Methods and Techniques for Capturing Music Concerts for Virtual Reality Experiences

Stephanie T. Benicek

Submitted in partial fulfillment of the requirements for the
Master of Music in Music Technology
in the Department of Music and Performing Arts Professions
The Steinhardt School
New York University

Advisor: Dr. Agnieszka Roginska
Second Reader: Paul Geluso

2018/05/14

Table of Contents

List of Figures

Abstract

As virtual reality gains popularity, it is worthwhile to explore applications outside of gaming and training purposes. As concert performances become more intricate in design, the shows become more of an experience – and consequently become more expensive to attend. A virtual reality concert experience is designed to be fully immersive so that the user feels as if they are actually in attendance even if they are not. Such a concert experience would allow those who may not have access to stadium tours or who may not be able to afford to go to one to experience the music and touring world.

3D and immersive audio contribute greatly to the believability of virtual reality scenes, yet there has not been much experimentation or quality research on optimizing the virtual reality concert experience. This research project examines recording and production techniques for capturing a live concert performance for virtual reality applications. Three distinct mixing implementations were then presented to human subjects to determine their preference for the most immersive and realistic experience. A jazz performance was recorded for subjective testing.

For recording, multiple audio components and a video component were set up and utilized. Three separate audio mixes were created, all using the same video. Each mix was evaluated in terms of immersiveness, naturalness, and audio quality. Results show that subjects preferred the mix of more indirect sound with some direct sound and the mix of more direct sound almost evenly.

This project only scratches the surface of what can be done using immersive and 3D audio with virtual reality for music concert applications. Due to equipment and venue limitations, this study would benefit from more research and experimentation. Future work would explore different virtual reality camera options, various venue sizes, and live streaming possibilities. Though 3D audio and virtual reality technology are advancing rapidly, this is only the beginning for the field.

Acknowledgments

First and foremost, I would like to thank my family for being so supportive.

To Dr. Agnieszka Roginska and Paul Geluso, thank you for your guidance, wisdom, advice, and encouragement throughout this whole project.

To Marta Olko, thank you a million times for all of your help and patience.

To the NYU Immersive Audio Group, particularly my recording team, thank you for your assistance and support.

Finally, to New York University, their facilities, and their event and theatre staff, thank you for your cooperation and for creating such a welcoming space for learning and research.

1.0 Introduction

In recent years, the demands for live sound, virtual reality, and 3D audio have been growing. In live sound, more intricate and expensive shows are becoming the norm. Virtual reality is becoming more accessible to the general population, and 3D audio is becoming a preferred playback system for those who have access to one. There has been some crossover in these areas, but with industry standards shifting, now is the time to explore new creative applications that involve all three of these areas of music technology.

As live shows have become more expensive, the divide between those who can afford the experience and those who cannot has grown tremendously. Recent technology could be used to narrow this gap, but only if sufficient research is done to optimize the listener's experience. Virtual reality has been used to create more realistic video game scenes, and the same concept can be applied to concerts. Sound is a large factor in how believable a virtual reality scene is, particularly with music. 3D audio is a better representation of how humans hear naturally, and when combined with high-quality virtual reality visuals, a realistic virtual reality concert experience can be accomplished. Moreover, those who may not have access to concerts can experience one as if they were actually in attendance. Though there have been some attempts at creating a virtual reality concert experience, none of them fully explore the technology available, and there is no standard or suggested protocol for how to properly capture a concert for this purpose. This project's approach was to design an experiment that brought together three of the key components of music technology to create a virtual reality concert experience.

Live sound, simply put, is the sound you hear when you attend a concert, performance, sports event, or other live event that requires the sound to be amplified for a crowd. For the purposes of this paper, live sound will refer to the concert application, whether it is a small show or a stadium show. Live sound engineers must properly mic instruments and

ensure a high-quality sound for everyone in the audience while staying on their toes for any issues that may arise during the performance.

The second key component is 3D and immersive sound. “A crucial aspect of your awareness and appreciation of sound in everyday life is that sound comes from all directions”, (Kendall, 1995). One can play recorded music on loudspeakers and the sound will seem as if it is coming from in front of you, whereas playing the same recorded music through headphones will make it seem as if the sound is inside your head. However, neither method allows one to listen to the music the way we would naturally perceive sound in a concert hall, for instance. 3D and immersive audio imitate the way we naturally hear, with sound coming from all directions. This can be rendered on multiple loudspeakers or through headphones, and when combined with mixing and recording techniques can sound very natural.

Third, virtual reality (VR) has been rapidly gaining popularity in recent years, particularly in gaming, though it is used in other settings as well. “VR is partly about increasing the level of detail in the audio environment”, which is extremely important in maintaining the virtual environment (Rumsey, 2016b). Audio has a large role in keeping the user immersed. “If someone’s watching a VR concert and moves toward the stage, the singer’s going to need to be louder, with fewer reflections from the concert hall walls. Turning your head will maybe make the guitars louder and the drums softer, or vice-versa”, (Anderton, 2016). Certain details like these are what make the VR scene seem believable.

In combining these three areas of music technology to create a virtual reality concert experience that utilizes 3D audio, new and growing consumer demands can be met. A virtual reality concert experience can make an artist’s performance available to an audience it may not have been able to reach before, such as those who may not be able to afford an expensive concert ticket or those who are in less accessible areas of the touring world. Artists can give their fans a new way to experience their music while finding another method to make a profit as physical media sales decline. While there has been some exploration of virtual reality concerts, these implementations have not fully utilized the

benefits 3D audio has to offer. By implementing a virtual reality concert experience with 3D audio, a more immersive, life-like, and enjoyable concert experience is possible and can be accomplished by using high-quality video capture equipment and 3D audio capturing methods, such as using microphone arrays with heights, to record the performance.

1.1 Motivation

In today's society, going to a concert or live performance is a fairly common activity. Many people jump at the opportunity to see their favorite artists perform live, and with physical media sales decreasing, more artists are hitting the road. "Because of the downfall of record sales, the artists gradually have to make money out of their live performances rather than by selling records: they go on tour to make profit whereas it used to be a way to promote their upcoming album", (Le Henaff, 2015). People no longer need to purchase a full album to hear the one song they like. Instead, one purchases the single from iTunes or listens to it on a streaming service like Spotify. As a result, not only are more artists touring, but the shows are also becoming more intricate – and more expensive. Yet, this does not seem to be a huge issue for consumers. "A big spectacle ranks up there for most concertgoers . . . [the shows] offered plenty of bang for the buck in the form of stadium shows . . . the rise in stadium shows points to a potentially brighter future ahead for the concert industry", (Young, 2014).

"Our senses are highly stimulated when attending a concert. While experiencing a record mainly remains in listening, attending a live show offers a wide sensory awakening, i.e. the staging (the scenery, the lights, the movements of the musicians), the physical sensations (subwoofers, high SPL) and the particular emotional state that is created by the event (the crowd, the restlessness), have a huge impact on how we feel music," (Le Henaff, 2015). Because of these heightened senses and the way music makes us feel, artists and concertgoers may want that experience captured. For concertgoers, listening to a captured live performance can help them relive their own experience, or get a sense of what that artist's energy is like on stage if they were not in attendance. This explains the popularity of

certain tour documentaries. Jonathan Demme's tour documentary of the Talking Heads' *Stop Making Sense* tour captures the performance so well that it gives "the film an uncanny sense of *being there*", and the documentary is still regarded as one of the best tour documentaries of all time (Watercutter, 2017). Artists may also want to hear their performances and adapt their technique for future performances, and that may lead to making a more ostentatious show.

Experiencing live music is a trend that is not disappearing any time soon, and the industry is adapting to make the future of live music greater than ever before. In combining live sound with virtual reality and a greater and more believable audio component, concertgoers can take part in an even more enhanced concert experience than previously thought imaginable.

1.2 Goals

The primary goal of this project was to create a unique and immersive virtual reality concert experience with 3D audio that improved upon current research and implementations. Secondary goals were to determine which recording technique works best for capturing live shows and reproducing them for virtual reality experiences and to determine a potential standard for implementing virtual reality concert experiences.

2.0 Literature Review

There are many independent resources describing live sound, immersive audio, and virtual reality. Although there is some crossover between these topics, there are only a few applications utilizing all three simultaneously. The goal of this section is to provide a general knowledge of these subjects of music technology, as well as to describe a few examples of prior applications.

2.1 Live Sound

To effectively capture a live sound event, proper recording and postproduction techniques must be considered and utilized.

2.1.1 Recording Techniques

Ultimately, it is the live sound engineer's job to capture high-quality audio of the performance while preserving the energy of the performance. As Clukey (2006) mentions, the type and placement of microphones, a mixing console, and a recording device are key to capturing a live show and should be chosen carefully for the highest quality recording.

Both Clukey and Nady suggest using two condenser cardioid microphones to capture live sound (for a stereo mix), as the microphones are flexible with a wide dynamic range and low distortion levels (Clukey, 2006), yet still pick up the ambience of the performance (Nady, 2007). A coincident pairing of microphones is better suited for smaller ensembles, and a near-coincident pairing is better for capturing larger ensembles (Clukey, 2006). A coincident pairing has the microphones pointing inward towards each other, while the near-coincident pair points outward away from each other. This utilizes the polar patterns in the best way for capturing their respective ensemble sizes.

It is also important to spot mic the instruments properly to avoid unwanted feedback. If guitar and bass amplifiers are miked, the mics are close enough to the sound source that

they will not pick up any unwanted signals. With drums, it is important to point the null end of the microphone towards any signal you do not want to pick up. It is always a good idea to reduce reverberation of the drums by adding some sort of insulation, if possible. Vocals are best captured with cardioid polar pattern mics as well, and it is important to point the null end towards their stage monitor to prevent feedback (Nady, 2007).

A proper live sound mixing console should be balanced to prevent unwanted noise and should have phantom power for condenser mics (Clukey, 2006). There are various different types of recording devices, both analog and digital, and it is better to choose the one you are familiar with that also is suitable for the delivery method. When recording, it is important to monitor volume levels. The meters should light up mostly in the green zone, with occasional peaks in the yellow and sometimes red zones. If your levels fall more in the red zone, you have a greater risk of issues such as clipping and distortion. It is recommended to record between -9 and -3 dB (Clukey, 2006). Nevertheless, it can be difficult to monitor levels live "because of excessive crowd noise, which [can be] captured by the stage mics, making the resulting sound quality too poor to warrant the release of [recordings]," (Rumsey, 2017c). A clean sound from the board will help later in postproduction if crowd levels get out of control. Still, the live sound engineer should be focused and watching the levels, adjusting them as needed throughout the performance.

2.1.2 Postproduction Techniques

Proper postproduction mixing is nearly as important as proper capturing of the sound. Normalization (or stabilizing dynamic ranges of the audio), trimming audio, editing crowd noise, reducing unwanted noise, and mixing are some of the tools used in postproduction of the audio. For proper, high-quality postproduction editing, your equipment should include a good pair of monitor speakers, headphones, and a good amplifier for playback (Clukey, 2006).

According to Clukey (2006), it is typical to have 95% normalization, though you must be cautious when normalizing live sound, as audience noise is often the loudest element and

can throw off the balance of volume levels. If it is the desire to keep applause in the recording, it is recommended to reduce the volume of it by 6 or 10 dB for smoother transitions into the actual music (Clukey, 2006). Consistency is key in the normalization process, and the loudest parts of songs should also be compared to the rest of the track when normalizing. Tiny fade ins and fade outs make for a better sounding track, and when keeping applause in, it is suggested to keep just several seconds of applause before applying a two second fadeout, followed by another two seconds of silence or "room tone" to accommodate burning the track for delivery (Clukey, 2006).

When it comes to a recorded live performance, there is a preference for its delivery. Pras (2016) understood the significance of a performer's character and expression during a live performance and conducted an experiment to see which editing process would better accommodate listener preferences of the recorded performance. Tonmeister students listened to three versions of classical performances: an edited and polished version, a raw studio take, and a performance in front of a live audience in the studio. Results showed most preferred the edited versions, and that the post-production live recordings were favored, which led Pras to believe that creative direction in the studio can positively impact a concert performance (Pras, 2016). Although this experiment was aligned with classical music performances where audience noise is generally minimal, the conclusion drawn can be that a tastefully edited version of a live performance is preferred among listeners.

Unwanted noise reduction should be dealt with carefully and only every so often. Sometimes, taking out too much unwanted noise makes a track sound synthetic or like it is missing some elements to it. A recorded live track, as mentioned in section 2.1.1, should preserve some of the character the performance, so it is best to keep the track sounding natural.

There are, however, some other mixing methods to keep the track from sounding synthetic. For instance, mixing in some overhead drum tracks with the rest of the drum set or mixing in height layers (discussed further in section 2.2 of this paper) can add to the ambience. Of course, basic tools like panning and reverberation can help the track as well (Nady, 2007).

2.2 Immersive and 3D Audio Recording Techniques

3D audio can add many layers and dimensionality to a recording. Surround sound technology took a step forward from stereo by adding in other channels around the listener. This resulted in 5.1 systems, which are composed of left, center, right, left surround, and right surround channels, and 7.1 systems, which add an extra two surround channels directly to the sides of the listener. What then upgrades a system from surround sound to 3D sound is the addition of height layers.

2.2.1 Ambisonics

Originally introduced by Michael Gerzon, Ambisonics is a spherical surround sound technique that is independent of any specific playback system. It uses "spherical harmonic basis functions that could encode the portions of a sound field originating from many different directions around a listener's position" and is meant as an alternative to surround sound and channel-based stereo systems (Roginska & Geluso, 2018). Ambisonics is meant to represent all directional sounds equally, in both the horizontal and vertical directions, and is ideally arbitrary in order (n^{th}- order Ambisonics). The first sound field microphone was a first-order 4-channel system composed of W (omnidirectional), X, Y, and Z (all bi-directional) channels. Ambisonic microphones are compatible with stereo systems but have "full spherical portrayal of directionality" (Roginska & Geluso, 2018). An example of these tetrahedral microphones that yield a full 3D sound field is the Sennheiser AMBEO microphone. While revolutionary, first-order Ambisonics can be improved by increasing the order number above one (High Order Ambisonics, or HOA). HOA match the original sound field as much as possible; increasing the number of spherical harmonics increases directionality resolution and spatial resolution.

There are different formats to Ambisonics: A, B, C, and D formats. A-format is the recording format representing the outputs (LF, LB, RF, RB) of the tetrahedral microphone. B-format is independent of speaker layout and recording setups, and yields the W, X, Y, and Z channels previously mentioned. Their decoding equations are represented below in Equation 1. C-format, also known as UHJ format, is meant more for broadcasting, diffusion, and consumer

purposes. It provides signals that are directly compatible with conventional systems of reproduction. D-format, also known as G-format, represents the set of input signals to loudspeaker configuration (Roginska & Geluso, 2018).

$$W = LF + LB + RF + RB$$
$$X = LF - LB + RF - RB$$
$$Y = LF + LB - RF - RB$$
$$Z = LF - LB - RF + RB$$

Equation 1: Decoding equations for B-format ambisonics. The W channel is omnidirectional, while the X, Y, and Z channels are bi-directional.

2.2.2 Immersion

"Immersive sound can give the listener an experience of *being there* through sound. Compared to vision, sound provides a fully immersive experience and can be perceived from all directions simultaneously," (Roginska & Geluso, 2018). "In VR there is a pressing need to ensure spatial correspondence between what you see and what you hear," (Rumsey, 2017b). Furthermore, "head tracking, efficient binaural synthesis, and individualization [can] contribute to a convincing listener experience," (Rumsey, 2017a). The terms 3D audio and immersive audio are often used interchangeably, and serve a common goal: "to create a sense of envelopment . . . [by creating] the feeling of surroundedness or engulfment . . . by immersing a listener in sound originating from all angles of azimuth and including elevations. One way to create envelopment is to position sources in 360° around the listener. This form of envelopment is particularly effective for music-only content," (Roginska & Geluso, 2018). Berg (2009) mentions that a sense of envelopment has been proven to be an important factor in listener preference and listener experience of "being present at the venue where the sound is located". Wide images and a combination of both direct and indirect (reflected) sound can contribute to the feeling of being surrounded by sound and a more natural representation of sound. Another useful

way to create this sense of envelopment is by using microphone arrays to capture the sound from the source.

2.2.3 Microphone Arrays

"The addition of height sound radically improves the listening experience" and is of better quality than upmixing channels (Geluso, 2012). When recording, there are certain microphone arrays, like the Bowles array and the Hamasaki cube, that are meant to capture the height layer. This typically consists of placing microphones up high to capture the sound that travels upward. For instance, when recording an ensemble, placing a microphone array above the conductor can yield in a more "natural" recording because the sound that naturally travels upward is captured and in turn, it adds a layer of depth to the recording (Bates & Boland, 2016).

The Bowles array is composed of four omnidirectional microphones and a cardioid microphone in the center of a horizontal array, with an added height layer of four super-cardioid microphones. "The height array was designed to capture sound reflections coming from the ceiling and higher areas of sidewalls," (Roginska & Geluso, 2018). The Hamasaki Cube is composed of four bi-directional microphones arranged in a square in a main horizontal plane and a height horizontal plane. It is an expansion of the Hamasaki Square, with height channels capturing additional reflected sound.

Michael Williams (2003) proposed using multichannel microphone array systems to better improve listener envelopment. A four channel system was proposed, with cardioid microphones positioned in each corner of a square, each at a 90° angle and 25 cm apart. This microphone array has been dubbed the Equal Segment Microphone Array (ESMA). Though this array and other proposed arrays were successful for capturing a complete surround sound recording, "realistic continuous sound field reproduction could only be obtained with these systems by the use of a loudspeaker configuration using equal segmentation of reproduced sound field," (Williams, 2003). Williams also mentioned that coincident microphone systems are not ideal for these multichannel arrays, as they lack some information necessary for optimal coverage.

Riaz et al. (2017) used an ESMA to capture a 360° recording in London's Abbey Road Studios. The authors mention that widening the distance between the cardioid microphones in the array from 25 cm to 50 cm can help improve location accuracy, and the array they used had an additional square of cardioid microphones positioned above the ESMA to capture the studio ambience. Though they also used other 3D and ambisonic microphones and arrays, such as an IRT Cross and the Sennheiser AMBEO microphone, they have not yet conducted subjective testing to determine a listener preference. Riaz et al. (2017) conducted their own listening tests, and found that "the combination of spot microphones and [the ESMA] works well to aid localization of individual sound sources and capture more of the room's ambience inducing a greater sense of the recording space."

Lee (2016) built off Williams' idea and found that a quadraphonic microphone array is suitable for virtual reality head tracked audio reproduction. A vertical extension of Williams' quadrophonic array was proposed, adding a figure-of-eight or bi-directional microphone to each cardioid microphone in the array. This created a mid/side (M/S) quadrophonic array, better known as an array of mid/Z (MZ) pairs introduced by Geluso (2012). The bi-directional microphones can then be decoded into height layers. Lee also found that near-coincident microphone systems provide better localization.

Geluso based the MZ microphone pair on ambisonics and MS recording techniques. Because the ambisonic format has W, X, Y, and Z channels, with X being forward facing, Y being sideways, and Z being up facing, it can be concluded that the Z channel is the raw height channel (Geluso, 2012). Geluso (2012) hypothesized that "complex height information can be captured by pairing horizontally oriented microphones with vertically oriented bi-directional microphones". Therefore, MZ is based on M/S with a Z channel bi-directional microphone oriented vertically and coincident to the horizontally oriented microphone to create a middle-Z pair. The MZ system is compact and doesn't require specialty microphones, which makes it more practical for recording engineers. "If the null of the Z microphone is facing the soundstage, excellent separation of the Z channel and its associated horizontal channel can be achieved without having to space the microphones far

apart from each other," (Geluso, 2012). Moreover, "using a standard MS decoder, the vertical pick-up angle for the MZ pair can be determined remotely or in post-production to create effective height channels," (Roginska & Geluso, 2018).

The array specifically used in this project for capture was an ESMA that also implemented Lee's expansion of using bi-directional microphones to create an ESMA of Geluso's proposed MZ pairs. A diagram representing the polar patterns and general concept is reproduced in Figure 1 below.

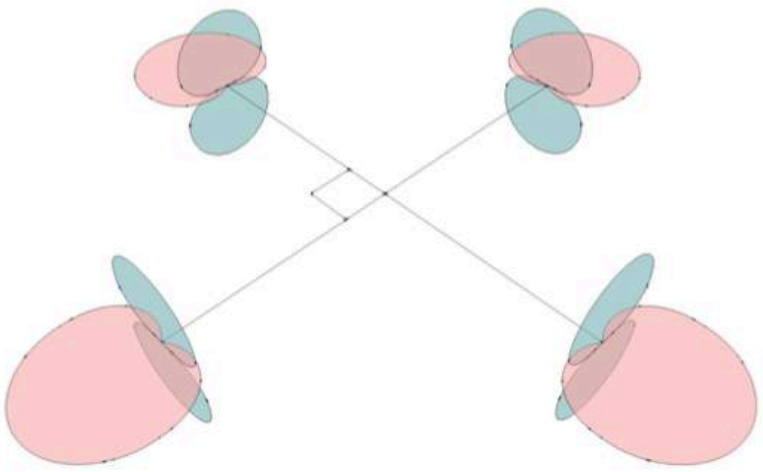

Figure 1: Polar patterns of ESMA with an MZ system in each corner. Pink represents the cardioid polar patterns, and blue represents the bi-directional microphones. Reprinted from "Capturing and rendering 360° VR audio using cardioid microphones," by H. Lee, 2016, Conference on Audio for Virtual and Augmented Reality. Copyright 2016 by the Audio Engineering Society.

2.2.4 Mixing

Just as it is important to properly capture the sound in the recording process, mixing and postproduction should be appropriately handled as well. Two key elements to mixing spatial (immersive) audio are panning and reverb. Panning sources can give the listener the sense of where around them the sound is coming from, while reverb, when used properly, can give the user a sense of how far away the source is from them. Reverb can also add a sense of ambience.

For example, if one wanted to give the impression that they were in the middle of a venue listening to a band, you could pan the instruments left, right, and center to illustrate where the performers were on stage. Different amounts of added reverb can not only give the listener a sense of how large or small the venue is, but also how far back the instruments are (e.g. drums are behind the other instruments and a lead vocalist may be in front of everything else). A small amount of all the instrument tracks could be directly to the left and right of the listener, mimicking the loudspeakers in the venue. Additionally, if a sound field was captured (e.g. general noise in the venue), this track could be positioned in the surround and height channels to give a more immersive and fuller feel. Ultimately, the goal of 3D audio is to have the listener hear from sources that are not heard in stereo but that do better reflect the way we naturally hear (Kendall, 1995).

2.2.5 Reproduction

The main benefit of 3D audio is that it is a closer representation of the way humans naturally hear and interpret sound. As mentioned, sound comes from all directions, not just two loudspeakers in front of you. When playing back 3D audio, multiple loudspeakers are necessary, however, you can still get a sense of sound being all around you in headphones. This is done using head-related transfer functions (HRTFs) and head tracking. With headphone virtualization, "the aim is to maximize perceived externalization by simulating reproduction in a virtual room, while maintaining a natural timbral balance . . . Headphones offer one possible destination for immersive content, without excessive hardware requirements, provided that suitably convincing signal processing can be developed to render material in a way that sounds natural to listeners," (Rumsey, 2016a).

When a sound wave approaches us, "high-frequency energy is specularly reflected away and low-frequency energy diffracts and bends around the listener . . . the sound waves that reach the listener's two eardrums are affected by the interaction the original sound wave with the listener's torso, head, pinnae (outer ears), and ear canals. The composite of these properties can be measured and captured as a head-related transfer function", (Kendall, 1995). HRTFs are a measure of how sound naturally moves around us. Head tracking is also

used. With head tracking, a sound source will change position based on where your head is positioned. For instance, if you hear a sound to the right and in front of you, but you turn your head towards the right, the sound will then seem as if it is coming from your left, rather than having the sound follow your right ear. The combination of HRTFs and head tracking is what yields a 3D sound through headphones.

"The ultimate goal of reproducing binaural signals is to recreate an acoustic signal at the eardrums of the listener that would be equivalent to the signal that a listener would hear under natural listening conditions . . . headphones provide a controlled listening environment," (Roginska & Geluso, 2018). While headphones can position a listener in a constant sweet spot, they also are less interactive with the listener's environment and are subject to "inside-the-head locatedness" (where all sounds seem to be originating from inside the head), which can lead to listener fatigue (Roginska & Geluso, 2018). Headphones also limit crosstalk, which is more present in loudspeaker playback. This can result in different listening experiences if the audio engineer mixed over loudspeakers and playback was intended to match the way the audio engineer was hearing sounds.

"Headphones with extended capabilities can create a spatial listening environment . . . virtual loudspeakers over headphones aim at emulating the experience of listening to one or more real loudspeakers in a real listening environment," (Roginska & Geluso, 2018). In order to implement virtual surround sound over headphones, "the signals to be output by the loudspeakers are processed where each speaker is treated as an individual sound source. The fundamental way to create the headphone representation of surround sound systems is to process each channel's audio signal by the impulse response corresponding to the location of the loudspeaker through which that channel would be played," (Roginska & Geluso, 2018). The addition of HRTFs and head tracking capability help make the listening environment more interactive and realistic for the listener. Though personalized HRTF measurements exist, they can be invasive and time-consuming, thus averaged HRTF measurements are often used in reproduction.

3D audio is more immersive because you have sound sources coming from every direction. Just like surround sound is an upgrade from stereo sound, 3D sound is an upgrade from surround sound and results in a better listening experience.

2.3 Virtual Reality

> *"A viewer enjoys a VR scene via his head mounted VR display from a fixed point in the scene. He can freely turn and move his head to observe other details in the scene and may follow the actors by listening and watching. He likes to change his view to explore the details in the scene, looking around freely. At the same time the user wants to have an accurate binaural representation of the scene, so that he can follow the story without the necessity to (visually) search the virtual scene to find the main action."* (Altman, Krauss, Susal, & Tsingos, 2016)

2.3.1 Believability

Perhaps the most important characteristic of virtual reality is believability. In order to feel fully immersed in a "different world", you need to believe that you are there. "A virtual auditory space is an acoustic environment created through the use of loudspeakers or headphones designed to replace or augment the natural listening environment," (Roginska & Geluso, 2018). "An essential aspect to the experience of attending a live performance in-person is the complete aural immersion. A listener becomes surrounded and engaged with the aural experience, including not only the performance on stage but also the noise and cheers of the audience and the acoustic environment of the venue," (Jacuzzi, Brazzola, & Kares, 2017). There are a few different approaches to accomplishing maximum believability in virtual reality, particularly with live sound applications.

"Three-dimensional sound needs to keep the illusion of reality intact, with accurate localization cues and spatial environments", (Rumsey, 2016b). As Rumsey (2016b) mentions, most VR experiences use headphones, and the appropriate way to keep the illusion intact is by using head-related transfer functions (HRTFs). HRTFs are unique to

individuals, as they represent time and spectral information of sounds that are carried to the ears (Rumsey, 2016b). They are a better representation of sound as it naturally occurs to us, and are used heavily in spatial audio and virtual reality because of this reason. However, you still must be cautious with headphones - if their response is poor or if the low-frequency response of the headphones is too low, the spatial cues can lose their natural quality and believability (Rumsey, 2016b). It is always best to consider these types of drawbacks when creating a VR project. Another suggestion made by Rumsey (2016b) is to leave open the opportunity to create an accurate soundscape by layering mono sources and adding spatial processing.

"Within surroundedness, the other attributes accounting for envelopment by reverberation, ambience, applause, wide single source, multiple sources, etc. are included," (Berg, 2009). These factors help contribute to a more immersive, and therefore believable, experience. A greater sense of immersion for the listener helps make the virtual soundscape more believable because it better mimics the environment that is being represented. The goal of VR is to make the user feel as if they are in the new environment that is being presented to them, as opposed to the environment they are actually in. To feel fully immersed in a VR application that places you in the middle of the audience of a concert, crowd noise and applause are significant factors.

As mentioned, for a concert experience crowd noise is significant. Stefanakis and Mouchtaris (2016) captured a crowded sports arena with a circular microphone array. Though they had a limited amount of sound sources, they concluded that perceived spatial impression and sound quality were improved using a circular microphone array for capture (Stefanakis & Mouchtaris, 2016). The circular microphone array is better applied to a large venue such as a stadium or arena, which due to limited resources is beyond the scope of this paper, but is still worth mentioning as a method to capture all sound, including crowd noise, in a large venue.

With VR, another contributing factor to believability is video quality. Previous research has already concluded that audio quality and video quality do affect each other. Gaston, Boley,

Selter, and Ratterman (2010) got more specific and conducted an experiment that evaluated certain audio artifacts and their effect on video quality. Their results showed that specific impairments affected perceived video quality in different ways (Gaston et al., 2010). Some impairments they focused on included brightness, pixelation, and desaturation of the video. In some cases, the impairment appeared to improve audio quality, and in others, it had the opposite effect. Regardless, they further prove there is a link between audio quality and video quality, and they must work well together to optimize maximum believability.

2.3.2 Immersive Audio with Virtual Reality

"Virtual reality is hungry for 3D audio", (Anderton, 2016). Because virtual reality can be seen as a submergence into another world, and because the visual aspect is limited to field of view, the audio aspect needs to not only reflect upon what is in the field of view but also what is not in order to fully enhance the experience (Altman et al., 2016). 3D audio goes beyond mimicking our perception of hearing by creating an environment around the audio as well (Anderton, 2016). As mentioned in section 2.2, immersive and 3D audio are a better representation of how we hear naturally, which is why they create a more immersive and believable VR experience. However, until recently, most commercial applications of virtual reality hardly touched upon on 3D audio.

A common immersive audio technique for VR is object-based audio (OBA). With OBA, sounds are assigned as arbitrary objects (with metadata) rather than assigned to specific loudspeakers. The metadata from these objects can then be decoded, and the objects are assigned to specific loudspeakers for each unique playback system, allowing flexibility across setups. This ideology can also be applied to headphones, as most VR applications use headphones and not multiple loudspeakers. OBA is already used in film and cinema applications, so the same mixing techniques can be used for audio and gaming VR purposes. OBA provides high spatial resolution and flexibility across platforms and setups (Altman et al., 2016).

There are certain programs and plugins like G-Audio Lab and Facebook 360 that help audio engineers implement 3D audio mixes for virtual reality purposes. This often includes a "drag and drop" type of operation that allows the audio engineer to drag sound sources to where they appear on video, with other tools within the application to aid in fine-tuning the mix and creating a more realistic and accurate representation of what was captured.

2.4 Significant Examples

There are some significant prior works relevant to the project (Virtual Reality Concert Experience, or VRCE) outlined in this paper, though many of the applications outlined below only relate to a small piece of it. The closest application to this project is NextVR and their innovative technology, though it still seems to have some drawbacks. These drawbacks served as items to address in the VRCE project outlined in this paper. Though all the below applications are relevant, the VRCE dives deeper into these concepts and updates them for a better experience. Still, these applications are worth mentioning, and at the very least, provide insight into the demand for VR experiences.

NextVR is a company that is dedicated to creating virtual reality experiences. In 2014, they teamed up with rock band Coldplay to stream a virtual reality concert experience. "The virtual reality concert puts Coldplay fans directly in the middle of the action, viewing the band members as if they are on stage with them. The concert film is the first ever broadcast quality VR experience", (NextVR, 2014). NextVR works with LiveNation to broadcast certain concerts and sporting events in VR and has worked with artists other than Coldplay as well. Though this is similar to the concept presented in this paper, NextVR hasn't provided any information as to how they record their audio for these VR experiences, or what their audio quality is like. The videos don't have a wide range, and the audio doesn't change positions to reflect when your head position changes.

A more recent application of 3D audio in a live sound setting is a 3D binaural performance of The Encounter, a play about Loren McIntyre who was a photographer for National

Geographic. “Audience members wear headphones throughout the performance to feel sensorially immersed through binaural sound (wherein each ear hears its own soundtrack, achieving a kind of “3D-listening” effect) . . . The production’s earphones work to both isolate audience members from one another and plunge them further inside a communal theater experience”, (McKee, 2017). The concept of incorporating 3D audio binaurally through headphones is often used in VR settings.

Augmented reality (AR) is similar to virtual reality except that it enhances the world as you are experiencing it, rather than acting as a means to escape the world as VR does. AR serves as a mix of real and virtual listening environments (Floros, Kapralos, & Moustakas, 2016). Floros, Kapralos, and Moustakas expand this concept by using virtual sources that produce synthetic 3D sounds that enrich the user’s listening environment. They utilized this ideology in an electroacoustic music concert and found that the majority of the people they surveyed after the performance enjoyed the AR musical experience and would attend another one given the opportunity.

Diaz and Koch (2016) captured a concert in VR using an Omnicam360 (ten individual cameras that make up an HD 360° panoramic picture) and a custom circular microphone array made up of eight hyper-cardioid mics placed around the camera. A Samsung GearVR was used for playback. “The concert was streamed in its entirety on the web and to multiple mobile devices . . . the current implementation is capable of streaming up to eight channels”, (Diaz & Koch, 2016). For the mobile devices, object-based audio was used, though the webstream had to downmix to stereo for rendering purposes. Diaz and Koch never discuss how listeners reacted to this technology, but it is considered a work in progress as they strive to improve this implementation.

2.5 Discussion

3D audio and virtual reality have been around longer than most people think, but are just now starting to gain more traction as technology is improving and more recreational

applications are now being implemented (previously, this technology was only available for training purposes, such as for the military). As mentioned, virtual reality has now become more popular because it is more accessible, and is largely used in gaming. 3D audio is also rising in popularity as more people are realizing it has a "better" sound (credited mostly to the fact that it better replicates the way we hear naturally). However, the two aren't being used together that often, as the focus with virtual reality lies mostly on graphics, and the implementations certainly shouldn't stop at gaming.

Each of these applications discussed in section 2.4 has been successful in different ways and have certainly paved the way for others to explore these areas of music technology, particularly in the world of live sound and performances. NextVR serves as the biggest inspiration for the VRCE discussed later in this paper, as it has the most relevant application, but other concepts such as the recording techniques mentioned in Williams (2003), Lee (2016), Geluso (2012), and Riaz et al. (2017) are significant as well. Still, there are improvements to be made as "production workflows for creating immersive musical experiences for VR are still in their infancy" (Riaz et al., 2017). The goal of the VRCE from the beginning was to create a virtual reality concert experience with the best quality video available that puts the user in the middle of the concert, surrounded by other concertgoers, that is accompanied by 3D audio that mirrors the head tracking of the video for maximum believability. This is discussed in detail in the following Methodology section.

3.0 Methodology

Ideally, the Virtual Reality Concert Experience (VRCE) proposed in this paper would give users an immersive virtual reality experience with 3D audio and head tracking capability. With the VR headset and headphones on, they should feel as if they are sitting in the audience of a concert. If the user turns their head, the audio should appropriately adapt as if the user weren't wearing headphones at all and they were hearing this sound happening in front of them, i.e. the VRCE sound should mimic the way humans naturally hear and perceive sound.

For capturing both the audio and the video of a performance, the setup must not block the view of any audience members and should be compact enough that it will not be knocked over or tripped over. This way, the safety of patrons and equipment will not be jeopardized. In larger venues, the setup could be placed by the mixing board, particularly if the soundboard is in the middle of the audience rather than at the back of the venue.

The methodology for this project can be broken down into four subcategories: capture, mixing, delivery, and testing.

3.1 Capture

The concert recorded took place February 10, 2018 in New York University's Frederick Loewe Theatre. The jazz performance covering Duke Ellington's song "Portrait of Louis Armstrong" was part of New York University's first annual Black History Month Celebration Through the Arts. Other performances were recorded as well, but only one was used for testing and evaluation to better control variables. The ensemble for the selected performance was composed of piano, trumpets, solo trumpet, upright bass, saxophones, trombones, electric guitar, and drums. Figure 5 shows the band setup. A jazz performance was used because it had a more interactive crowd than other performances, which in turn yielded a more immersive playback environment because audience noises added believability. Additionally, the jazz performance most closely represented a standard

concert setup that could be applied to many musical genres. Recordings happened on the New York University campus for accessibility purposes.

To capture the performance, a GoPro Fusion was used for videography. This device offers high-quality 360° video specifically for virtual reality purposes. Audio was captured with two Sennheiser AMBEO ambisonic microphones, an Equal Segment Microphone Array (ESMA), and an omnidirectional microphone, as well as spot microphones set up by venue staff.

As mentioned in section 2.2.3 of this paper, the ESMA in this project used Lee's (2016) proposed addition of Geluso's (2012) MZ pairs to Williams' ESMA. This project's ESMA was composed of four Sennheiser MKH 800 microphones with cardioid polar patterns, with one microphone in each corner of the array. They were oriented at 0°, 90° (facing the stage), 180°, and 270° (facing the back of the venue). Although the ESMA should ideally be a square with all microphones equidistant apart, the tightest possible array for this project was a 12-inch by 14-inch rectangle, rather than the model 12-inch by 12-inch square. This was due to a large amount of equipment being so close together, and rather than overlapping equipment to get the perfect square array, it was preferred for all the microphones to be in the same plane. While the ESMA spacing can vary between 25 and 50 cm, a tighter array was used because it better mimicked a human head and the sound arriving on all sides of it. To create the MZ pairs, a Schoeps CMC6 MK6 microphone with a bi-directional polar pattern pointing up (+) and down (-) was placed in each corner as well. This setup allows sound to be captured in a way that is ideal for 3D audio playback because sound is being captured at all angles and the tightness of the array closer resembles the sound arriving at the ears around a human head. The array was approximately four feet four inches off the ground and was placed in the third row of the audience next to the aisle. It was important not to obstruct the views of audience members. Figures 6 and 7 show diagrams of the array setup.

Just above the modified ESMA, a Schoeps CMC6 MK6 microphone set to an omnidirectional polar pattern was positioned in the center of the array. Right above the omnidirectional

mic was one of the Sennheiser AMBEO microphones, running parallel to the floor and pointed towards the stage like the omnidirectional microphone. The GoPro Fusion was placed over these microphones in the center of the ESMA and stood approximately four feet ten inches off the floor. Photos of the audience setup are represented in Figures 2, 3, and 4. The second Sennheiser AMBEO microphone was positioned along the front and center of the stage, perpendicular to the ground and pointing upwards. It stood approximately three feet and five inches off the floor, just above the height of the stage. The entire setup for the recording process is detailed in Figures 5, 6, and 7.

Figure 2: Photo of the setup in the third row of the audience. From top to bottom: GoPro Fusion camera, Sennheiser AMBEO microphone, omnidirectional microphone for the modified ESMA, the modified ESMA.

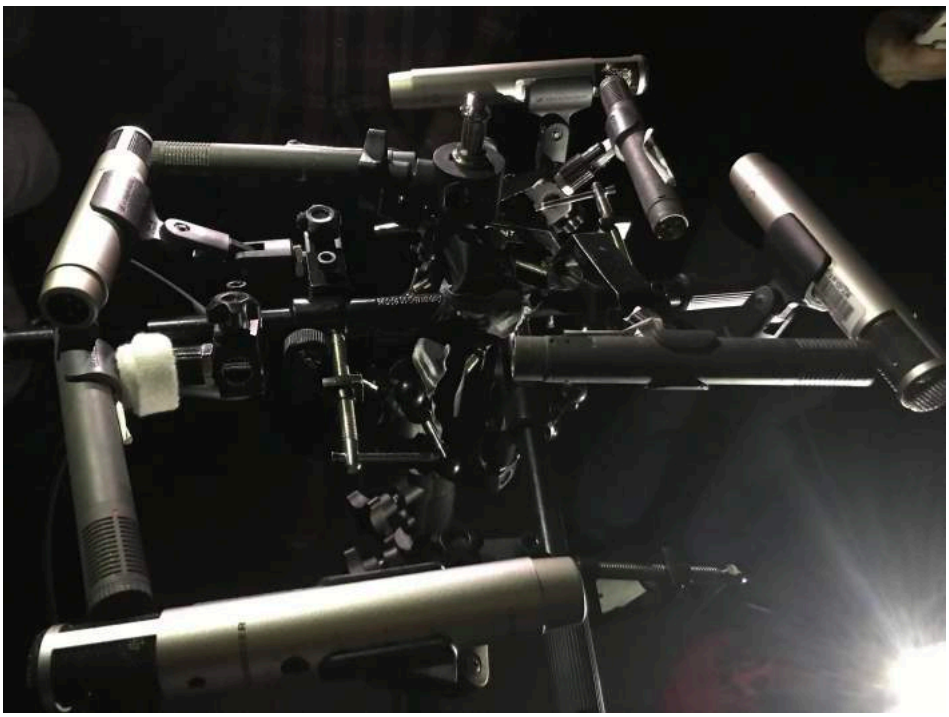

Figure 3: Another view of the modified ESMA. For recording, the microphones were pulled in tighter to form a 12-inch by 14-inch rectangle that stood approximately four feet four inches off the floor.

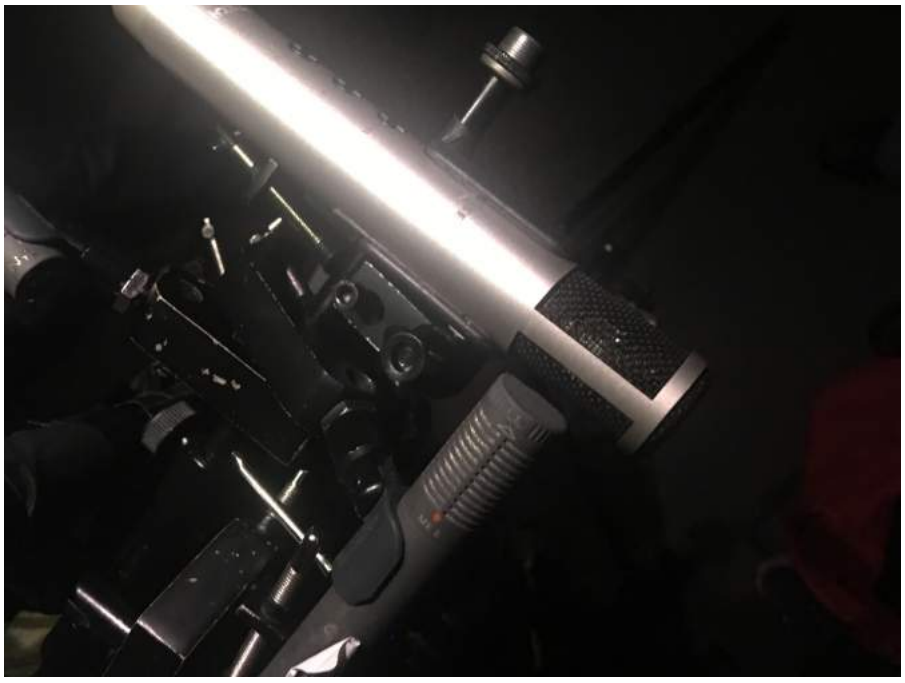

Figure 4: A closer top-view of a corner of the modified ESMA (an MZ pair). The silver microphone is the Sennheiser MKH 800 in a cardioid polar pattern, and the black microphone is the Schoeps CMC6 MK6 in a bi-directional polar pattern, pointing up and down.

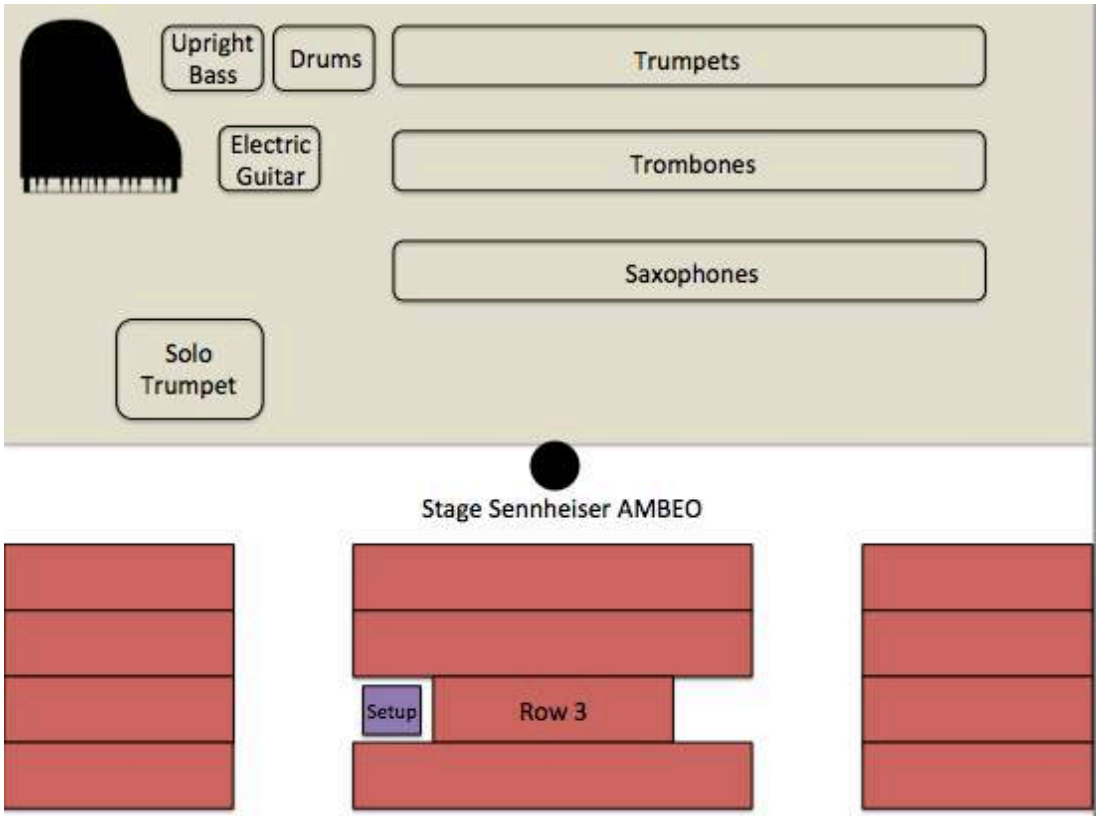

Figure 5: Diagram of the setup in the theater, with the contents of the purple setup square being further detailed in Figures 6 and 7 below.

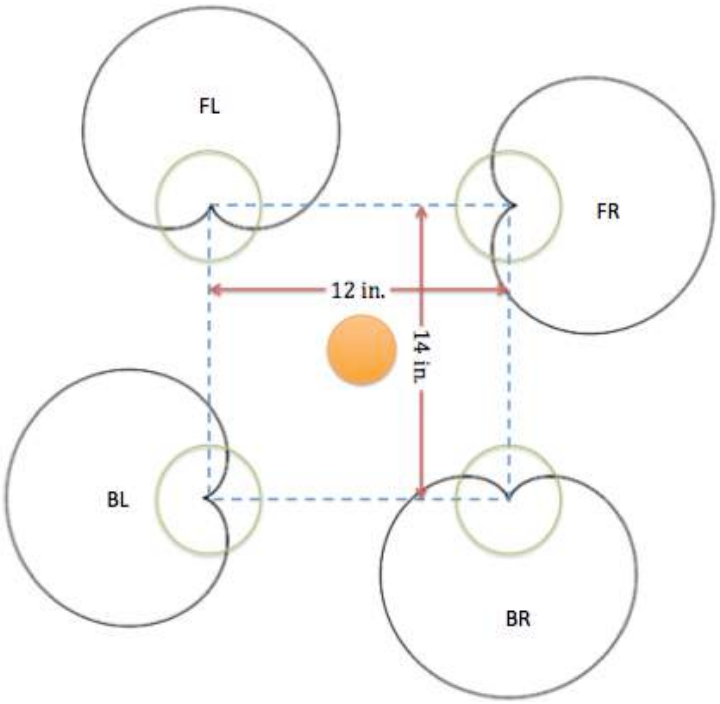

Figure 6: Top-view diagram of the modified ESMA setup. The green outlined circles represent the bi-directional mics, and the orange circle represents where the omni and Sennheiser AMBEO mics were placed, as well as the GoPro Fusion camera.

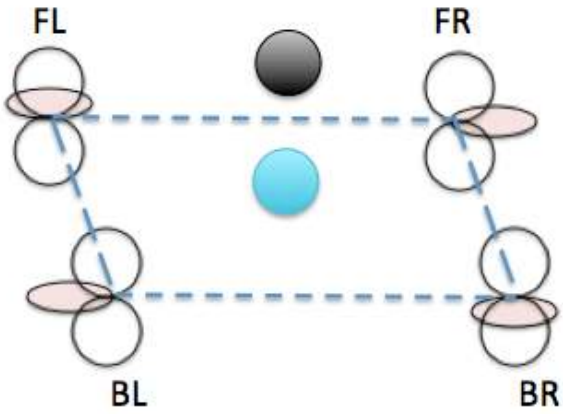

Figure 7: Side-view diagram of the modified ESMA setup, with the blue circle in the middle representing the omnidirectional mic and the black filled circle representing the Sennheiser AMBEO mic. The light pink shaded circles represent the cardioid microphones in the modified ESMA.

Two Zoom F8 Multi-Track Field Recorders and a Tascam DR-680 Field Recorder (8 tracks each) were used to capture the audio. The Zoom F8s captured 96 kHz 24-bit wav files and the Tascam captured 48 kHz 24-bit wav files. The difference in sample rates didn't matter once the audio files were mixed and rendered for the Samsung Gear VR for delivery.

3.2 Mixing

The performance captured had three separate mixes for subjective testing, each mix being a distinct combination of the microphone systems used. Mix 1 was composed of the Sennheiser AMBEO microphone placed in the audience with some added spot mics, Mix 2 was the modified ESMA array with the omnidirectional mic placed in the middle and some added spot mics, and Mix 3 the Sennheiser AMBEO microphone that was placed along the stage. The spot mics used in mixes 1 and 2 were consistent across both mixes. Spot mics were placed by the venue staff and were placed in the front left and right and back left and right of the stage. Facebook 360's program was used. All audio was mixed in ProTools HD.

The first step in the mixing process was to stitch and render the video. Using GoPro Fusion Studio, the video was stitched and rendered for Facebook at 5.2k with 360° audio. The process took several hours for an approximately three and a half minute long video, which is not ideal. This video was then converted to the recommended format without changing the resolution for the Facebook 360 plugin by using the FFmpeg command in Terminal (Audio360, n.d.).

All the audio files were imported into ProTools and were then sorted and grouped by system, and trimmed with fade ins and fade outs. The Sennheiser AMBEO microphones were decoded into B-format and into multichannel tracks via the AMBEO A-B Format Converter plugin. The modified ESMA also had to be decoded. Each corner of the array already had their respective cardioid microphones positioned properly (Left Front (LF), Right Front (RF), Left Back (LB), and Right Back (RB)), but the bi-directional channels needed to be duplicated and these duplications needed to be phase inverted. This was done to get the proper up and down directionality for each corner of the array. The following

equations summarize the modified ESMA decoding process for the bi-directional microphones:

$$LF_{up} = LF_{card} + LF_{bi}$$
$$LF_{down} = LF_{card} - LF_{bi}$$
$$RF_{up} = RF_{card} + RF_{bi}$$
$$RF_{down} = RF_{card} - RF_{bi}$$
$$LB_{up} = LB_{card} + LB_{bi}$$
$$LB_{down} = LB_{card} - LB_{bi}$$
$$RB_{up} = RB_{card} + RB_{bi}$$
$$RB_{down} = RB_{card} - RB_{bi}$$

Equation 2: Equations for decoding the modified ESMA.

In decoding these channels, the cardioid channel and its respective up or down (+ or -) channel were routed to an aux channel in ProTools. For the down facing channel, the duplicated phase inverted channel was used. This process was repeated for each corner of the array, creating a decoded cube with the omnidirectional microphone head-locked in the center of the cube. Figure 8 below represents a diagram of the decoded modified ESMA cube, and Figure 9 further details the routing that matches the decoding equations above.

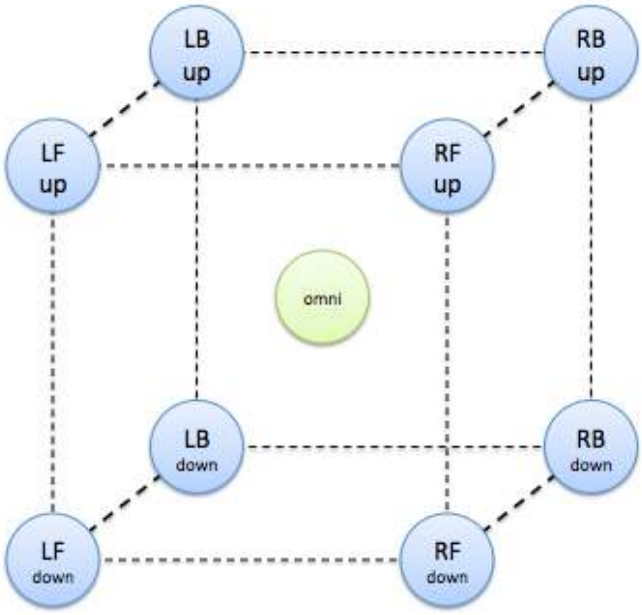

Figure 8: Diagram of the decoded modified ESMA with the omnidirectional microphone placed in the center where the listener's head will be.

Figure 9 shows the modified ESMA tracks used for mixing. The purple, red, and blue tracks on the left are the modified ESMA channels, including the phase inverted bi-directional channels in blue. The channels are sent through busses, which resulted in the equations listed on the previous page. The green and orange tracks are the corners of the decoded modified ESMA cube outlined in Figure 8. The green and orange channels labeled ESMA_Up and ESMA_Down are quad audio tracks with the spatializer plugin applied to them. Their inputs are the routed aux channels previously listed.

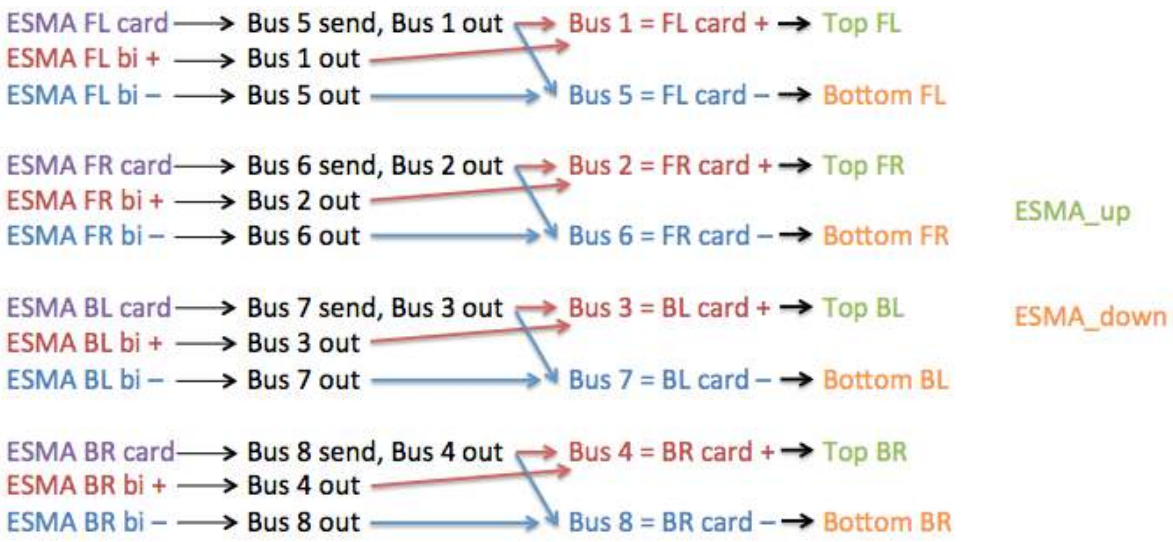

Figure 9: Diagram of the modified ESMA decoding process. The cardioid channels were sent to multiple busses so they could be added to the bi-directional channels. The ESMA_up and ESMA_down tracks are quad tracks that correspond to the blue circles in Figure 6 above.

The AMBEO audience and stage microphones are both quad audio tracks in ProTools with 8 outputs for B-format sounds. The Facebook 360 plugin creates its own tracks in ProTools, including a 3D master channel and a head locked master channel. The omnidirectional channel for the modified ESMA was routed to the non-spatialized head-locked track for more definition. The Facebook 360 Spatializer insert was applied to the tracks that were being routed to the 3D Master channel (ESMA and Sennheiser AMBEO tracks).

Once all of the routing and decoding was done, some basic mixing was done – little EQ and compression were added to the channels. Spot microphones were used in mixes 1 and 2 to further define the audio sources, and the same spot mics and effects were used in both mixes for consistency. Because the VRCE is from the audience's point of view, it was important to analyze how the sound was being delivered to audience members. For mixes 1 and 2, the emphasis was on methods of capturing the ambience of the performance and venue. Spot microphone tracks were captured by venue staff and generously shared for this project. Adding some spot mic channels to the ambience choices (Sennheiser AMBEO in the audience and the modified ESMA + omni) brought a better balance and more natural mix.

When viewing concerts, it is common that loudspeaker arrays are facing the audience and provide more definition to the sound. Audience members hear mostly ambient noise, plus some direct sound from loudspeakers. Mixes 1 and 2 of this project were aimed at recreating that sound. Mix 3 used only the Sennheiser AMBEO microphone placed along the front and center of the stage and did not include spot mics. This was because this microphone captured more direct sound due to its placement in the venue, and also captured some ambience sound because it was not placed on stage as a spot mic for the sole purpose of capturing the direct sound. Theoretically, all three mixes are a combination of direct sound and ambient sound capture, but each mix presents the combination in a different way.

Mixing techniques varied slightly from what was proposed in the live sound section of the Literature Review for several reasons. The capturing implementation of this project didn't follow a more traditional stereo live mix setup, and instead used 3D audio recording techniques combined with some live sound mixing techniques. Additionally, the newer approach with the point of view being set from the audience's point of view rather than it being on stage with the performers as previous similar projects have done diverged from what listeners expected (a more polished mix, as Pras (2016) researched). Though a newer approach, this project did implement common live sound and 3D audio mixing techniques to create the experience presented to test subjects.

The Facebook 360 plugin was used to position the sounds in the virtual soundscape. Once mixing was completed, each separate mix was printed and sent to the Facebook 360 encoder. Mix 2 was the only mix that had a head-locked track (the omni mic in the middle of the modified ESMA). The spatial audio mixes were encoded with the properly formatted (MPEG-4 and H.264) 360° video for Facebook 360 format. The videos were then uploaded to Facebook and downloaded to the Gear VR device. Finally, each of the three mixes was ready for delivery.

3.3 Delivery

Audio was presented to test subjects binaurally over Sennheiser HD 650 headphones for optimal audio quality because binaural 3D playback is typical for VR. Audio was head tracked and linked to the video of the performance and presented to test subjects on the Samsung Gear VR Virtual Reality Headset for Samsung Galaxy. The Samsung Gear VR is a common device used for virtual reality applications. The Facebook 360 plugin implements an average HRTF measurement and head tracking, though the Facebook 360 documentation does not elaborate on how this is done.

3.2 Testing

There was an open call for New York University music technology students, both graduates and undergraduates, and staff to volunteer to be test subjects for this project. This presented subjects with various musical backgrounds and musical experience who have at least basic knowledge of music technology. Testing was carried out on New York University Steinhardt's campus, specifically in music technology studios D1 and E. A total of 19 subjects participated. The performance being evaluated was approximately three and a half minutes long and the three mixes of this performance were presented to test subjects. This resulted in at about 10 minutes of playback time, with additional time being used to explain the test and testing equipment, and for subjects to fill out the preliminary survey provided to them. In total, the test ran from around 20 minutes to no more than 30 minutes in length. This helped prevent listener fatigue.

As mentioned, test subjects wore the Samsung Gear VR headset and Sennheiser HD 650 headphones to carry out the test. The three mixes were presented to them. After each mix, a couple of basic questions were asked, such as "How natural do you find the audio mix?" and "On a scale of 1 to 5 with 5 being fully immersed, how immersed do you feel in this performance?". A basic definition of "immersed" was provided as how much you feel the sound and video is surrounding you. The same questions were asked for each mix. After all

three mixes had been presented to them, test subjects were asked to compare the mixes and were asked more subjective questions such as “Which performance do you prefer?” and “Which performance do you find to be the most immersive?”. This survey was carried out on paper and collected for evaluation. No records were shared and no personal information was stored. All testing met IRB standards. Once subjective testing was over, the data collected was analyzed.

4.0 Results

Of the 19 total subjective testing participants, 14 said they were familiar with virtual reality in some way, 11 rated themselves above average (a 4 or higher out of a 5 point scale, with 5 being the greatest) in technical ear training skills, and 12 rated themselves above average in music technology skills.

When subjects rated each individual audio mix in relation to immersion, naturalness, and quality of the audio, and their overall impression, the average score across the three mixes was similar. However, when asked to rank the mixes against each other, mix 2 (the modified ESMA with the omnidirectional microphone in the center and some spot mics) and mix 3 (the Sennheiser AMBEO microphone places along the front and center of the stage) were the strong contenders for favorite mix, immersion, and audio quality. When comparing mixes 2 and 3, there was an almost even split, with mix 2 being slightly ranked higher in terms of audio quality. These results are represented in Figure 10 below.

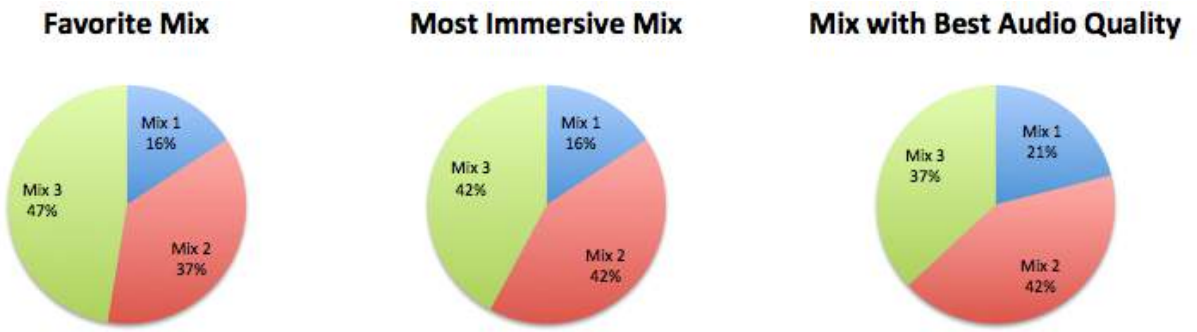

Figure 10: Pie charts representing the overall results of ranking the three mixes against each other. Mix 3 was chosen as the favorite mix just over mix 2, mix 2 was ranked as having slightly better audio quality than mix 3, and mixes 2 and 3 were equally chosen as the most immersive mix.

Subjective testing results are represented by Figures 11, 12, 13, and 14 below. The results for each participant is compared to their self-reported ear training and music technology skill levels to provide better insight as to why some participants may have preferred one mix over the other.

Because mix 2 was more ambient sound with some direct sound added for clarity and mix 3 was more direct sound with minimal ambient sound, and both were preferred amongst listeners, the results of this subjective testing are intriguing.

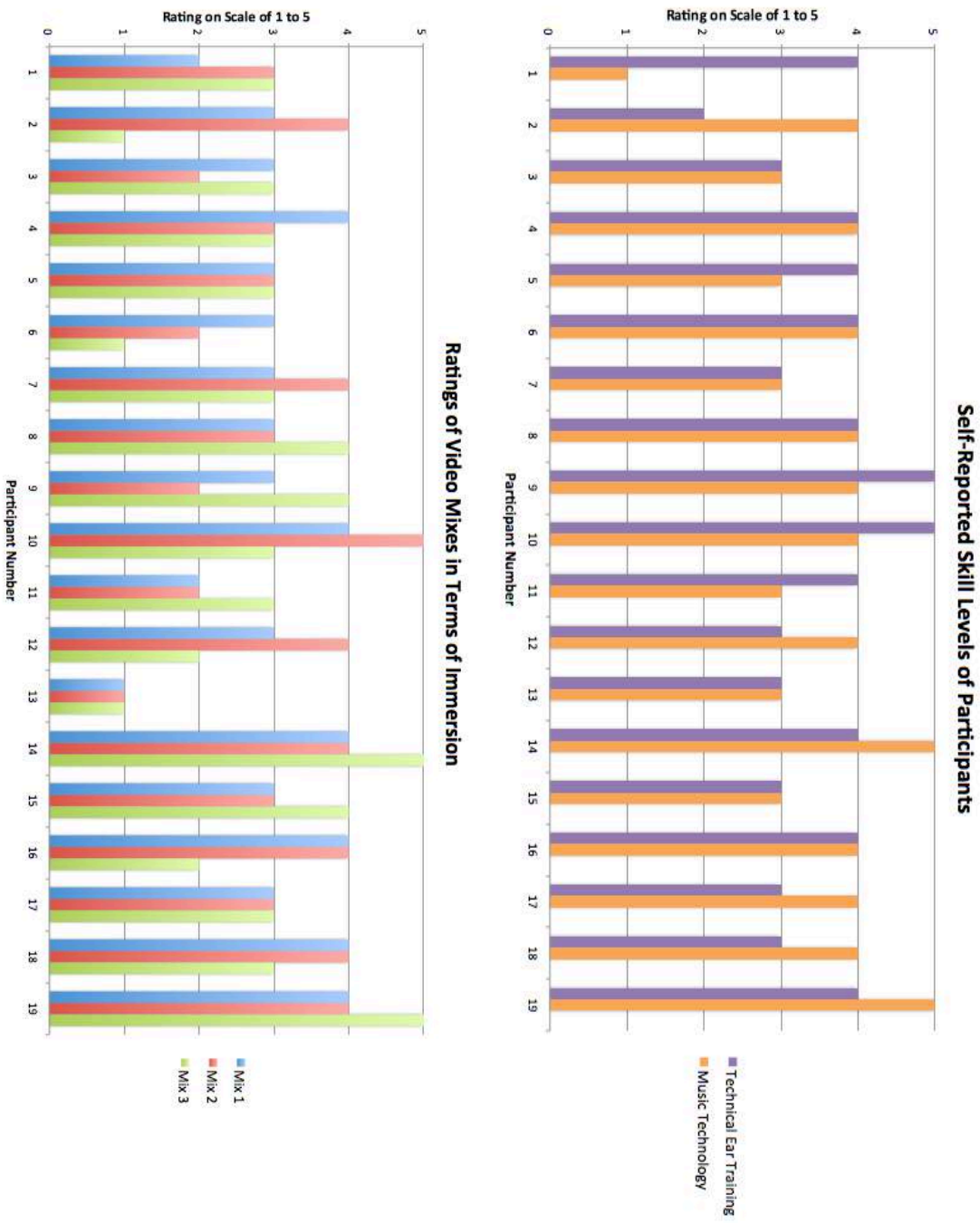

Figure 11: Graph representing subjects' responses to which mix was the most immersive to them, lined up with their self-reported skill level in music technology and technical ear training.

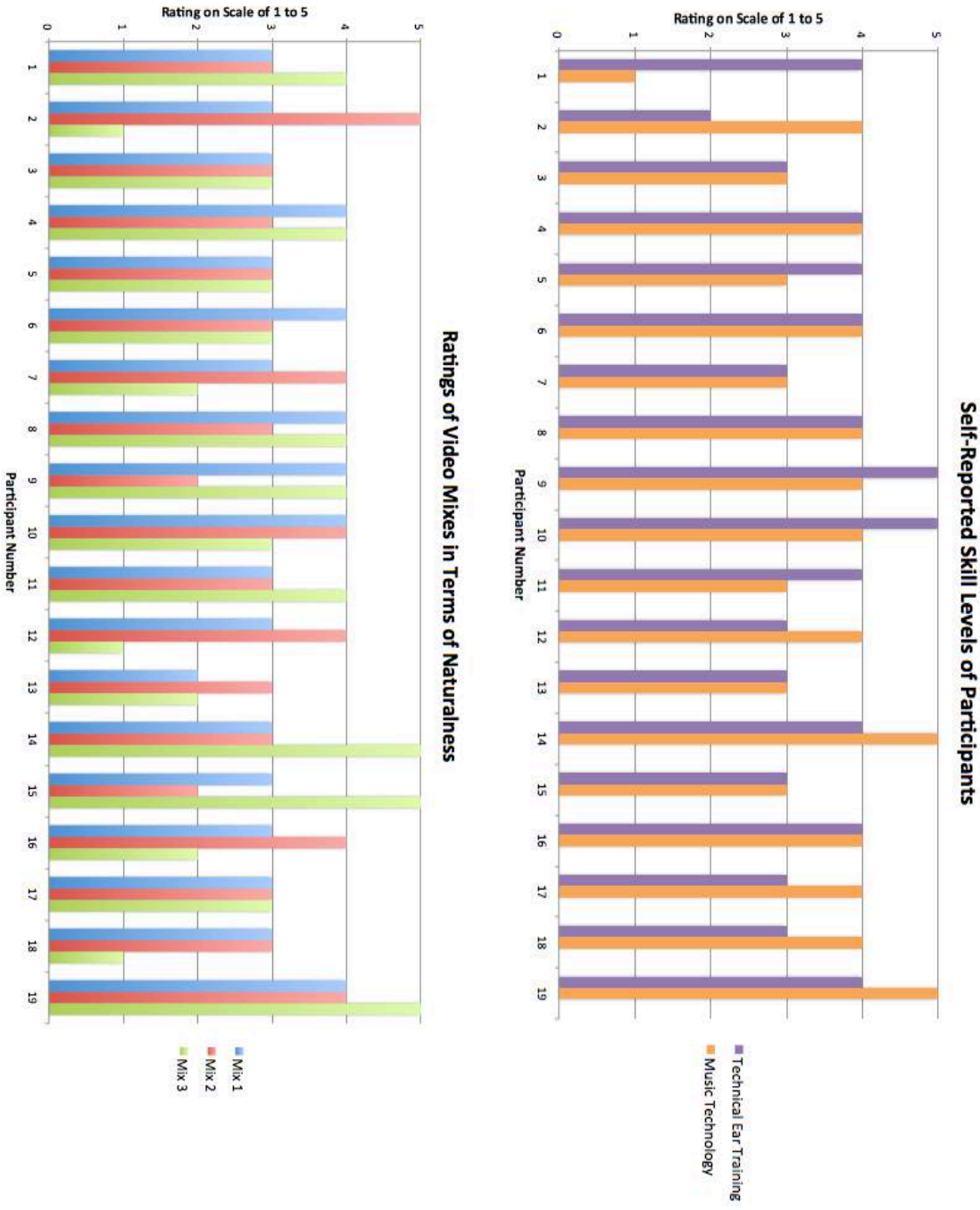

Figure 12: Graph representing subjects' responses to which mix was the most natural sounding to them, lined up with their self-reported skill level in music technology and technical ear training.

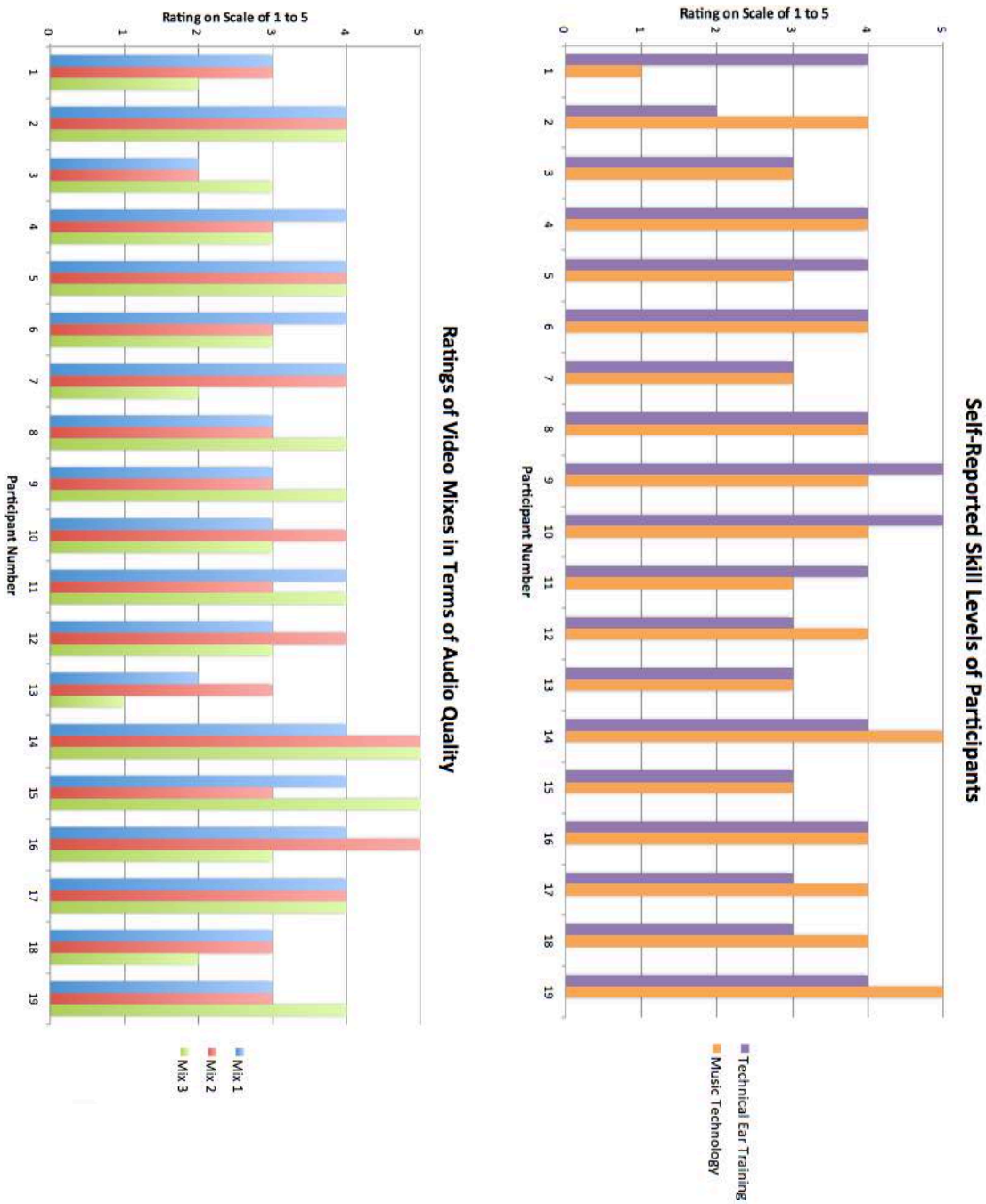

Figure 13: Graph representing subjects' responses to which mix had the best audio quality, lined up with their self-reported skill level in music technology and technical ear training.

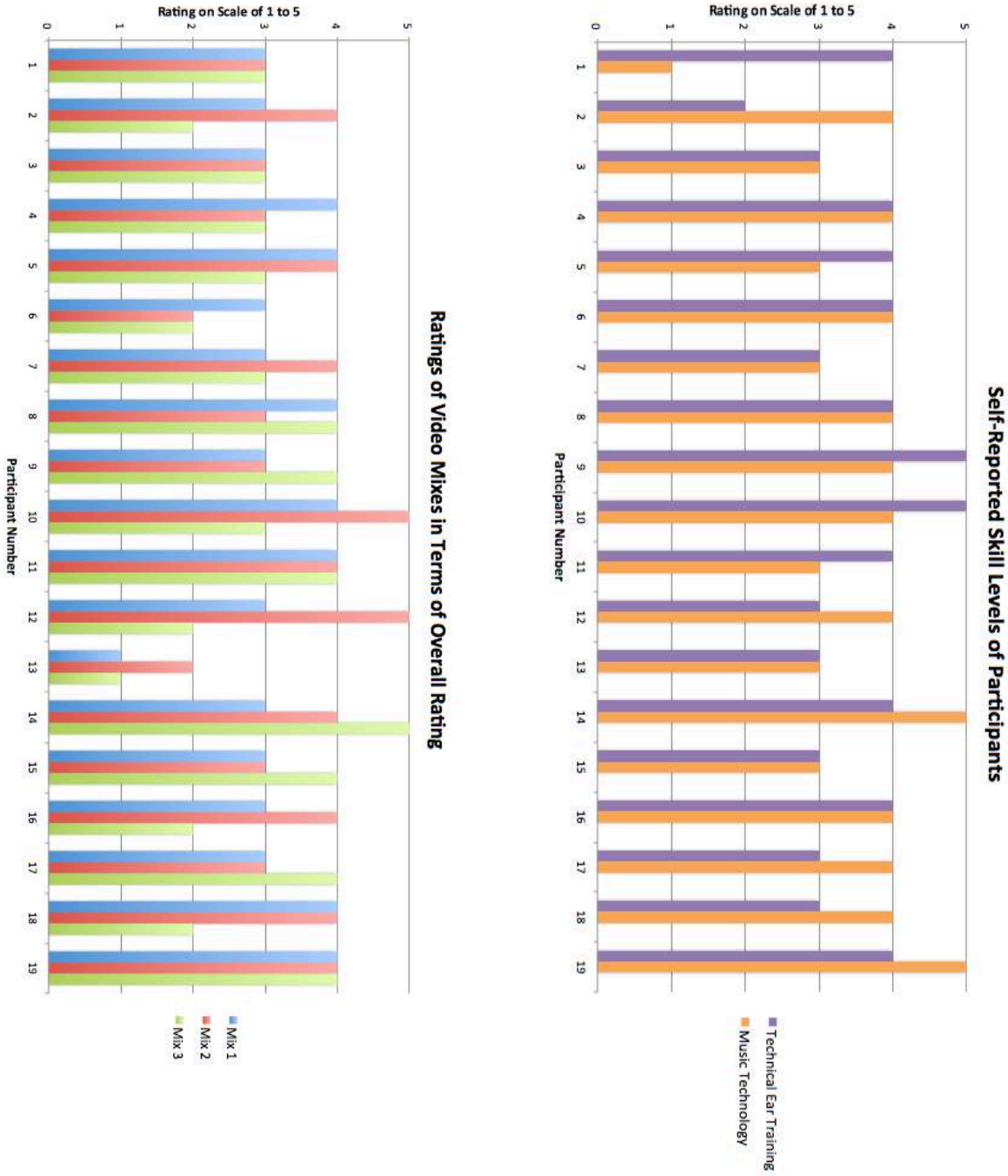

Figure 14: Graph representing subjects' responses to what their overall perception of each mix was, lined up with their self-reported skill level in music technology and technical ear training.

5.0 Analysis and Discussion

It is intriguing that mixes 2 and 3, although being very different in terms of mixing, were preferred almost evenly across subjects because when the test was over, subjects tended to be vocal about why they were against the mix they did not prefer. Mix 1 was preferred and ranked best in terms of audio quality and immersion by very few people (less than 22%), so while it would be one of the simplest and most compact equipment setups, it is probably not a good choice for VR concert experiences, particularly if the goal is to envelop the listener.

While the modified ESMA is the most complicated in terms of equipment setup, it is still a strong contender for virtual reality concert experiences. Some participants found the modified ESMA mix to be slightly more difficult to localize sounds, though localization wasn't the primary goal of this array. Rather, the goal of the modified ESMA in this project was to best recreate the sound around a listener's head if they were in the audience (when created by Williams, the ESMA is meant to provide better localization). However, capturing audio from the audience's perspective for virtual reality purposes is rather uncommon and is not typically what listeners will expect to hear due to what they are already accustomed to. This is likely the reason why mix 3 was also a strong contender, despite it being very direct, especially with the solo trumpet being placed where it was (front stage right). It could be argued that mix 3 is more of what listeners would expect to hear at a live concert.

In addition, previous research suggests that "spatial enhancement features are rather less important to listeners than other aspects of sound quality," (Rumsey, 2013) and that many audiophiles will agree that "a concert is about expression whereas a recording is about perfection," (Blier-Carruthers, 2010). This, along with the Pras (2016) experiment determining that listeners prefer polished and edited audio, could further add to the argument that mix 2, with its more ambient sounds, was not what listeners expected upon watching a VR video and could have contributed to more test subjects choosing mix 3, which had more direct sound than the other mixes presented. However, participants who preferred mix 2 over mix 3 commented that the harshness and forwardness of the solo

trumpet made it difficult for them to focus on other aspects of the performance and made the mix seem less balanced. Further evaluation should be done to determine why participants would prefer one mix over the other, both by asking them and by evaluating certain psychoacoustic aspects that could contribute to these results.

The primary commentary from participants was actually an evaluation of the video quality rather than the actual audio content. “Earlier studies have shown that when sound and picture do not form a single spatiotemporal coherent stream, the overall perceived quality is degraded,” (Beerends & De Caluwe, 1999). Beerends & De Caluwe (1999) also mention that when an average is taken, no clear influence of perceived audio quality on perceived video quality (and the other way around) is determined, though there is a small, significant mutual influence, particularly “for the influence of the view quality on the perceived audio quality”. They also mention that “the influence of video on audio is stronger than the reverse”, (Beerends & De Caluwe, 1999). They suggest this could be because audio quality varies less than video quality.

This correlation and the fact that the video quality of this project was not ideal (musicians’ faces could not even be made out, which is not realistic) could suggest the results from this experiment could be different if presented again with a better quality video, such as from a more expensive, less consumer-accessible 360° camera like the Nokia OZO, which has been discontinued (to focus on other areas, but could be because the camera comes with a hefty price tag and a complicated stitching process, even though the video quality is superb). Still, “the video quality will contribute significantly to the subjectively perceived audio quality . . . the video quality dominates the overall perceived audiovisual quality,” (Beerends & De Caluwe, 1999).

Participants mentioned that being surrounded by the audience and seeing audience members did add a factor of believability when evaluating if they felt like they attended the performance or not. Despite video quality being an issue, all but one participant reported that they would or would possibly view concerts performances in a similar way again. More people chose that they would “maybe” view performances like this in the future than

those who chose a definitive “yes”, and it is concluded that this is most likely because of the video quality of the performance. However, because the main focus of this project was to explore how to properly capture 3D audio recordings for virtual reality concert experiences, this experiment could be determined as moderately successful, with two viable recording setups to choose from. Still, much work is yet to be done and this project only touches upon what the future holds in this area of music technology.

6.0 Conclusion and Future Work

The Virtual Reality Concert Experience (VRCE) implemented in this paper is designed to explore newer applications of virtual reality beyond those of gaming and immersive training applications and improve upon prior research and implementations in the crossover area of 3D audio, live sound, and virtual reality. As touring concert shows become more elaborate and expensive to make up for artists' lack of revenue from physical media sales, concerts are becoming less accessible to the general population. A virtual reality concert experience could provide more accessibility to concertgoers as well as an additional source of revenue for artists.

Incorporating 3D and immersive audio with virtual reality adds a sense of believability and envelopment and more accurately represents the way humans hear and perceive sound naturally, resulting in a more engaging and enjoyable virtual reality experience. With much to explore and improve upon, this project's aim was to create some basic guidelines and a solid starting point for creating an immersive and believable virtual reality concert experience from capture to delivery.

A live jazz performance was captured in 360° using virtual reality and 3D audio recording techniques. A modified Equal Segment Microphone Array (ESMA) composed of four cardioid microphones and corresponding bi-directional microphones (pointing up and down) arranged in a small 12-inch by 14-inch rectangle that was positioned slightly above a sitting audience's line of sight. In the middle of this array, an omnidirectional microphone and a Sennheiser AMBEO ambisonic microphone were placed for additional capture. On top of this setup, a GoPro Fusion camera was mounted to capture a 360° video of the performance. This setup was positioned in the third row of the venue, slightly off center and along the stage right aisle. A second Sennheiser AMBEO microphone was placed along the front and center of the stage to capture both direct and ambient sounds of the performance.

Three separate audio mixes were implemented and presented to a volunteer group of subjects to determine which captured sound technique was preferred among listeners. Mix 1 was the Sennheiser AMBEO microphone placed in the audience, with some spot mics added for clarity and balance. Mix 2 was the modified ESMA with the omni mic head-locked in the middle, with the same spot mic implementation as in Mix 1 for consistency. These two mixes explored methods to capture the ambient sound of a concert from a concertgoer's perspective. Mix 3 was the Sennheiser AMBEO microphone placed along the front and center of the stage and was set up as an alternative to the previous mixes' balance of direct and ambient sound.

Upon comparing these three audio mixes, which were all presented with the same video clip, it was concluded that listeners prefer the audio two mixes: the modified ESMA with an omnidirectional microphone placed in the center of the array of MZ pairs and spot microphones added for additional definition, and the Sennheiser AMBEO microphone placed along the front and center of the stage. Though a definitive method of capture wasn't favored among testing subjects, this project as a whole is meant to create a guideline for the ideal capture and reproduction of a music concert for virtual reality experiences.

6.1 Limitations

Though this virtual reality concert experience (VRCE) project provides a solid template for future similar projects, several limitations prevented the project from reaching its full potential. Ideally, the VRCE would be used in larger venues with more interactive performers than the one recorded in this project. The venue used for this project also had less crowd noise than stadium shows would have, for instance, and therefore unwanted venue noise such as the air conditioner could be heard throughout the mixes. A higher quality 360° camera should be used as well. Though the GoPro Fusion provided higher quality than other VR cameras and was very compatible for the scope of this project, it also did a poor job capturing the stage lighting, which skewed the image of the performers on stage (the stage lighting blurred musicians' faces, and testing subjects commented on this

often), leading to a decrease in believability. Additionally, the camera captured the audio engineers monitoring the sound levels and the microphone array, which also could have contributed to decreased believability.

The venue and staff were very cooperative in the addition of this project's equipment to their event, though some compromises had to be made. It would have been ideal to have had the modified ESMA and other audience microphones placed in the center of the rows of seats as opposed to right along the aisle. Ideal placement would be close enough to feel engaged in the performance, but back far enough to capture the whole stage and its lighting and special effects. The placement near the aisle presented some issues in the mixing process, particularly those involving localization with the stage Sennheiser AMBEO microphone and spot microphones. Additionally, the investigator of this project did not have control of the spot microphone placement and setup. The venue staff set up the spot mics and shared their recordings, and the spot mics were not placed on each instrument, as they should have been. The spot mics did not provide enough information to create a fully balanced mix.

Finally, the mixing process for this project was complex and time-consuming, which may not be ideal for industry professionals on a deadline. Furthermore, more popular music genres would have been ideal to explore, such as pop and rock performances, which are more likely to be large, intricate, and more interactive shows that some concertgoers may not have access to. Despite all of these challenges, this VRCE project was carried out with the best equipment and resources available to the investigator and the project provides some solid groundwork for future implementations.

6.2 Future Work

The project implemented in this paper is considered small-scale when compared to the performances this project would primarily benefit. For instance, this VRCE would ideally be better suited for larger venues with more complex lighting and stage design and larger and

more interactive audiences. Future work would experiment with this project's implementation and make any necessary adjustments to better fit the specific circumstances of different sized venues. Future work would also address more of the limitations previously mentioned, such as adjusting placement of the equipment (for instance, in the center of the audience as opposed to the aisle or near the live sound engineer and the venue's mixing board), and exploring 360° cameras that are better suited for low-lighting environments and that better capture subjects when under stage lighting. Additionally, other genres of music should be explored, and future subjects should be asked a more open-ended question of why they prefer one mix over the other so that a possible psychological explanation for distinguishing choices between mix 2 and mix 3 can be determined.

Other future work goals include incorporating bass vibrations and rendering moving sources. "Tactile bass vibrations can . . . be added [with spatial audio] to increase the sense of realism," (Rumsey, 2011). With this project, bass vibrations could be implemented by placing a device on listeners' chairs that helps mimic the low-frequency vibrations many concertgoers experience, especially if they are standing near loudspeakers. Particularly with concert performances and musicians moving around on stage, moving sources should be evaluated. For instance, if a VRCE user is watching a lead singer move from stage left to stage right, and the audience point of view isn't shifting along with it, the sound source of the lead singer should move and be mirrored in the user's headphones. This should be done in an "efficient and smoothly changing way . . . many of the extant approaches involve interpolation or crossfading of HRTFs for different discrete locations in order to synthesize intermediate locations where no stored HRTF exists," (Rumsey, 2008), though more work is to be done in this area as well.

It is a long-term goal for this project to be improved and prepared for live streaming applications, though it is likely that this goal lies far in the future. Though some projects have already explored the area of live streaming 3D audio and virtual reality, improvements first need to be made in the area of capturing musical concerts for virtual

reality and 3D audio purposes. Still, the projects that explore live streaming applications should be mentioned.

In one application, basic stereo image streaming factors were applied to immersive audio streaming applications. As immersive audio can be transmitted in channel-based, scene-based, or object-based formats. In channel-based systems, each audio channel represents a different loudspeaker. Scene-based formats are better known as previously mentioned High Order Ambisonics, and object-based formats were covered in detail back in section 2.3.2 of this paper. Head tracking and HRTF technology should also be considered, and if VR headsets don't already implement this, it could be difficult to accurately represent the 3D audio soundscape. It is suggested by Kares & Larcher (2016) to use a static binaural stream for immersive audio "because it can build on current infrastructure". Object-based audio, MPEG-H, and Dolby Atmos are the better candidates currently, though live streaming immersive audio content is still considered a work in progress (Kares & Larcher, 2016). "There is [currently] no publicly available commercial platform that enables live streaming of 360 video with ambisonics. But as this is already available in non-realtime applications, the landscape surrounding this field of audio production may change in the near future", (Jacuzzi, Brazzola, & Kares, 2017).

6.3 Final Remarks

With this Virtual Reality Concert Experience project, it is anticipated that the methods and recording techniques used will lay the groundwork for future and improved implementations. Virtual reality concert experiences could provide an additional source of revenue for artists and make concerts more accessible. Though much work remains to be done, and while technology in the areas of virtual reality and 3D audio is rapidly improving, this project hopefully will serve as some modest inspiration and insight into what future applications of music technology will look like.

Bibliography

Altman, M., Krauss, K., Susal, J., & Tsingos, N. (2016). Immersive audio for VR, presented at 2016 AES Conference on Audio for Virtual and Augmented Reality, Los Angeles. Retrieved from *AES*.

Anderton, C. (2016, February). Are you ready for virtual reality audio?. *Pro Sound News*, *38*(2) pp. 27-28. EBSCO Publishing.

Audio360. (n.d.). Video format guidelines. Facebook 360 Spatial Workstation User Guide. Retrieved from https://facebookincubator.github.io/facebook-360-spatial-workstation/KB/VideoFormatGuidelines.html#converting-videos-to-dnxhd-with-iffmpeg.

Bates, E. & Boland, F. (2016). Spatial music, virtual reality, and 360 media, presented at 2016 AES International Conference on Audio for Virtual and Augmented Reality, Los Angeles. AES e-Library.

Beerends, J. & De Caluwe, F. (1999). The influence of video quality on perceived audio quality and vice versa, *Journal of the Audio Engineering Society*, 47(5), pp. 355-362. AES e-Library.

Berg, J. (2009). The contrasting and conflicting definitions of envelopment, presented at the 126th Audio Engineering Society Convention, Munich. AES E-Library.

Blier-Carruthers, A. (2010). Live performance-studio recording: An ethnographic and analytical study of Sir Charles Mackerras. Unpublished doctoral thesis, King's College, University of London.

Clukey, T. (2006). Capturing your sound: A guide to live recording. *Music Educators Journal*, 92, 3, 26. Retrieved from *ProQuest*.

Diaz, R. & Koch, T. (2016). Live panorama and 3D audio streaming to mobile VR, presented at AES International Conference on Headphone Technology, Denmark. AES e-Library.

Floros, A., Kapralos, B. & Moustakas, N. (2016). An augmented reality audio live network for live electroacoustic music concerts, presented at 2016 AES Conference on Audio for Virtual and Augmented Reality, Los Angeles. Retrieved from *AES*.

Gaston, L., Boley, J., Setler, S., & Ratterman, J. (2010). The influence of individual audio impairments on perceived video quality. *E-briefs of the Audio Engineering Society Convention 128, 8151*. AES e-Library.

Geluso, P. (2012). Capturing height: The addition of Z microphones to stereo and surround microphone arrays, presented at the 132nd Audio Engineering Society Convention, Budapest. AES E-Library.

Jacuzzi, G., Brazzola, S., & Kares, J. (2017). Approaching immersive 3D audio broadcast streams of live performances, presented at the 142nd Audio Engineering Society Convention, Berlin. AES E-Library.

Kares, J. & Larcher, V. (2016). Streaming immersive audio content, presented at the Audio Engineering Society Conference on Audio for Virtual and Augmented Reality, Los Angeles. AES E-Library.

Kendall, G. (Winter 1995). A 3-D sound primer: Directional hearing and stereo reproduction. *Computer Music Journal*, 19(4), pp. 23-31.

Le Henaff, G. (2015). From studio to stage, presented at the 139th Audio Engineering Society Convention, New York. AES E-Library.

Lee, H. (2016). Capturing and rendering 360° VR audio using cardioid microphones, presented at the Conference on Audio for Virtual and Augmented Reality, Los Angeles. AES E-Library.

McKee, J. (2017, March 29). Audiences experience 'The Encounter' via headphones. *Detroit Free Press*. Retrieved from http://www.freep.com/story/entertainment/arts/2017/03/29/the-encounter-simon-mcburney-university-musical-society/99739688/.

Nady, J., (2007). Captured live! Simultaneous live sound and recording techniques. *EQ Magazine,* 18(4)*.* Retrieved from *ProQuest*.

NextVR. (2014, December 10). NextVR and Coldplay release virtual reality concert for Samsung Gear VR. *PR Newswire*. Retrieved from http://www.prnewswire.com/news-releases/nextvr-and-coldplay-release-virtual-reality-concert-for-samsung-gear-vr-300007406.html.

Pras, A. (2016). Live vs. edited studio recordings: what do we prefer?, presented at the 141st Audio Engineering Society Convention, Los Angeles. AES E-Library.

Riaz, H., Stiles, M., Armstrong, C., Chadwick, A., Lee, H., & Kearney, G. (2017). Multichannel microphone array recording for popular music production in virtual reality, presented at the 143rd Audio Engineering Society Convention, New York. AES E-Library.

Roginska, A. & Geluso, P. (2018). *Immersive sound: The art and science of binaural and multi-channel audio*. New York, NY: Routledge, an imprint of the Taylor & Francis Group.

Rumsey, F. (2008). Signal processing for 3-D audio, *Journal of the Audio Engineering Society*, 56(7/8), pp. 640-645. AES e-Library.

Rumsey, F. (2011). Spatial audio: Eighty years after Blumlein, *Journal of the Audio Engineering Society*, 61(6), pp. 474-478. AES e-Library.

Rumsey, F. (2013). Spatial audio processing: Upmix, downmix, shake it all about, *Journal of the Audio Engineering Society*, 59(1/2), pp. 57-65. AES e-Library.

Rumsey, F. (2016a). Immersive audio: Objects, mixing, and rendering, *Journal of the Audio Engineering Society*, 64(7/8), pp. 584-588. AES e-Library.

Rumsey, F. (2016b). Headphone technology: Personalization, perception, preference, *Journal of the Audio Engineering Society*, 64(11), pp. 940-944. AES e-Library.

Rumsey, F. (2017a). Binaural audio and virtual acoustics, *Journal of the Audio Engineering Society*, 65(6), pp. 524-528. AES e-Library.

Rumsey, F. (2017b). Broadcast and streaming: Immersive audio, objects, and OTT TV, *Journal of the Audio Engineering Society*, 65(4), pp. 338-341. AES e-Library.

Rumsey, F. (2017c). Recording and reproduction: Studio myths and new technology, *Journal of the Audio Engineering Society*, 65(9), pp. 776-780. AES e-Library.

Stefanakis, N. & Mouchtaris, A. (2016). Capturing and reproduction of a crowded sound scene using a circular microphone array. *European Signal Processing Conference 24*, pp. 1673-1677. IEEE Xplore.

Watercutter, A. (2017). Jonathan Demme's 'Stop Making Sense' is still the concert film all others try to be. *Wired*. Retrieved from https://www.wired.com/2017/04/jonathan-demme-stop-making-sense/.

Williams, M. (2003). Multichannel sound recording practice using microphone arrays, presented at the 24th International Conference on Multichannel Audio. AES e-Library.

Young, C. (2014). State of the industry: Sound reinforcement. *Pro Sound News*, 36(10), pp. 53.

Appendix A: Glossary

- Ambisonics: A spherical surround-sound application that has equal distribution across all sound sources and is independent of playback systems.
- Azimuth: The angle from the listener to the sound source in the horizontal plane.
- Equal Segment Microphone Array (ESMA): For the purposes of this paper, the ESMA is composed of four cardioid microphones with corresponding bi-directional microphones (pointing up and down) arranged in a small 12-inch by 14-inch rectangle, creating M/S pairs in each corner of the rectangle.
- Head-Related Transfer Function (HRTF): A measurement of how the ear perceives sound as it arrives to the ear canal; is specific to each person, as head shape, ears, shoulders and other factors are unique to individuals and affect how sound arrives to the ears.
- Immersive: Immersion refers to how much you feel the sound is around you, or how enveloped you feel in the audio.
- Virtual Soundscape: The virtual sound environment created by 3D sound.

Appendix B: Recording Documentation

Figure 15: An alternate view of the recording setup. The ESMA and GoPro Fusion camera are clearly visible.

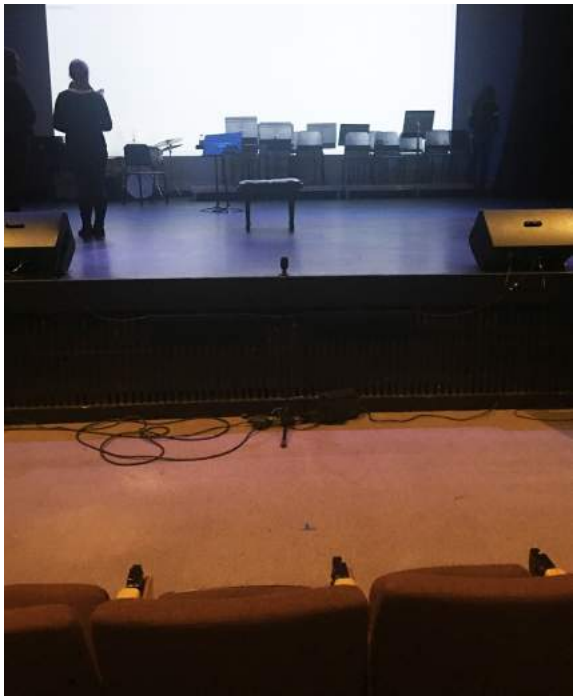

Figure 16: The Sennheiser AMBEO microphone positioned along the front and center of the stage, standing approximately 3 feet 5 inches off the floor.