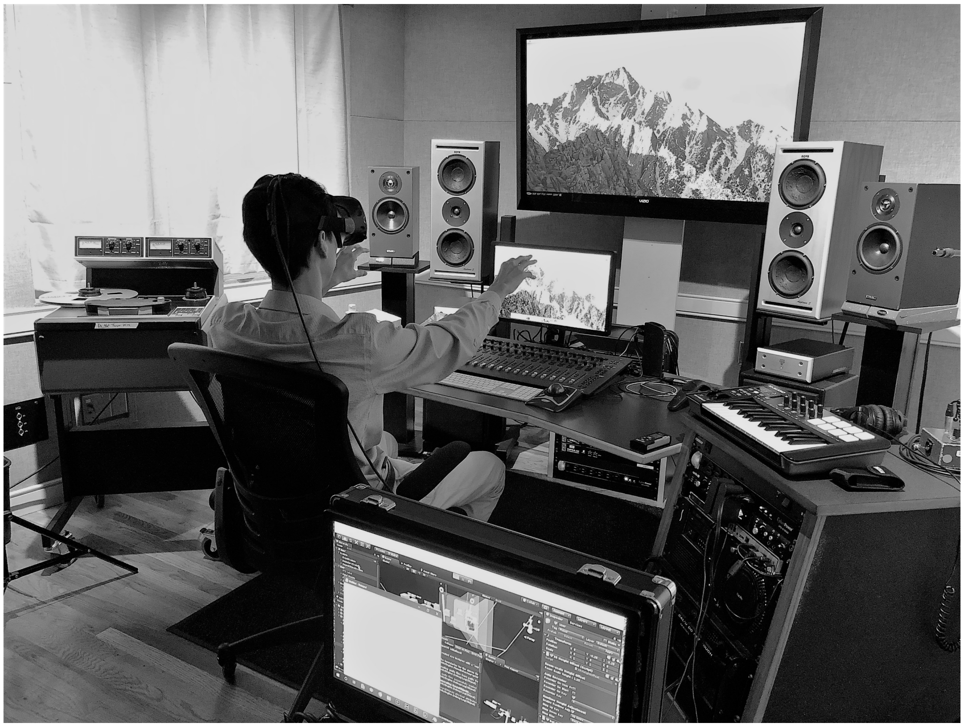


Figure 4. The HTC VIVE™ handheld controllers, adapted from [22].

The second system utilized the Leap Motion optical hand-detection and gesture-recognition controller, pictured earlier in this paper in section **2.4**. Both control methods were integrated with the Leap Motion Interaction Engine, as it provided an adequate method for both the physical controllers and hand-and-gesture controller to interact with virtual objects without any differences in latency and inconsequential differences in controller resolution [23]. The system allowed the user to interact with virtual sound objects as if they were physically presented in front of them, drawing inspiration from the GUI in Wakefield, Dewey, and Gale's LAMI system. The subjects could use any number of the controllers to push and pull individual sound sources, or they could grab and release objects either using the triggers on the physical controllers or natural grab-like hand gestures detected by the Leap Motion controller.

Virtual spheres could be moved in three dimensions to control the gain adjustment (depth, as a gain adjustment reading in decibels) and stereophonic position (left/right, as a panning percentage) of individual sound sources assigned to each sphere. This was designed to build upon the representational objects presented in the *MotionMix*, LAMI, and VESPERS systems [16, 19, 20]. The height of the objects was left unassigned to allow users to utilize vertical placement of the objects to prevent occlusion or crowding, allowing all eight objects to have identical gain adjustment and stereophonic parameters when placed in a vertical column. A demonstration of a participant using the hand-and-gesture control scheme to change the gain adjustment and stereophonic position of multiple sound sources is pictured on the next page in Photo 1.



Photograph 1. A participant using the Leap Motion controller to interact with sound objects.

When the scene was played, the program would create invisible colliders which were used to calculate variables for determining the stereophonic position and gain adjustment of each individual sound object. Two vertical planar collider provided minima and maxima, which were used to calculate gain adjustment, and were placed from the front edge of the detection radius of the optical controller (about 0.1 meters forward from the user's headset position) to a plane 0.5 meters from the front edge of the monitors. The left and right colliders, used to determine stereophonic position of the sound sources, were placed approximately at the center position of the left and right monitors. This prevented the user from placing virtual objects outside of their range of motion.

The software would first load the monophonic audio files into the program. Next, it would generate virtual representations of the monitoring system in the testing room, based on measurements of the distance of the user to the speakers, the distance between the speakers, and their height from the floor. A script would then procedurally generate several representational sound objects equal to the number of audio files in a resource folder, spacing them evenly across the distance spanning the left and right monitors and naming each object identical to the corresponding audio file's name. When users came into close contact with any sound object with their controls, a text display reading would appear so long as they were in range, recalculating the gain adjustment and stereophonic position every frame until the user stopped hovering over or interacting with the virtual object. The gain adjustment range, displayed as "vol", was calculated using linear scaling between a vertical planar collider 0.1 meters in front of the headset's initial position (0.0 dB) to the far vertical planar collider boundary about 0.5 meters away from the front edges of the monitors. The stereophonic position of each signal, displayed as "pan", ranged from 100% left to centered to 100% right, mimicking the design of common digital audio workstations. The gain adjustment and stereophonic position values were determined by the software engine's built-in audio system, using values of 0.0 to 1.0 for linear gain adjustment and -1.0 to 1.0 for panning following a constant-power panning law. A diagram of the boundaries is pictured below in Figure 5.

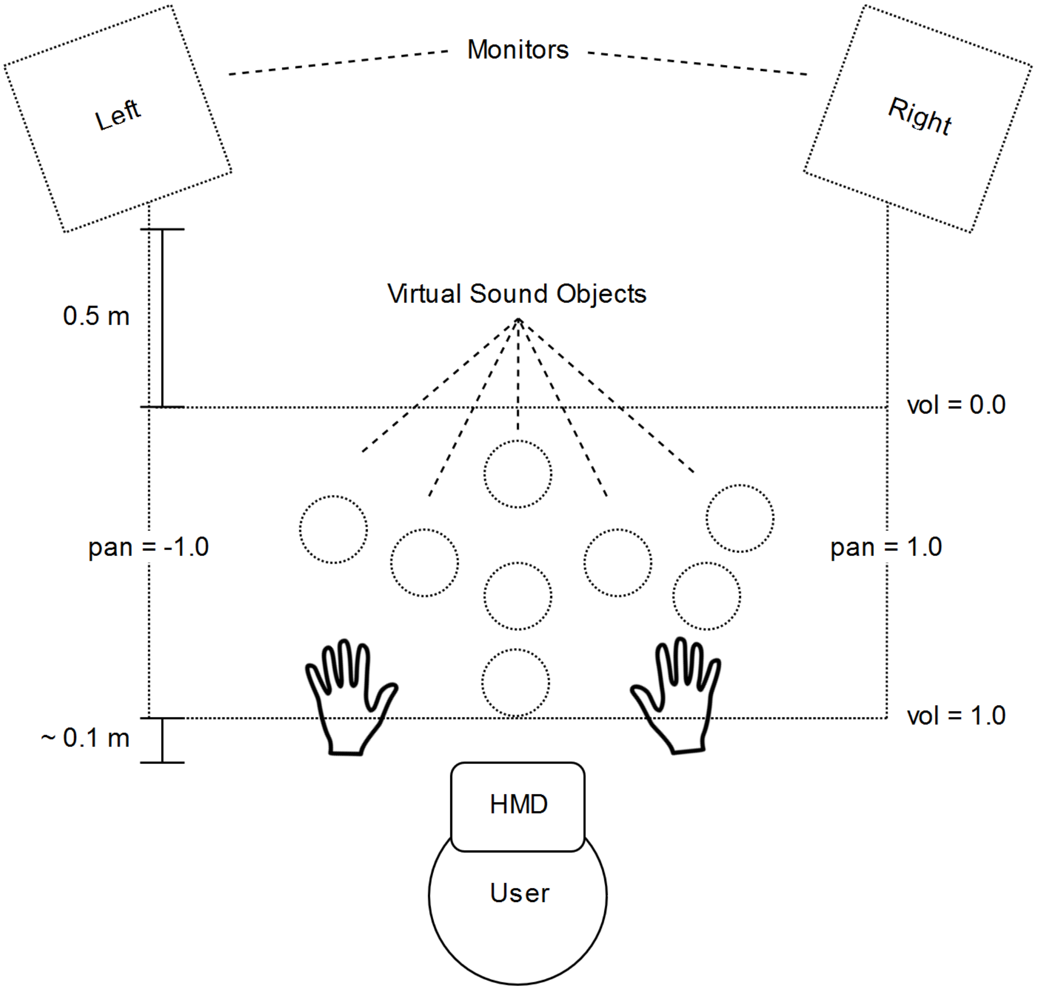


Figure 5. The boundary placement and limit values within the virtual reality environment for the control of audio sources, represented by virtual objects.

3.2 Research Environment

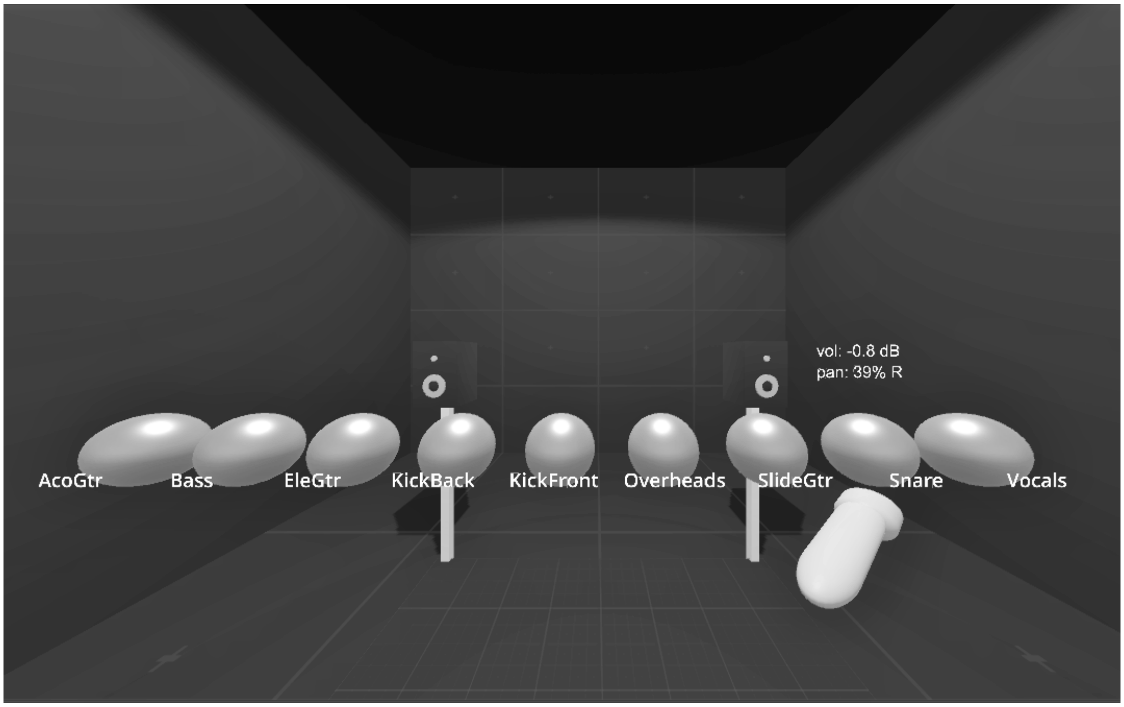
The testing was executed in a small recording studio commonly used for audio mixing and mastering. Two PMC® TB-1 nearfield studio monitors were aligned using a laser measure in an equilateral triangle between the speakers and the listening position [24]. A Genelec 7050B subwoofer was placed in position equal distance between the left and right monitors and the same distance from the listening position as the monitors [25]. The studio monitors and the subwoofer were time and level (74 dB, A-weighting) calibrated to ensure that the centered panning position was presented directly in the middle of the two monitors, and so that the listening level at unity gain was reasonable according to OSHA noise standard 29 CFR 1910.95 [26]. A Focusrite® 2i2 USB audio interface's outputs were plugged directly into the first and second inputs of a Universal Audio® 8P with the gain set to unity and calibrated to +4 dBu. The adjustment of the USB audio interface's master gain adjustment knob allowed the researcher to lower the listening level below the calibration level at the start of the test to between 60-65 dB (A-weighting), adjusting upon request of the subjects to match their preferred listening volume without exceeding the safe listening level. Before each round of testing, the virtual reality headset and both types of controllers were calibrated according to the recommendations by each manufacturer, and the base stations were placed in the same place during each trial to ensure that virtual object positioning was as consistent as possible across different testing days.

3.3 Stimuli

The stimuli for the test evaluation were comprised of eight mono 48kHz / 24bit audio tracks with minimal editing and processing and of similar average perceived loudness, derived from a free online resource for multitrack audio mixing practice and education. The tracks were derived from a 9-track recording of an acoustic rock song, with the right mono channel of the overhead drums removed from the test in order to convert the stereo overhead track to a mono overhead track. The total length of the song was approximately 3 minutes and 25 seconds, allowing the song to "loop" just under three times during each 10-minute evaluation period. The filenames of the audio samples were processed to provide simple object names which persistently appeared in front of each object over the duration of the test. Table 1 below shows the file name to object name conversion. A visual representation of what the subjects would see in their headset (with slight lens warping due to the difference between VR headset display and on-screen display) is also presented on the next page in Photo 2, demonstrating visualization of all eight sound sources and an example interaction with one of the sound sources using a physical controller.

Table 1. The stimuli presented in the evaluation.

Audio .wav File Name	Object Title
01_KickBack.wav	KickBack
02_KickFront.wav	KickFront
03_Snare.wav	Snare
04_Overheads.wav	Overheads
05_Bass.wav	Bass
06_AcGtr.wav	AcoGtr
07_ElecGtr.wav	EleGtr
08_SlideGtr.wav	SlideGtr
09_LeadVox.wav	Vocals



Photograph 2. A user hovering near a sound source in the virtual environment using the physical controllers.

3.4 Subjects

Two undergraduate students, four graduate students, and four instructors participated in the evaluation. The 10 participants were either enrolled as undergraduate/graduate students or an instructor within the audio engineering technology department, and were between 20 and 65 years old, with a mean age of 39.4 years. The mean audio engineering/mixing experience of the subjects was 21.2 years. The least experienced subject reported 2 years of experience and the most experienced subject reported 45 years of experience.

3.5 Experimental Procedure

Subjects were given the task of achieving their perception of an "optimal mix balance" between the sound sources using both randomly-ordered control schemes in two, ten-minute trials. The subjects were informed of the focus of the study and the duration of the trial period. Subjects were instructed to focus solely on the accuracy as it related to their perception of virtual objects responding properly to controller input, efficiency as a subjective measurement of the ease of control and quickness to achieve the intended result, and their overall satisfaction rating of each individual control scheme.

Before each ten-minute evaluation period, the placement and focus of the VR headset was adjusted to be comfortable for each subject. The subjects were given up to five minutes before each ten-minute trial period to familiarize themselves with the assigned control scheme. When the subject indicated they were ready, the testing period began. All sound sources would begin to simultaneously play, repeating after the song ended. Each subject was instructed to end the test once they believed they had achieved a satisfactory mix.

3.6 Survey Questions

After each ten-minute trial period had concluded, the subjects were asked to fill out a survey rating their perception of the accuracy, efficiency, and satisfaction related to both the volume and panning of their assigned control scheme. The first survey's questions and their corresponding response options are detailed respectively in Tables 2a and 2b.

Table 2a. The survey given to participants after each control scheme's trial period.

QUESTIONS
Which control scheme did you use?
On a scale of 1-10, how ACCURATE were the controls for adjusting volume?
On a scale of 1-10, how EFFICIENT were the controls for adjusting volume?
On a scale of 1-10, how SATISFYING were the controls for adjusting volume?
On a scale of 1-10, how ACCURATE were the controls for adjusting panning?
On a scale of 1-10, how EFFICIENT were the controls for adjusting panning?
On a scale of 1-10, how SATISFYING were the controls for adjusting panning?

Table 2b. The first survey's response choices, corresponding to the questions in Table 2a.

RESPONSES									
Hand-Detection Controls					Physical Controllers				
1 - Not accurate at all	2	3	4	5	6	7	8	9	10 - Very accurate
1 - Not efficient at all	2	3	4	5	6	7	8	9	10 - Very efficient
1 - Not satisfying at all	2	3	4	5	6	7	8	9	10 - Very satisfying
1 - Not accurate at all	2	3	4	5	6	7	8	9	10 - Very accurate
1 - Not efficient at all	2	3	4	5	6	7	8	9	10 - Very efficient
1 - Not satisfying at all	2	3	4	5	6	7	8	9	10 - Very satisfying

The subjects were given the option to revise their first survey's answers after the second ten-minute trial, and after both ten-minute trials and surveys were completed, they were given another additional question asking them for their overall preference between the two control schemes, or to indicate a lack of preference between either control scheme. The survey given to participants at the end of both trials is detailed in Table 3.

Table 3. The exit survey of overall control scheme preference.

Survey 2. Controller preference exit survey		
Q1. Which control scheme did you prefer to use?		
Hand-Detection Controls	Physical Controllers	I did not prefer any individual control scheme over the other.

A period in which subjects could provide verbal feedback about the control schemes was given, their responses were recorded, as well as the time they took in each ten-minute evaluation period from the start of the test until its termination. In total, subjects participated in one trial per control scheme. A total of 20 trials were performed.

4. RESULTS

4.1 Subject Response Differences Between Controllers, All Subjects

Tables 4a and 4b show the response means for both the hand controls and physical controllers for all subjects who participated in the evaluation. Although the physical controls seemed to show higher mean evaluation ratings, only some ratings were found to have statistically significant differences within the 95% confidence interval (p < .05).

Table 4a. Hand controller response ratings for all subjects.

HAND CONTROLS RESPONSE (ALL SUBJECTS)	MEAN	*SD*	*N*	*n*
Volume Accuracy	7.3	1.252	10	10
Volume Efficiency	6.2	1.751	10	10
Volume Satisfaction	6.4	2.633	10	10
Panning Accuracy	7.3	2.111	10	10
Panning Efficiency	6.6	2.319	10	10
Panning Satisfaction	7.3	2.669	10	10
Time Spent During Evaluation(s)	448.9	90.285	10	10

Table 4b. Physical controller response ratings for all subjects.

PHYSICAL CONTROLS RESPONSE (ALL SUBJECTS)	MEAN	*SD*	*N*	*n*
Volume Accuracy	8.6	0.966	10	10
Volume Efficiency	8.2	1.687	10	10
Volume Satisfaction	8.4	1.647	10	10
Panning Accuracy	8.3	1.252	10	10
Panning Efficiency	8.6	1.506	10	10
Panning Satisfaction	9.2	1.135	10	10
Time Spent During Evaluation(s)	448.5	122.616	10	10

Independent samples t-tests for differences performed comparing mean subject ratings and time spent to complete the task between the two control schemes showed statistically significant differences between several ratings. The ratings for volume accuracy (t = -2.600, df = 19, p = .018), the ratings for volume efficiency (t = -2.601, df = 19, p = .018), and the ratings for panning efficiency (*t* = -2.287, *df* = 19, *p* = .034) were found to be statistically significant when used to compare differences between the hand controls and the physical controllers. No significant difference in the time spent during each 10-minute evaluation period was found (*t* = 0.008, *df* = 19, *p* = 0.993). The results of the *t*-tests and a one-way ANOVA between the two control schemes is found on the next page in Table 5.

Table 5. Comparisons between controls (independent samples t-test, one-way ANOVA).

DIFFERENCES BETWEEN CONTROL SCHEME RESPONSES (ALL SUBJECTS)	*t*	*p*
Volume Accuracy	-2.600	.018
Volume Efficiency	-2.601	.018
Volume Satisfaction	-2.037	.057
Panning Accuracy	-1.289	.214
Panning Efficiency	-2.287	.034
Panning Satisfaction	-2.072	.053
Time Spent During Evaluation (s)	0.008	.993

A comparison of means between the six individual ratings provided by all subjects is displayed below in Figure 6. The physical controller scheme scored higher, on average, in each categorical rating than the hand-controlled system, however not all differences were found to be statistically significant.

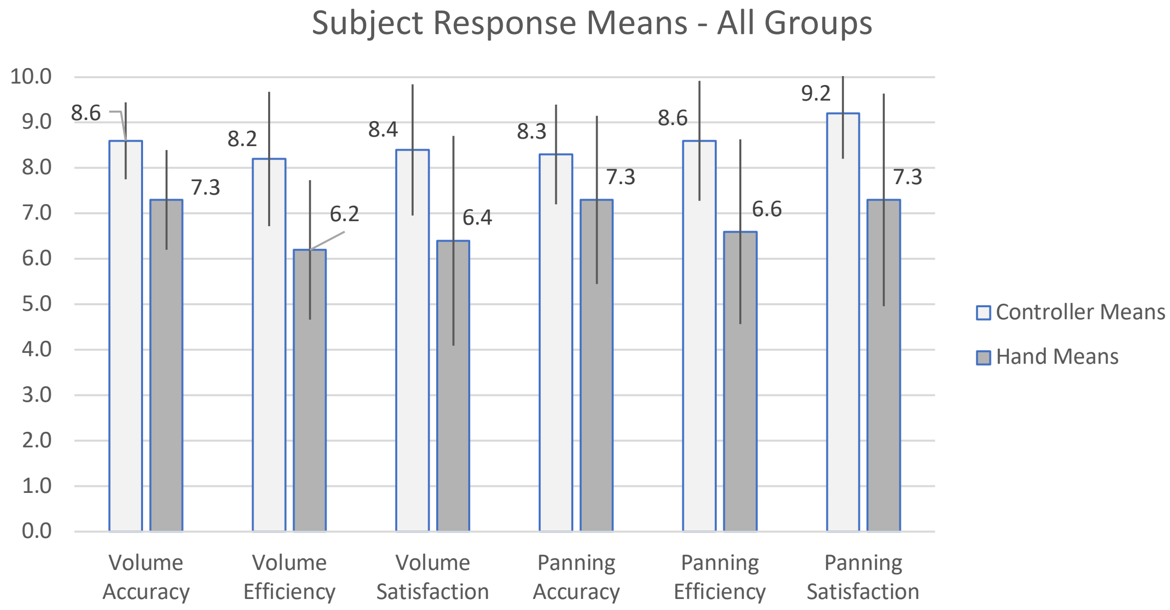


Figure 6. Subject response means & 95% confidence interval between all groups.

4.2 Differences Between Inexperienced & Experienced Subjects

Differences between two groups of respondents, those with under 10 years reported experience in audio engineering/mixing and those with over 10 years reported experience, were further investigated. The response means for the hand controls showed no statistically significant difference in reported rating for any category, and no statistically significant difference for the time spent during the evaluation. However, the Volume Accuracy *p*-value (.073) was found to approach the significance level ($p < .05$), indicating a difference in the Volume Accuracy ratings provided between experience groups for the hand-and-gesture controls. The response means and repeated analysis procedure for the hand controls are shown in tables 6, 7, and 8.

Table 6. Hand controls subject responses, subject experience under 10 years.

HAND CONTROLS RESPONSE (Experience < 10 years)	MEAN	*SD*	*N*	*n*
Volume Adjustment Accuracy	8.0	1.000	5	5
Volume Adjustment Efficiency	6.0	2.121	5	5
Volume Satisfaction	6.8	2.049	5	5
Panning Accuracy	8.2	0.837	5	5
Panning Efficiency	6.6	2.191	5	5
Panning Satisfaction	8.2	1.304	5	5
Time Spent During Evaluation (s)	444.8	101.758	5	5

Table 7. Hand controls subject responses, subject experience above 10 years.

HAND CONTROLS RESPONSE (Experience > 10 years)	MEAN	*SD*	*N*	*n*
Volume Accuracy	6.6	1.140	5	5
Volume Efficiency	6.4	1.517	5	5
Volume Satisfaction	6.0	3.317	5	5
Panning Accuracy	6.4	2.702	5	5
Panning Efficiency	6.6	2.702	5	5
Panning Satisfaction	6.4	3.507	5	5
Time Spent During Evaluation (s)	453	89.129	5	5

Table 8. Comparisons between experience groups (independent samples t-test, one-way ANOVA) when using hand controls.

GROUP (HAND CONTROLS, ALL SUBJECTS)	*t*	*p*
Volume Accuracy	2.064	**.073**
Volume Efficiency	-0.343	.740
Volume Satisfaction	0.459	.659
Panning Accuracy	1.423	.193
Panning Efficiency	0.000	1.000
Panning Satisfaction	1.076	.313
Time Spent During Evaluation (s)	-0.136	.896

Additional analysis was performed to investigate if there were statistically significant differences between the ratings between both experience groups for the physical controller scheme. No statistically significant differences were found between any subject rating category, as well as the time spent during the evaluation period. However, the Panning Accuracy *p*-value

(0.073) was found to approach the significance level (p < .05), indicating a difference in the Panning Accuracy ratings provided between experience groups for the physical controllers. The mean and analysis displayed below in tables 9, 10, and 11.

Table 9. Physical controls subject responses, subject experience under 10 years.

PHYSICAL CONTROLS RESPONSE (Experience < 10 years)	MEAN	*SD*	*N*	*n*
Volume Accuracy	9.0	1.000	5	5
Volume Efficiency	8.6	2.191	5	5
Volume Satisfaction	8.2	1.924	5	5
Panning Accuracy	9.0	1.000	5	5
Panning Efficiency	9.0	1.732	5	5
Panning Satisfaction	9.4	0.894	5	5
Time Spent During Evaluation (s)	452.2	113.438	5	5

Table 10. Physical controls subject responses, subject experience above 10 years.

PHYSICAL CONTROLS RESPONSE (Experience > 10 years)	MEAN	*SD*	*N*	*n*
Volume Accuracy	8.2	0.837	5	5
Volume Efficiency	7.8	1.095	5	5
Volume Satisfaction	8.6	1.517	5	5
Panning Accuracy	7.6	1.140	5	5
Panning Efficiency	8.2	1.304	5	5
Panning Satisfaction	9.0	1.414	5	5
Time Spent During Evaluation (s)	444.8	144.657	5	5

Table 11. Comparisons between experience groups (independent samples t-test, one-way ANOVA), physical controls.

GROUP (PHYSICAL CONTROLS, ALL SUBJECTS)	*t*	*p*
Volume Accuracy	1.372	.207
Volume Efficiency	0.730	.486
Volume Satisfaction	-0.365	.724
Panning Accuracy	2.064	**.073**
Panning Efficiency	0.825	.433
Panning Satisfaction	0.535	.608
Time Spent During Evaluation (s)	0.090	.930

The groups comparison of means was visualized in a bar chart in figures 7 and 8 on the next page, displaying both the means and 95% confidence intervals for the subject ratings split into the under 10 years' experience group and the over 10 years' experience group.

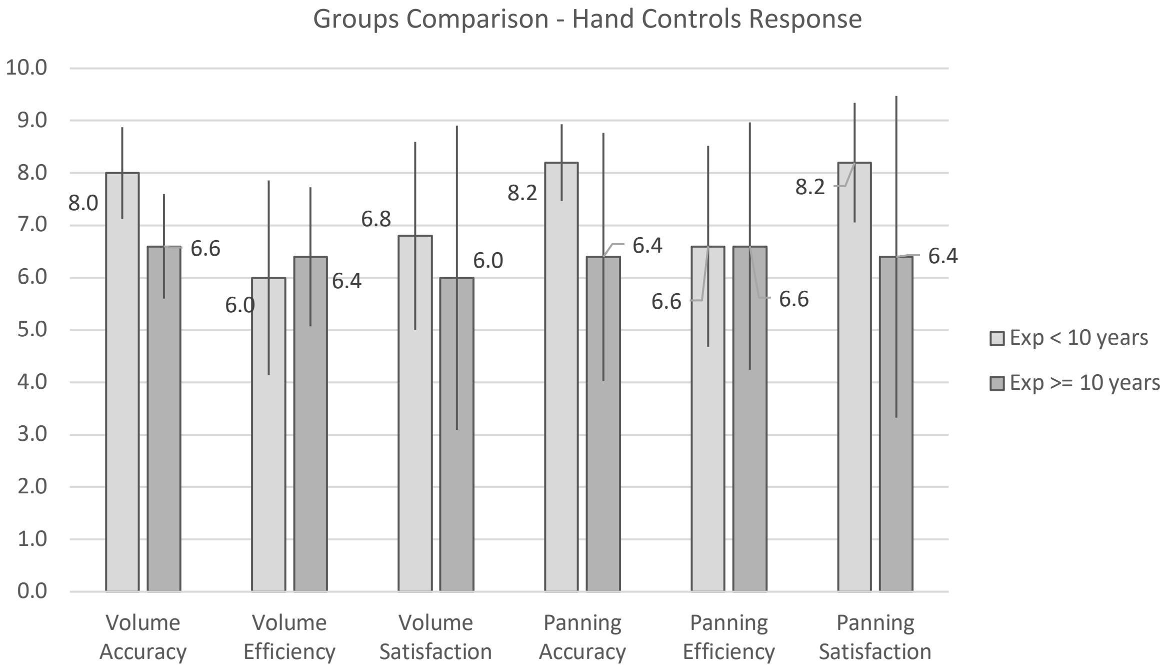


Figure 7. Comparison of means between less and more experienced groups' responses for the hand controls, with 95% CI.

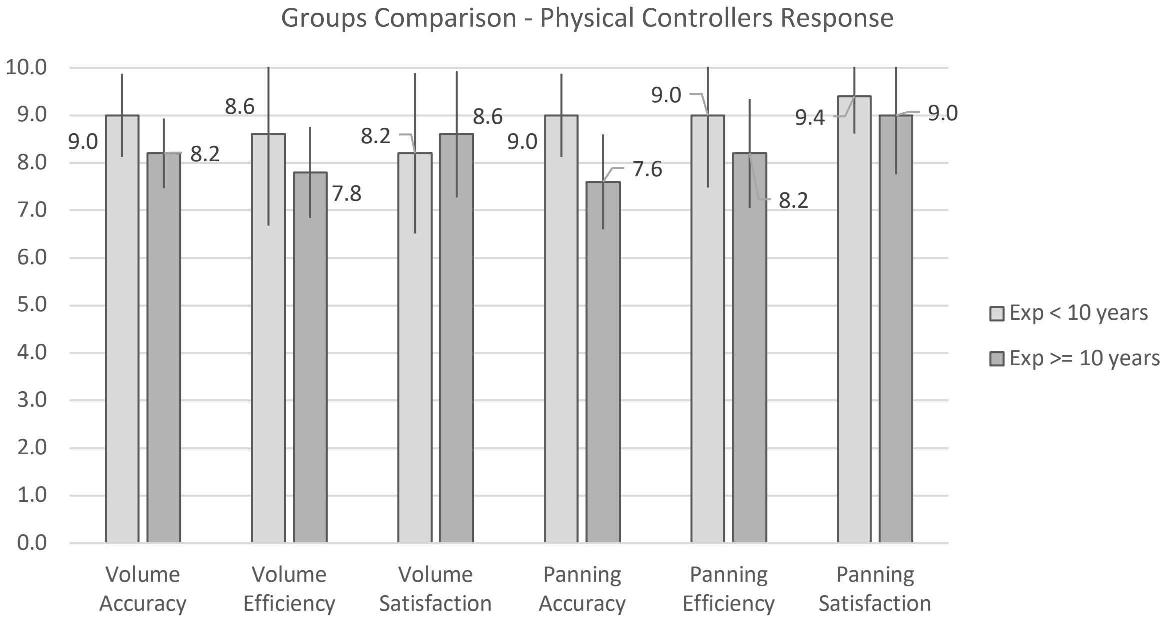


Figure 8. Comparison of means between less and more experienced groups' responses for the physical controllers, with 95% CI

5. DISCUSSION

5.1 Quantitative Results

Mean ratings for the physical controllers were slightly higher than those of the hand-and-gesture controls in every category amongst the total participant group. Additionally, these results were repeated when the sample group was split into more experienced (exp. >= 10 yrs.) and less experienced (exp. < 10 yrs.) users, with both groups reporting higher mean ratings for the physical controllers' than for the hand-and-gesture controls. Several differences garnered statistically significant results. It was not possible to test for differences between control schemes for the experience-split groups due to the small sample size (*n* = 5).

In the data set for all subject responses, the mean subject ratings between the physical controllers and hand-and-gesture controls for the categories of Volume Accuracy (*p* = **.018**), Volume Efficiency (*p* = **.018**), and Panning Efficiency (*p* = **.034**) were found to be statistically significant (*p* < .05). Additionally, the p-values for Volume Satisfaction and Panning Satisfaction (*p* = .057, *p* = .053 respectively) closely approached the threshold of statistical significance. The *p*-value for Panning Accuracy was not found to be statistically significant (*p* = .214).

When compared to the hand-and-gesture controls, the higher average ratings for the physical controllers for all subjects and the experience-split groups can be reflected in the results of the final preference survey in figure 9. Seven subjects preferred the physical control scheme, one subject preferred the hand-and-gesture controls, and two subjects did not maintain any preference for one scheme over the other for the evaluation task.

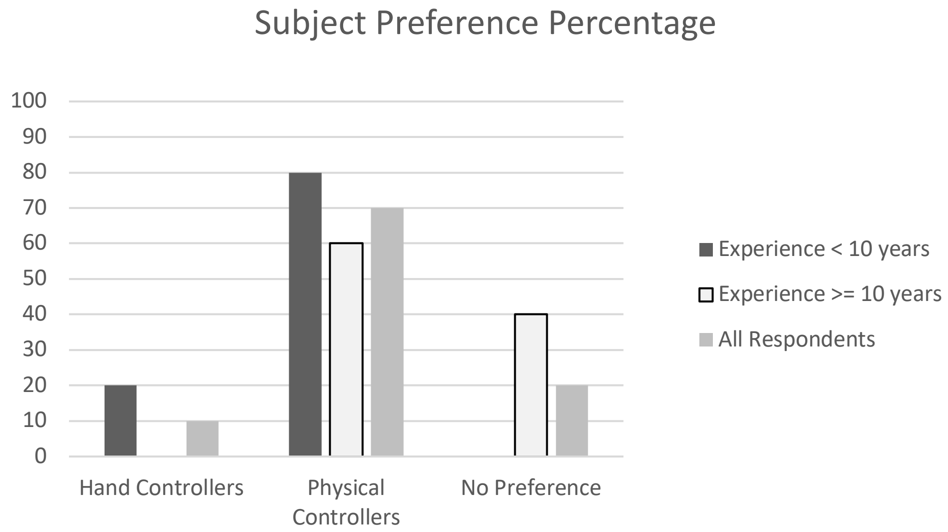


Figure 9. The percentage of subjects' preference between the control schemes.

There was not any statistically significant finding of difference, nor any indication towards a trend in the difference between time-on-task when comparing any of the control schemes or experience groups to each other.

When the sample group was split into two sets of five subjects, one with 10 or more years of experience and one with less than 10 years of experience, both groups were found to rate the physical controllers higher on average than the hand-and-gesture controls. The researcher set out to additionally investigate if there were significant differences in the ratings when compared between the experience groups. None of the categories met the threshold of statistical significance ($p < .05$), however, the p-values for both Volume Accuracy for the hand-and-gesture controls ($p = .073$) and Panning Accuracy for the physical controls ($p = .073$) showed a difference between the two experience groups which neared the threshold of significance.

5.2 Subject Verbal Response

Many subjects mentioned the practicality of being able to directly interact with sound channels using their hands. All subjects expressed that they would like to see more features in future iterations of the testing program. When asked about their reasoning behind the preference of one scheme over the other, the most common reason reported was an improved responsiveness of the physical controls in comparison to their hand-and-gesture counterparts. Some of the test participants mentioned that the physical controllers' triggers being used to "drag and drop" sound objects were more effective than the grabbing gestural detection provided by the hand-and-gesture controls. A few subjects mentioned problems with responsiveness of the hand-and-gesture controls, namely in their ability to grab objects without accidentally colliding with other objects they had already put into place. A few subjects elaborated further, suggesting a form of "focus" or "locking" mechanic for sound sources they had finished placing in the scene. A few of the test subjects expressed that they would like to use this system to mix their own records.

5.3 Comparison to Prior Research

In the past, research related to hand or gestural controls for audio applications has primarily focused on comparing them to schemes such as a keyboard and mouse or a MIDI controller, utilizing traditional visualization methods like a screen, and with findings generally indicating user preference of hand-and-gesture controls. The data gathered here presents some evidence that physical controllers were preferred by the subjects who participated in this study for mixing audio within a virtual reality soundstage more than the hand-and-gesture control system.

6. CONCLUSIONS

Physical controls were preferred over optical hand-and-gesture detection controls for the purpose of mixing multichannel audio within a VR representation of the stage metaphor, and the subject ratings of each control scheme possessed some statistically significant differences which reflected this preference. The study gathered evidence supporting the subjects of the study, who largely preferred physical controllers over hand-and-gesture detection-based controls when interacting with objects in a basic VR audio mixing environment. Despite both control schemes having differences in subject-reported efficiency, the task to completion time between the two schemes did not possess a great enough difference to be deemed significant.

Physical controllers scored higher than the hand-and-gesture controls in every single individual category: mean accuracy, panning, and satisfaction ratings for both volume and panning were higher for the physical control scheme than the hand-and-detection control scheme. Even with a small sample size, many individual differences between these interfaces were found to be statistically significant, and nearly all other differences closely approached the threshold of statistical significance. Nearly all subjects maintained a preference for the physical controls, describing them as more perceivably accurate, more efficient, and more satisfying than their hand-and-gesture counterpart, although a few subjects reported isolated moments of frustration with the testing software itself. There was not found to be any difference in the times recorded for the subjects to complete the evaluation task. The researcher concludes from this and the verbal feedback that the lack of difference in completion times may have been due to the novelty of mixing in virtual reality for many of the subjects, and the lack of difference in tracking latency between the two schemes while used in the program.

Even though more experienced subjects tended to rate individual metrics lower on average than the less experienced subject group, none of these differences were found to be statistically significant. However, it was found that even when split into two different experience groups, subjects still preferred the physical controllers over the hand-and-gesture controls and rated the individual categories for physical controls higher than their counterpart.

The researcher concludes: there were some significant differences in subject-reported accuracy, satisfaction ratings and the overall preference between the hand-detection controls and the physical controller systems. The researcher partially rejects the null hypothesis: this study provides evidence in support of some differences in preference, subject-reported accuracy, efficiency, and satisfaction ratings between the two control schemes evaluated in this study. Most subjects preferred physical controller systems for the experimental task.

Further Research

Further research into the differences between control schemes for VR-based stage metaphor audio mixers could include a more thorough investigation into some differences between more complex control mappings between two sets of physical controllers. Including more subjects to allow for additional analyses will allow for more conclusive evidence. Additionally, the design and testing of a hybrid between the two control schemes used in this study, such as a glove which might provide both hand-and-gesture tracking as well as more responsive drag-and-drop functionality of the physical controllers and may solve some of the functionality issues present in the schemes tested over the course of this study. If the test were to be repeated, creating a system for detailed user interaction logging may be useful for gaining more insight into the ways that users interact with control schemes for virtual reality multichannel audio mixing. The inclusion of time-on-task as a measure may have been irrelevant to the study due to the open-ended nature of the mixing task. A more focused study, such as tasking users to match values of individual channels to a reference, may be worth investigating.

It may also be useful to the scope of virtual reality audio workstation design to compare these control methods to other ways of mixing multichannel audio, performing a broader test comparing methods such as mixing on a console, or changing software parameters with a keyboard and mouse, to a virtual reality soundstage utilizing physical controllers and/or hand-and-gesture controls. As some researchers have used VR and gestural controllers together to creatively augment physical instruments such as a keyboard, there is plenty of exploration to be done in the realm of new and innovative control schemes for audio, whether creative or corrective adjustments are to be made [28].

The researcher plans to continue to develop features for the test program used in this study based on the feedback provided by the subjects in the evaluation, and has provided the software open-source under the MIT License for use in any future projects in order to help facilitate the advancement of audio engineering practice and research.

BIBLIOGRAPHY

Citations

[1] M. Walther-Hansen, "New and Old User Interface Metaphors In Music Production." Journal on the Art of Record Production (JARP 2017). Issue 11, (2017) DOI: http://www.arpjournal.com/asarpwp/content/issue-11/

[2] D. Daley, "The Engineers Who Changed Recording: Fathers Of Invention." Sound On Sound (Oct 2004) DOI: https://www.soundonsound.com/people/engineers-who-changed-recording

[3] D. Gibson, "The Art Of Mixing: A Visual Guide To Recording, Engineering, And Production." ArtistPro Press. (1997)

[4] S. Gelineck, D. Korsgaard and M. Büchert, "Stage- vs. Channel-strip Metaphor - Comparing Performance when Adjusting Volume and Panning of a Single Channel in a Stereo Mix." Proceedings of the International Conference on New Interfaces for Musical Expression (NIME 2015). pp. 343-346, (2015)

[5] J. Mycroft, T. Stockman and J. Reiss, "Visual Information Search in Digital Audio Workstations." Presented at the 140th AES Convention, Convention Paper 9510. (May 2016)

[6] R. Selfridge and J. Reiss, "Interactive Mixing Using Wii Controller." Presented at the 130th AES Convention, Convention Paper 8396. (May 2011)

[7] M. Lech and B. Kostek, "Testing a Novel Gesture-Based Mixing Interface." J. Audio Eng. Soc., Vol. 61, No. 5, pp. 301-313. (May 2013)

[8] S. Bryson, "Virtual Reality in Scientific Visualization." Communications of the ACM, Vol. 39, No. 5, pp. 62-71. (May 1996)

[9] A. Kuzminski, "These Fascinating New Tools Let You Do 3D Sound Mixing – Directly In VR." A Sound Effect. (Aug 2018) DOI: https://www.asoundeffect.com/vr-3d-sound-mixing/

[10] T. Mäki-Patola, J. Laitinen, A. Kanerva and T. Takala, "Experiments with virtual reality instruments." Proceedings of the International Conference on New Interfaces for Musical Expression (NIME 2005). pp. 11-16. (May 2005)

[11] "DearVR," DearVR. Retrieved from Web. DOI: http://dearvr.com/

[12] J. Kelly and D. Quiroz, "The Mixing Glove and Leap Motion Controller: Exploratory Research and Development of Gesture Controllers for Audio Mixing." Presented at the 142nd AES Convention, Convention e-Brief 314. (May 2017)

[13] F. Rumsey, "Virtual reality: Mixing, rendering, believability." J. Audio Eng. Soc., Vol. 64, No. 12, pp. 1073-1077. (Dec 2012)

[14] R. Campbell, "Behind the Gear," Tape Op – The Creative Music Recording Magazine, No. 81, pp. 12-13. (Mar 2011)

[15] B. Owsinski, "The Mixing Engineer's Handbook: Second Edition." Boston: Thomson Course Technology PTR. (2006)

[16] J. Ratcliffe, "MotionMix: A Gestural Audio Mixing Controller." Presented at the 137th AES Convention, Convention Paper 9215. (Oct 2014)

[17] K. Göttling, "What is Skeuomorphism?" The Interaction Design Foundation. (2018)

[18] M. Young-Lae and C. Yong-Chul, "Virtual arthroscopic surgery system using Leap Motion." Korean Patent KR101872006B1 issued June 27, 2018. DOI:

https://patentimages.storage.googleapis.com/4c/8d/85/55932cf18e50d9/112017000213110-pat00001.png

[19] J. Wakefield, C. Dewey and W. Gale, "LAMI: A Gesturally Controlled Three-Dimensional Stage Leap (Motion-Based) Audio Mixing Interface." Presented at the 142nd AES Convention, Convention Paper 9785. (May 2017)

[20] R. Graham and S. Cluett, "The Soundfield as Sound Object: Virtual Reality Environments as a Three-Dimensional Canvas for Music Composition." Presented at the Conference on Audio for Virtual and Augmented Reality, No. 7-3. (Sep/Oct 2016)

[21] C. Dewey and J. Wakefield, "A Guide to the Design and Evaluation of New User Interfaces for the Audio Industry." Presented at the 136th AES Convention, Convention Paper 9071. (Apr 2014)

[22] "About the VIVE™ Controllers." HTC Corporation. (2019) DOI: https://www.vive.com/media/filer_public/17/5d/175d4252-dde3-49a2-aa86-c0b05ab4d445/guid-2d5454b7-1225-449c-b5e5-50a5ea4184d6-web.png

[23] "Interaction Engine 1.2.0." Leap Motion. (Jun 2018) DOI: https://developer.leapmotion.com/releases/interaction-engine-120

[24] "TB1." The Professional Monitor Company Ltd. (2019) DOI: https://pmc-speakers.com/products/archive/archive/tb1

[25] "Genelec 7050B Studio Subwoofer." GENELEC (2018) DOI: https://www.genelec.com/studio-monitors/7000-series-studio-subwoofers/7050b-studio-subwoofer

[26] "OSHA Noise Regulations (Standards-29 CFR): Occupational noise exposure.-1910.95." Occupational Safety and Health Administration, Appendix E – Acoustic Calibration of Audiometers. OSHA, Vol. 1, No. 9, p. 9. (1996)

[27] "Leap Motion Orion." Leap Motion. (Jun 2018) DOI: https://developer.leapmotion.com/orion/

[28] J. Desnoyers-Stewart, D. Gerhard, and M.L. Smith, "Augmenting a MIDI Keyboard Using Virtual Interfaces." J. Audio Eng. Soc., Vol 66, No. 6, pp. 439-447. (Jun 2018)

Resources

"Channel (audio) – Glossary" Federal Agencies Digitization Guidelines Initiative. (n.d.) Retrieved from Web site: http://www.digitizationguidelines.gov:8081/term.php?term=channelaudio

"HTC Vive," Wikipedia. (n.d.) Retrieved from Web site: https://en.wikipedia.org/wiki/HTC_Vive

J. F. Hair, R. E. Anderson, R. L. Tatham, and W. C. Black, "Multivariate Data Analysis", 5th ed. (1998)

"Unity User Manual (2018.3)." Unity Technologies. (2018)

"GitHub: Tactile Mix." Justin Bennington. (2018) Retrieved from Web Site: https://github.com/justin-bennington/tactile-mix/

APPENDIX

A. Virtual Environment Programming

The virtual environment used in the test was primarily written in C# and is comprised of several components: the system controller, the user elements, and the environment entities. It utilized both the Leap Motion Interaction Engine and the Leap Motion Orion software development kit, both open-source software, to handle the physical interactions between the control schemes and the objects within the scene [27]. A diagram of the essential components of the program is presented below in figure 8.

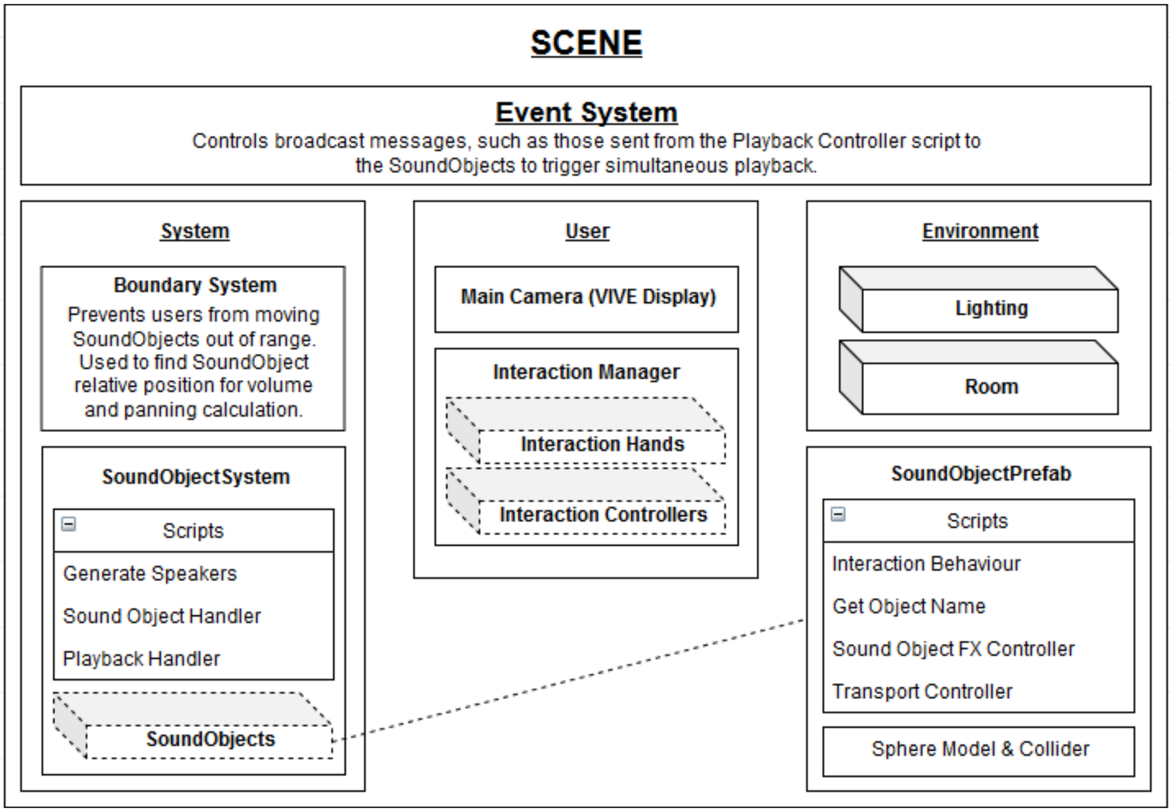
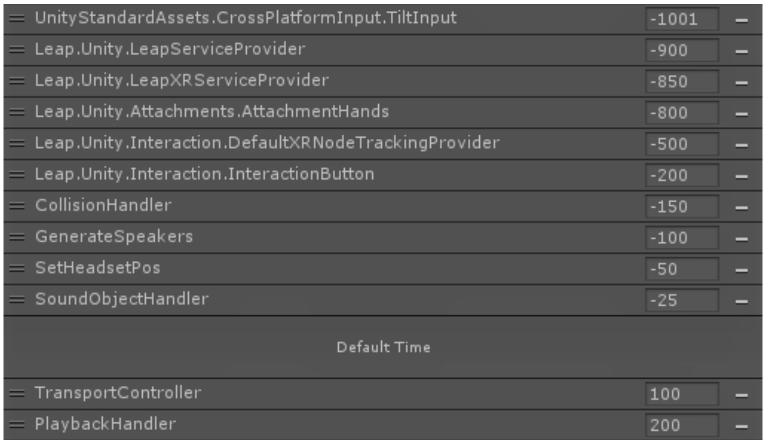

SCENE
Event System
Controls broadcast messages, such as those sent from the Playback Controller script to the SoundObjects to trigger simultaneous playback.
System
Boundary System
Prevents users from moving SoundObjects out of range. Used to find SoundObject relative position for volume and panning calculation.
SoundObjectSystem
Scripts
Generate Speakers
Sound Object Handler
Playback Handler
SoundObjects
User
Main Camera (VIVE Display)
Interaction Manager
Interaction Hands
Interaction Controllers
Environment
Lighting
Room
SoundObjectPrefab
Scripts
Interaction Behaviour
Get Object Name
Sound Object FX Controller
Transport Controller
Sphere Model & Collider

Figure 10. The programming structure within Unity.

The SoundObjectSystem would first load the mono audio files from the Resources/AudioFiles folder in the Unity project. Instead of hard-coding the audio files into the program in anticipation for using different stimuli, this allowed the test administrator to hot-swap the files in the resource folder within seconds, saving time in the event of redesigning the test.

The SoundObjectSystem additionally contained a Playback Controller script. By "arming" the Playback Controller using a radio button, the administrator of the evaluation could additionally trigger a ToggleChange button within the same interface to start or stop playback on all objects simultaneously. This prevented sound sources from being played sequentially, which would have led to timing or phase issues during playback.

The script execution order was important for proper function of the system, pictured below in photo 3.



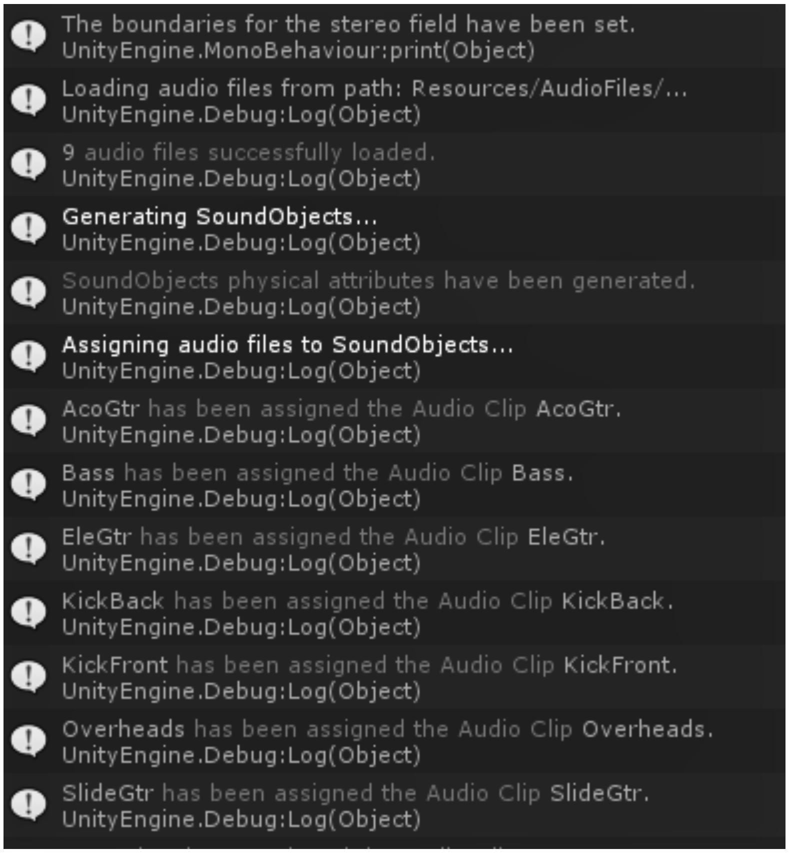
UnityStandardAssets.CrossPlatformInput.TiltInput	-1001
Leap.Unity.LeapServiceProvider	-900
Leap.Unity.LeapXRServiceProvider	-850
Leap.Unity.Attachments.AttachmentHands	-800
Leap.Unity.Interaction.DefaultXRNodeTrackingProvider	-500
Leap.Unity.Interaction.InteractionButton	-200
CollisionHandler	-150
GenerateSpeakers	-100
SetHeadsetPos	-50
SoundObjectHandler	-25
Default Time	
TransportController	100
PlaybackHandler	200

Photograph 3. The script execution order.

In order to alleviate some risk of sequential programming causing audio sources to be played out of time, the PlaybackHandler system employed use of message broadcasting to play each sound source. Each SoundObject would actively "listen" for messages related to playback, and upon an update frame where the administrator triggered the song to play, a message would be broadcast and the start time for all sound sources would begin on the same frame.

Additionally, to mitigate CPU usage and offer an efficient way for panning and volume to be updated, the SoundObjectSystem would start coroutines for updating panning and volume as soon as the scene would play and updated whenever an object changed position or was touched by one of the controllers. This allowed each panning and volume update to run independently of each other utilizing instancing within Unity.

The program displayed console messages assigned to individual routines, allowing the test administrator to ensure the routines properly ran in the correct order, pictured in Photo 4 on the next page.



```
The boundaries for the stereo field have been set.
UnityEngine.MonoBehaviour:print(Object)
Loading audio files from path: Resources/AudioFiles/...
UnityEngine.Debug:Log(Object)
9 audio files successfully loaded.
UnityEngine.Debug:Log(Object)
Generating SoundObjects...
UnityEngine.Debug:Log(Object)
SoundObjects physical attributes have been generated.
UnityEngine.Debug:Log(Object)
Assigning audio files to SoundObjects...
UnityEngine.Debug:Log(Object)
AcoGtr has been assigned the Audio Clip AcoGtr.
UnityEngine.Debug:Log(Object)
Bass has been assigned the Audio Clip Bass.
UnityEngine.Debug:Log(Object)
EleGtr has been assigned the Audio Clip EleGtr.
UnityEngine.Debug:Log(Object)
KickBack has been assigned the Audio Clip KickBack.
UnityEngine.Debug:Log(Object)
KickFront has been assigned the Audio Clip KickFront.
UnityEngine.Debug:Log(Object)
Overheads has been assigned the Audio Clip Overheads.
UnityEngine.Debug:Log(Object)
SlideGtr has been assigned the Audio Clip SlideGtr.
UnityEngine.Debug:Log(Object)
```

Photograph 4. A screenshot of the console window inside the Unity Editor for Tactile Mix.

The usage of the Unity Editor as part of the test allowed the test administrator to ensure that each subject during the test was seated at the same position and was exposed to the same stimuli at the same starting position. It also allowed the researcher to move objects back into the field of view in the rare case that a subject would knock them out of comfortable reach.

The full Unity project files, including all scripts for Tactile Mix can be accessed via GitHub, provided by a link in the Resources section of this paper.

B.1 Full Subject Survey Response Data

In table 12 below, the subject response data recorded in both surveys, the exit survey, and the task completion time in seconds is shown. This dataset was used for the analysis in section **4.0** of this paper.

Table 12. The full subject response data set.

			Exp			Vol	Vol	Vol	Pan	Pan	Pan	Time
ID	Age	Role	(yrs)	Preference	Scheme	Acc	Eff	Sat	Acc	Eff	Sat	(s)
2	27	Graduate	5.0	Controllers	Controllers	8	8	8	8	6	8	494
3	21	Undergrad	6.0	Controllers	Controllers	9	10	9	10	9	10	490
4	26	Graduate	6.0	Controllers	Controllers	10	10	9	9	10	10	412
6	27	Graduate	4.5	Controllers	Controllers	10	10	10	10	10	10	281
8	20	Undergrad	2.0	Hands	Controllers	8	5	5	8	10	9	584
1	65	Instructor	45.0	Controllers	Controllers	7	7	7	6	7	7	593
7	54	Instructor	34.0	Controllers	Controllers	9	9	10	8	8	10	292
5	24	Graduate	10.0	Neither	Controllers	8	7	10	7	9	10	331
9	60	Instructor	35.0	Controllers	Controllers	9	9	9	9	10	10	600
10	34	Instructor	20.0	Neither	Controllers	8	7	7	8	7	8	408
2	27	Graduate	5.0	Controllers	Hands	8	6	7	7	3	7	458
3	21	Undergrad	6.0	Controllers	Hands	9	8	10	8	9	10	449
4	26	Graduate	6.0	Controllers	Hands	9	5	5	9	7	7	392
6	27	Graduate	4.5	Controllers	Hands	7	8	7	8	7	8	325
8	20	Undergrad	2.0	Hands	Hands	7	3	5	9	7	9	600
1	65	Instructor	45.0	Controllers	Hands	6	6	5	6	7	5	472
7	54	Instructor	34.0	Controllers	Hands	5	4	1	2	2	1	412
5	24	Graduate	10.0	Neither	Hands	8	7	10	9	9	10	403
9	60	Instructor	35.0	Controllers	Hands	7	7	7	8	8	8	600
10	34	Instructor	20.0	Neither	Hands	7	8	7	7	7	8	378

B.2 Full Subject Verbal Response Data

Subject 1:

“Should be able to solo / lock channels. It seems narrow – the field is too narrow. The parameter update needs to happen faster. Would be cool to have mute and solo. Would be nice to have reverb zones. The hands are a little harder because of their tactile element. The hand controls were a little sticky sometimes. The hands are more novel but controllers worked a bit better, they were cool, but if it worked as good as the physical controllers, I would enjoy it better. The hands don't perform as well as the physical controls. What would be cool is if he were looking at the audio sources, it would be great to have a joystick panner. The interface reminds me of the same quality as early Pro Tools.”

Subject 2:

“For my first time in Virtual Reality, it was cool, and functioned a lot better than expected. As it is, it's beneficial for younger students. It's a lot easier to understand, or glance and get an idea of what the stereo field image is like. Putting it all in its own "world" makes a lot of sense. I like the controllers more than the hands. The hands would have trouble with proximity.”

Subject 3:

“I enjoyed the hand controls, cool to see hands in the Virtual Reality space. Weird because there was no haptic response. Visual icons instead of labels would be helpful, or in addition to the labels. Object collision was expected, but the hand would collide with other objects when interacting with an object. Easier to manipulate controllers because they had a lower profile than the hands. Some sort of tap to mute or tap to solo function would have been nice. Super enjoyable experience and cool to see it put into use.”

Subject 4:

“I can't wait until I can mix records like that. The system could use a solo button. There should be more processing options. This program is the inevitable future, and the demonstration makes me feel that's more so the case than I believed before. A laser pointer style control design would be more effective for interaction, but tactile was satisfying. The system felt crowded at times. I was wondering if the speakers were in the same place in their physical position as they were virtually.”

Subject 5:

“I wish for this to be in the modern studio environment. Awesome. Hand gesture control was awesome but took getting used to. The testing task was limited. I loved it a lot. I am impressed.”

Subject 6:

“I want to use this system to mix their own music. For the future of this, I would love to see spectral effects, reverb, and maybe trigger to indicate the instantiation of effects.”

Subject 7:

“In the trial, the physical controllers were far more superior and had far more control. Click and drag is easier. Other controllers, if it had to be hand detection, physical sensors would be the way to go. The struggle with the hands was figuring out when you could touch it. I also had trouble knocking things around. The physical controllers, because the controllers have no effect until you click the button, were much more like you could get what you wanted out of it. Much more satisfying. I could see using the physical controllers. It was odd how hot the face got.”

Subject 8:

“Weird to get used to, especially the hands. Sort of distracting, didn't look at visual labels, used ears.”

Subject 9:

“Very interesting mixing in VR. I preferred physical controllers. There were times when using hands that it would push away. That was distracting. I felt like I couldn’t accurately place the objects with the hand detection controls. I liked the physical controllers click and drag, felt easier to do the thing I was trying to do. More than one thing at a time was useful, with the physical controllers behaving more dynamically than the hand controllers. I immediately got used to the physical controllers.”

Subject 10:

“Impressive. I could not see difference between the two control schemes other than the lack of drag-and-drop functionality in the hand-controlled method.”

ACKNOWLEDGMENTS

This body of work is firstly dedicated to my parents, Bud and Donna, to my talented sister Olivia, and to Jake. Out of all the potentialities I could exist, observe, create and learn in – I am glad to share this one with the greatest family I could imagine.

To the experts at Warner Music Group, whose wisdom, gregariousness and expertise gave me a generous start to a fulfilling lifelong career – thank you for believing in me and giving me such great opportunities in the music industry.

To my instructors and faculty – thank you for imparting your wisdom which facilitated my goal to pursue the highest standard of academic success.

To my pioneering classmates and colleagues at Belmont University – Paul, Morgan, Austin, Owen, Tyler, Chris, and Jim – you all embraced me as an outsider and showed me a love which words cannot easily express.

To Will Wright and Andrew Gower, who taught me at an early age that experimenting with complex systems and navigating your way through them in an intuitive way was more valuable than winning or losing. This forged me into the person I am today at my earliest opportunity to learn.

To all the subjects who took part in the evaluation, thank you for your participation, timeliness, and enthusiasm.

To Clarke Schleicher and Paul Worley, who taught me the value in seeing the "forest for the trees".

AUTHOR BIOGRAPHY

Justin Bennington is a 24-year old audio engineer, musician, multimedia artist and futurist currently residing in his hometown of Windermere, Florida.

His most notable achievements include writing this paper and working with musicians, artists, and companies as the leader of his creative services company, Somewhere Systems.