Title

Psychoacoustic Studies of Music Spatialization Strategies in Different Listening Contexts

Author

Aswathanarayana, Shashank

Publication Date

2022

UNIVERSITY OF CALIFORNIA

Santa Barbara

Psychoacoustic Studies of Music Spatialization Strategies in Different Listening Contexts

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy
in Media Arts and Technology

by

Shashank Aswathanarayana

Committee in charge:

Professor Curtis Roads, Chair

Professor Yon Visell

Professor Andrés Cabrera

Professor Braxton Boren

December 2022

The dissertation of Shashank Aswathanarayana is approved.

Braxton Boren

Andrés Cabrera

Yon Visell

Curtis Roads, Committee Chair

December 2022

Psychoacoustic Studies of Music Spatialization Strategies

in Different Listening Contexts

by

Shashank Aswathanarayana

ACKNOWLEDGMENTS

First of all, a huge thanks to my Appa and Amma. Without their encouragement and support, I would never have been able to enter the field of Music Technology all those years ago.

I am very grateful to my committee chair, Prof. Curtis Roads for allowing me complete freedom to pursue the research questions that interested me and gave me all the support and guidance throughout this journey. This PhD would not have been possible without him. Thank you also to the rest of my committee members, Professors Yon Visell, Andrés Cabrera and Braxton Boren. They were always approachable and guided me so ably.

Thankful for the equipment and facilities support from Center for Research in Electronic Art Technology (CREATE), The Allosphere Research Group, Hochshule für Gestaltgung and the Zentrum für Kunst und Medien (ZKM). This PhD would not have been possible without this support.

Thank you to my childhood friend and now collaborator, AD Narayan. It's been great fun working on the prototype with him. Thanks to Professor Paul Modler for the support and guidance during my time in Germany and thanks to Professor Paul Geluso for providing me with the audio files to conduct the experiments. Thank also to Areti Andreopoulou and Karl Yerkes for being such wonderful people I could brainstorm ideas with.

I'm also extremely grateful to Scott Ji and Zaveeni Ji who have been like parents to me during my time here in Santa Barbara.

Finally, there are many friends to thank from Santa Barbara to New York to Washington DC to Bengaluru who have helped and supported me during my good and bad times. All my lovely friends, Meghana Chandra, Nachiket Panse, Pratik Shah, Kruthika Sivakumar, Nitye Sood, Varsha Srinivasan, Anisha Thomas, a huge thanks!

Curriculum Vitae

SHASHANK ASWATHANARAYANA

EDUCATION

2022	Ph.D. in Media Arts and Technology, University of California, Santa Barbara
2014	M.M in Music Technology, New York University
2012	B.E in Electronics and Communication Engineering, Visvesvaraya Technological University

EXPERIENCE

2016 - 2022	Teaching Assistant, Media Arts and Technology & Physics, University of California, Santa Barbara
2014 – 2016	Research Engineer, Humtap Inc., San Francisco
2013 – 2014	Research Assistant, Music Technology, New York University

PUBLICATIONS

Aswathanarayana, S. (2021, May). Comparison of Spatialization Techniques with Different Music Genres II. In Audio Engineering Society Convention 150. Audio Engineering Society.

Aswathanarayana, S. (2020, October). Comparison of Spatialization Techniques with Different Music Genres. In Audio Engineering Society Convention 149. Audio Engineering Society.

Aswathanarayana, S. (2017, May). Effect of a Known Environment on the Estimation of Sound Source Distance. In Audio Engineering Society Convention 142. Audio Engineering Society.

Aswathanarayana, S., & Roginska, A. (2014). I Hear Bangalore3D: Capture and Reproduction of urban sounds of Bangalore using an Ambisonic Microphone. Proceedings of the International Conference on Auditory Display (ICAD), New York.

AWARDS

Corwin Award for Solo Percussion in 2019 and 2017 at the University of California, Santa Barbara.

Received the Baden Württemberg Stipendium to conduct research in 3D Audio at the Hochshule für Gestaltung, Karlsruhe, Germany during the summer of 2017.

Achiever Award at Sri Bhagawan Mahaveer Jain College of Engineering in 2010 for outstanding achievements during the year 2009-2010.

ABSTRACT

Psychoacoustic Studies on Music Spatialization strategies in different listening contexts

by

Shashank Aswathanarayana

The use of space in music is a complex issue which involves several different, yet interrelated factors. The technical means of performance, the sonic material, and the overall musical aesthetic should all work together to produce a truly immersive experience and one which the listener can comprehend the spatial impression as independent and yet musically significant. Performances of spatial music typically involve a distributed audience and often take place in an acoustically reverberant space. This situation is quite different from the case of a single listener at home, or the composer in the studio or in a virtual reality setting with head mounted displays. As a result, spatial strategies which are effective in one context may not be transferable to another context. This thesis attempts to study the psychoacoustic factors of various music spatialization strategies in different contexts and in that process attempts to answer the questions, what psychoacoustic factors affect listening to music in non-virtual listening contexts and how could they transform when listening in a virtual listening context? How do these psychoacoustic factors work?

This question is answered through three psychoacoustic studies. The first one on Sound Source Distance, the second and third are comparative studies of music spatialization algorithms with different music genres – a loudspeaker study and a binaural analysis study. The last part of the thesis is to build a prototype of a virtual concert hall that uses some of the learnings of the aforementioned studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. Introduction

Spatialization of sound is becoming increasingly popular. As Roads notes, “the art of spatialization has emerged as one of the most important topics in composition today” [1]. Spatialization as a concept isn’t widely prevalent in all forms of music, almost non-existent in genres like Indian music, but composers of all genres, including Indian music (as I talk about my own personal motivations in the next section) are beginning to warm up to the idea of adding a spatial element to their compositions. The use of space in music is a complex issue which involves several different, yet interrelated factors. The technical means of performance, the sonic material, and the overall musical aesthetic should all work together to produce a truly immersive experience and one which the listener can comprehend the spatial impression as independent and yet musically significant. Performances of spatial music typically involve a distributed audience and often take place in an acoustically reverberant space. This situation is quite different from the case of a single listener at home, or the composer in the studio or in a virtual reality setting with head mounted displays. As a result, spatial strategies which are effective in one context may not be transferable to another context. Thus, the study of spatialization of music in a performance venue/concert hall is quite different to the study of spatialization in a binaural listening environment. However, research is emerging which shows how these two seemingly different settings can be merged in virtual environments.

An efficient workflow for sound spatialization requires structure, flexibility, and interoperability across all involved components. In addition, one requires a solid understanding of the spatialization systems, their capabilities, and limitations. These factors are looked at with particular attention towards the creation of an effective virtual environment. An effective virtual environment can be defined as one which is fully immersive and the listening environment as close to the natural listening environment as possible. To create such an environment, the factors such as room acoustics which affect the listener must be carefully examined.

1.1 Motivation

My first encounter with the world of electronic music and spatial music was through the works of Joel Chadabe [2] and Karlheinz Stockhausen. This opened me to a whole new world of audio. Up until then, I had only been a performer playing Indian classical and semi-classical music. It led to me thinking with a very fresh new perspective and it changed me as a musician. Now, I identify myself as a musician who primarily plays Indian Classical music but has also spent a significant amount of time exploring music from the middle east as well as attempting to create fusion music. Within the large realm of fusion music, I have explored creating pieces that have influences from Indian music, western pop/dance music, electronic music, and middle east music.

Indian Classical music is heavily improvisational in nature and therefore, trains a performer to inherently be a composer as well. But the composition doesn't happen with a blank slate. There is some form of base composition over which the artist learns to improvise and show his/her creativity on stage. With me being introduced electronic music and spatial music, turned a new leaf to my creative and compositional process and thinking. Some of the questions I have started asking myself when I compose nowadays are: Can I add any sampled sounds to the composition to add more depth to the composition? Can I organize the sounds spatially to give a more immersive experience? Can I use any spatializing techniques to enhance the experience? When amazing effects can be achieved with spatialization of electronic sounds, why can't we do it with Indian classical music? Will it not add a new dimension and make it more dynamic? These questions formed the primary motivation to my research pursuit and the base to this dissertation as well.

1.2 The research question

As I began to explore the field of spatial audio in search of the answers to the questions mentioned in the previous section, I realized that there is indeed a gap in the field and there is quite a lot of work that can be done. The topic of sound source distance was not fully explored, and distance effects are a key aspect of spatial compositions and the first part of my research focused on that. There was also much work to be done studying spatialization techniques and different music genres. There are so many variables that come into the picture when different genres are involved as we add many layers of complexity. The instruments used are different that results in timbral changes in the audio. Different genres' approach to pitch and rhythm change. As an example, Western classical music works a lot with harmony and chords whereas the concept of harmony is not present in Indian classical music, which works or horizontal expansion of melody rather than adding harmonies to the base note.

With the onset of the pandemic in early 2020, everything moved virtual, and this added a new challenge to the lives of musicians, composers and performers. With the research I had already done with spatialization techniques and music genres, I began to think of ways in which I could take this research to the virtual space, create new music material for virtual spaces, take existing material and transform it to fit within this virtual realm and so on. All in all, the questions I had been attempting to answer over the course of my research resurfaced and began to prod me to think again with virtual listening in mind.

Putting all of this together, in summary, this dissertation attempts to answer the questions, **"What psychoacoustic factors affect listening to music in non-virtual listening contexts and how could they transform when listening in a virtual listening context? How do these psychoacoustic factors work?"**

1.3 Breakdown of the research question

The research question mentioned in the previous section is extremely loaded with many moving parts. Several phrases used in the research question have fluid definitions. Therefore, it is important to describe it in more detail to bring greater clarity and to set the scope of this dissertation. This section attempts to do that.

1.3.1 Psychoacoustic factors

This is a very broad term, and it encompasses anything, and everything related to sound perception and psychological responses to sound. Within this dissertation, there are two user studies. One study focused on sound localization (Chapter 3) and the second study focused on rating various attributes based on musical stimuli (Chapter 4). A third quantitative study (Chapter 5) based on binaural recordings was carried out. While this was not a user study, many attributes that measured and analyzed in this study relate to perceptual parameters (such as brightness of sound) that could also be brought under the umbrella of psychoacoustic factors for this dissertation.

1.3.2 Listening contexts

A listening context can be defined as any and every setting in which we can listen to audio. By definition, this is extremely broad. We can further categorize this into real, non-virtual and virtual listening contexts.

Real listening Contexts

It is interesting to note that real and non-virtual (seemingly alluding to the same listening context/setting) are categorized separately. This is because, we want to separate the natural listening context (without sound enhancements) from the ones with sound enhancements. Therefore, a real listening context is defined as what we would experience at most times during a day. A context where we hear nature sounds, people talking to one another live, in-person, listening to acoustic music without the use of enhancements such as microphones, loudspeakers, or headphones. One of the goals of any music spatialization strategy can be to try and achieve this natural, real listening context. Since real listening contexts do not process or tamper with sound in any way, it is beyond the scope of this dissertation. However, given that every music spatialization strategy tries to achieve this ultimately, it was important to define it here.

Non-virtual listening contexts

Non-virtual listening contexts can be defined as those in which one is listening to the music and seeing the live visuals of the performer/musician, but the music is being heard via loudspeakers without the use of crosstalk cancellation.

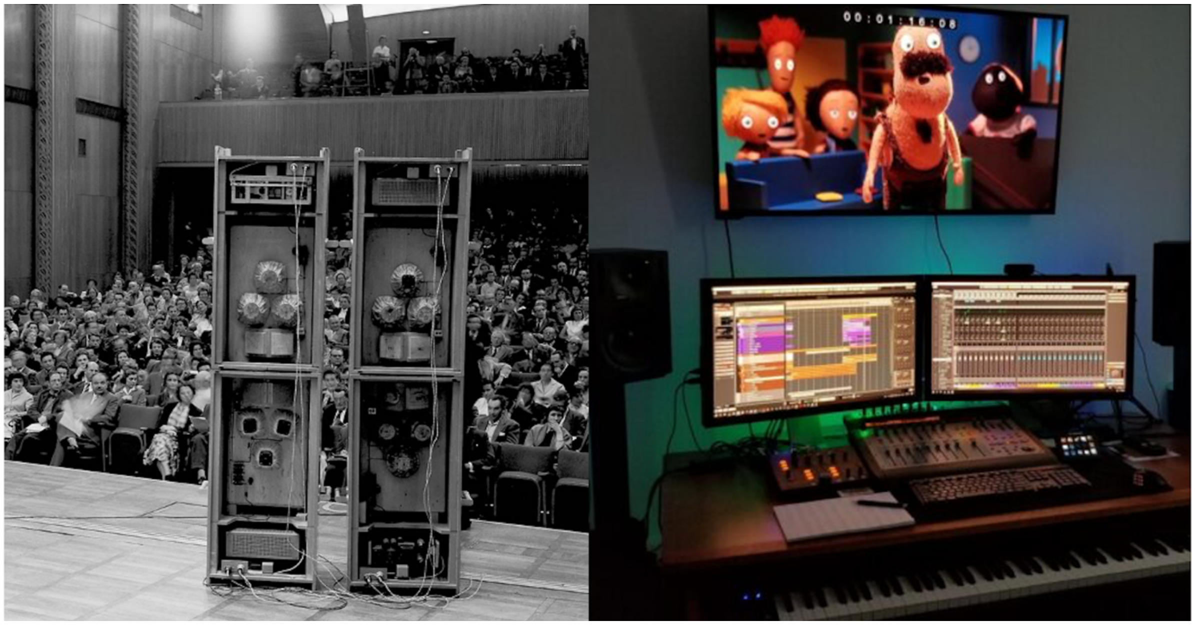

Figure 1: Examples of non-virtual listening contexts. On the left, a distributed audience and on the right a recording studio mixing audio for visuals.

Examples of these would include listening to a live performance in a concert hall (Figure 1), previously noted as a distributed audience environment or listening to the music while recording the musicians in a studio. Again, these definitions are fluid in nature and for the purposes of this dissertation, non-virtual listening contexts also include those listening to recorded music over loudspeakers as described in Chapters 3 and 4.

Virtual listening contexts

In contrast virtual listening contexts are those in which audio is primarily listened through headphones. It is possible for virtual listening contexts to include loudspeaker listening incase binaural audio is delivered through loudspeakers with crosstalk cancellation, but for this dissertation such a listening context is out of scope and not considered. Figure 2 shows examples of virtual listening contexts which could range from an audio-visual experience in virtual reality to listening and mixing audio over headphones. For the remainder of this dissertation, virtual listening contexts and binaural audio over headphones or simply binaural audio or headphone listening will be phrases that would be used interchangeably.

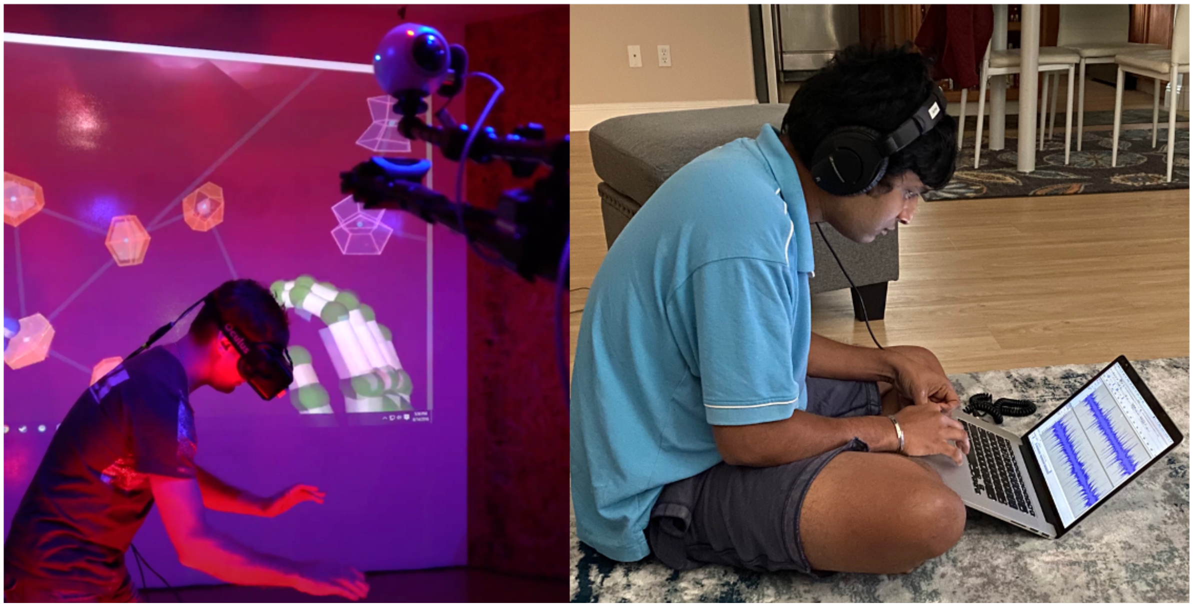

Figure 2: Examples of virtual listening contexts. On the left, an audio-visual VR experience and on the right, music mixing over headphones.

Dolby Atmos

While talking about listening contexts, it is important to look at a recent trend in the industry, Dolby Atmos. Dolby Atmos is a surround sound technology developed by Dolby Laboratories [3]. The reason Dolby Atmos is important to be discussed is because Dolby is selling its technology as a one size fits all. To quote a sentence from their website [3], “No matter what you’re listening to, Dolby Atmos adapts automatically to your audio device and system, so you’ll get the highest quality experience.”

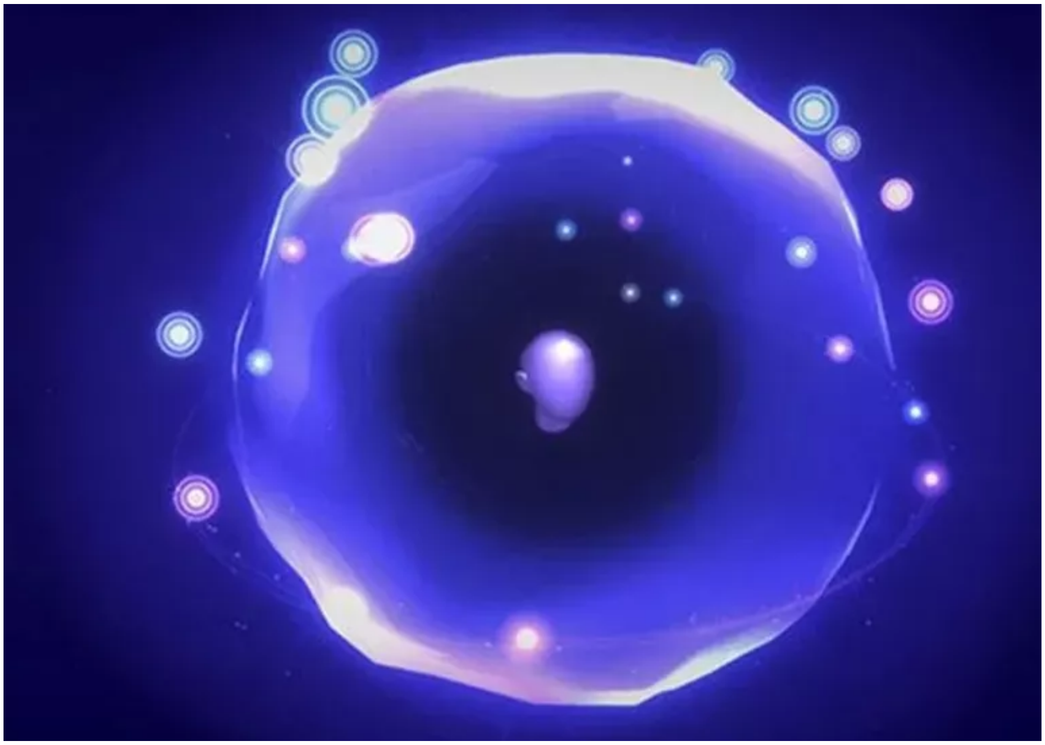

Figure 3: Pictorial representation of Dolby Atmos technology. Taken from [3].

This dissertation is shaped in a way that is in stark contrast to the above sentence wherein I make a claim that different listening contexts and settings should be treated independent of one another and a spatialization strategy that works in one context may not work in another context. I elaborate over the course of this dissertation how different listening contexts change perceptual music parameters and why we should not go for a one size fits all strategy, however, comparative studies with Dolby Atmos material have been left out in the discussion as that is beyond the scope of this dissertation. Over the course of the dissertation, I shall also highlight other material (interview with Prof. Paul Geluso) that highlight my school of thought, but at the end of the day, I believe it is up to the reader and the listener to take their own call on the matter. Even though comparative studies were beyond the scope of this dissertation, I felt it’s important to mention the current industry trend and comparative studies will be proposed in future work.

1.3.3 Research approach

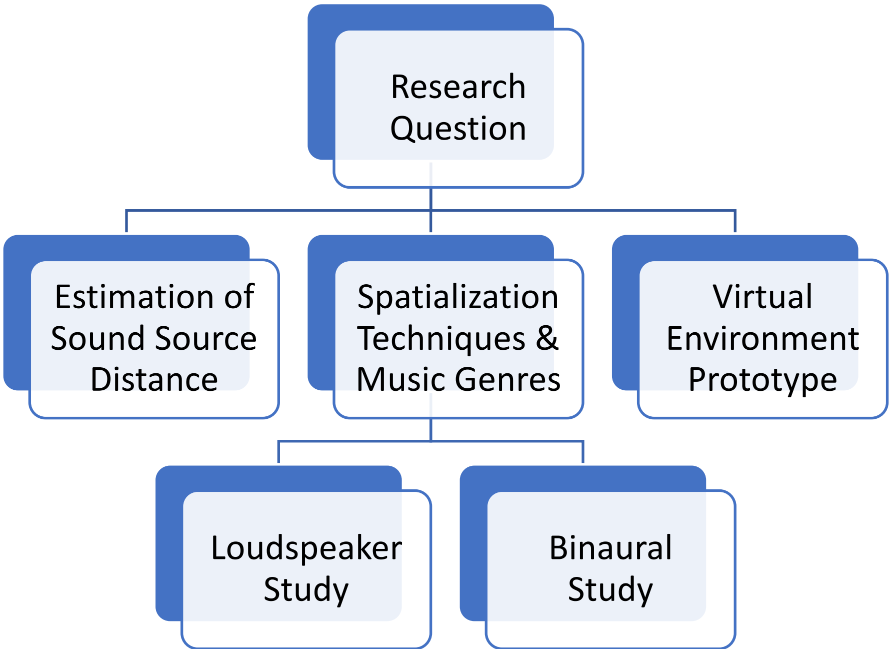
Research Question
Estimation of Sound Source Distance
Spatialization Techniques & Music Genres
Virtual Environment Prototype
Loudspeaker Study
Binaural Study

Figure 4: Hierarchal breakdown of the research question.

In my approach to answer the research questions posed in the previous section, I broke down the question into sub-questions/sub-topics that could be visualized in the figure shown above. The studies on sound source distance and the loudspeaker study of the spatialization techniques & music genres address the psychoacoustic factors of music spatialization strategies in non-virtual listening contexts part of the research question. With the Binaural study, we start moving toward virtual listening contexts. The binaural study as mentioned earlier is a quantitative study of several perceptual parameters of music and would also cover some aspects of the psychoacoustic factors part of the question. The virtual environment prototype as the name suggests looks only at the virtual listening context side of things.

1.4 Organization

This dissertation is organized in the following manner:

- Chapter 2, titled, “Development of spatial audio, virtual environments, and their intersection” gives an extensive background of the larger field of study. It covers the development of the field of spatial audio in the 20th century with separate sections discussing how the field has evolved in loudspeaker reproduction and headphone reproduction. The chapter also reviews virtual environments, what they are, how are they defined and evaluated, and the final section of the chapter reviews a few landmark case studies in the field. Through this chapter the reader will get a broad sense of the entire field while setting up for the specific sub-questions that would be addressed in the future chapters.

- Chapter 3 talks about the first sub-question, sound source distance. This chapter introduces what sound source distance estimation is all about, what studies have been done and then details the gap in the field this dissertation tried to bridge by designing, conducting, and analyzing a study on the estimation of sound source distance. Finally, the chapter concludes with a word on the significance of this research to the overall dissertation.

- Chapter 4 focuses on the loudspeaker study of the spatialization techniques and music genres sub-question. A user study in which different music genres are compared with different music spatialization algorithms is described. The chapter also includes an interview with audio engineer, Prof. Paul Geluso, who mixed and mastered all the sound stimuli used in the study.

- Chapter 5 continues with the spatialization techniques and music genres topic but shifts focus to the binaural aspect of the study. In this chapter, various melodic, rhythmic, and spectral features are extracted from binaural recordings and analyzed.

- Chapter 6 describes the creation of a virtual environment prototype. It describes the environment created, the features included and talks about how this contributes towards a unique tool that can potentially be used in future by music educators, electronic musicians, music composers and consumers.

- Chapter 7 summarizes the main concepts and contributions of the dissertation, the timeline of the dissertation as well as discusses the potential future work that would benefit the field, and which follows from the work presented.

2. Development of Spatial Audio, Virtual Environments and their intersection

Over the past few decades, audio research toward improved sound quality was a problem focused on the compression of the audio signal. With the improvements to hardware infrastructure, the consumer demand for better audio has shifted away from the issue of compression to more immersive environments with formats such as 5.1 surround sound coming widely into use [4]. The popularity of surround sound led to research and development of more immersive audio formats such as 7.1, 9.1. Nippon Hoso Krokai (NHK) has developed a 22.2 channel system while Audyssey has a 11.1 system. This unprecedented growth in increased number of channels providing immersive audio experiences, however, does come with a huge drawback. There is lack of wide consumer usage due to the high complexity of the system coupled with the high setup costs. Therein comes the idea of virtual reality and use of binaural technology to provide immersive listening experiences.

Now, what is virtual reality and what are immersive experiences? Researchers over the years have provided many definitions for these. The differences in the descriptions of these phrases gives us insight into the different mindsets of the researchers. The common theme across all definitions is that virtual reality and an immersive experience gives a sense of an alternate reality. A sense of telepresence. The histories of virtual reality systems draw heavily from foundational work in telepresence, robotics, cinema, and gaming [5]. Modern virtual experiences, however, go beyond casual entertainment, gaming and into music, design, communication, engineering, medicine and more [6].

Right from the early days of virtual reality (VR) systems, sound has played an integral role in the establishment of a convincing sense of immersion. In fact, the founding father of VR, Jaron Lanier thought of VR as a medium within which we would be able to collectively 'improvise reality' [7, 8]; a vision drawn from an explicitly collaborative and moreover musical origin. One of the earliest concept designs of a VR system, Morton Heilig's stereoscopic-television apparatus describes the use of earphones and binaural sound [9]. Sutherland, creator of the what is considered as the world first head-mounted display commented in 1965 [10] that computers did not have the capability of producing meaningful sounds that can be integrated into "the ultimate display", thus further highlighting the importance of sound in a VR system.

Despite the apparent importance of sound in VR systems, the primary focus of the development of VR systems has revolved around the visual component. Much of the existing discussion on VR systems focuses on technological advances ranging from higher pixel density for head-mounted displays (HMDs) to the simulation of real-world environment in real-time [5]. This is also apparent from the concept of the reality-virtuality (RV) continuum introduced by Milgram et al. [11] to describe real, virtual, and augmented environments. They said that rather than regarding augmented reality (AR) and VR as antitheses concepts, they can be illustrated as lying on opposite ends of the same RV continuum.

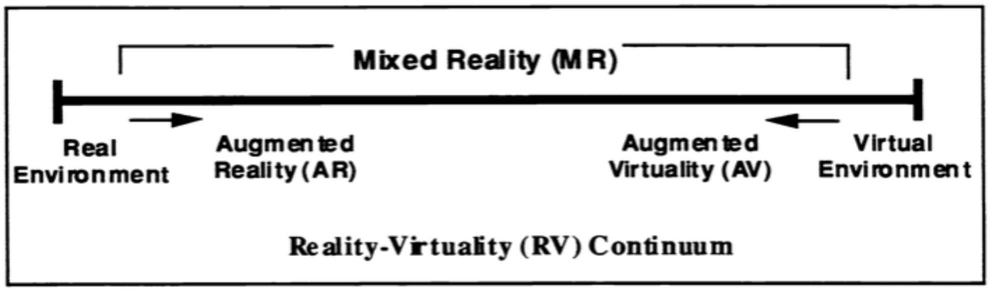

Figure 5: Simplified Representation of the RV Continuum based on [11]

From figure 1, Milgram et al. [11] described the case on the left as any environment consisting solely of real objects and includes anything observed directly in-person or viewing it through a viewing window as in a scene in a video display. The extreme right is an environment consisting solely of virtual objects, examples including computer graphic simulations either monitor based or immersive. In both cases, we see that the visual component is the driving force behind the framework and the auditory component is completely left out.

Graham and Bridges [12] observed this apparent lack of consideration toward the auditory side of things when they raised interesting questions based on the RV continuum. Regarding the integration of musical instrument-based performance VR systems, they say, “Where does this combination fall on the continuum and what are the implications for the composer/improviser? Are we dealing with the augmentation of the listener’s real environment through a virtual environment? The performer’s experience is potentially and solely based within a virtual environment.” Graham and Bridges [12] go on to consider Smalley’s ideas surrounding spectromorphology [13] and space forms [14] to help answer the questions they raised above and to guide them in the development of new mappings between real and virtual worlds. The research attempts such as [12] are an indication that the investigation of immersive sound in the context of VR systems is not completely absent or new, rather its simply an underused modality.

This underused modality possesses unique features. Unlike the visual counterpart, auditory perception is always on since we cannot close our ears. Thus, this channel always provides information about the surrounding environment, regardless of whether we pay attention to it or not [15].

Music and thereby audio also has another very unique connection with virtual reality and with performances within the confines of VR spaces. According to composer Henry Brant, space is “an essential aspect of musical composition.” [16]. Since 1950, spatialization has become an inevitable part of musical composition and a major part of academic research [17]. While Brant’s idea of space in music and the spatialization as spoken about in [17] do not directly refer to VR spaces, a performance within the context of a VR environment inherently has a new spatial element within it whether the composers/performers wish for it to be present or not. This space gives the composition a completely different musical aesthetic. Further, depending on how this aesthetic is incorporated in a performance could then impact the performer and the audience member in completely different ways – musician immersion and audience immersion.

Traditionally, musical performances have been constrained by physicality, i.e., being in the physical space. This gives the musicians an ability to attend to and communicate with one another as well as with the audience. As noted in [18], musician performance practices have evolved throughout history taking advantage of our natural understanding of the physical world, with the evolution of specific meta-languages of musical gesture and subtle communication as the end result. With VR being incorporated into the performance, this naturally goes out of the window and a whole new form of communication and art-form emerges. But this artform has a fullness to it. To build a VR system with musical performance requires not just careful attention to the sound, but also the visuals and interaction.

The rest of this chapter is organized in the following manner. Section 2.1 provides an extensive review of Spatial Audio for Immersive Environments. This section will elaborate on the different algorithms that can be used to spatialize sound, as well as delivery of the sound (loudspeakers and headphones). The focus will be more on the delivery through headphones as it is the medium of delivery for VR setups. Section 2.2 will give an outline on the creation of virtual environments, what they are, how are they defined and evaluated. Finally, section 2.3 will proceed with reviewing some of the landmark case studies that intersects spatial audio with virtual environments in ways that inspire and become the foundations over which this dissertation is built.

2.1 Spatial audio in immersive environments

A natural listening environment is inherently three dimensional [19] or as Blauert famously stated, “there is no non-spatial hearing” [20]. Therefore, the need for perceptually believable sound is fundamental to creating an immersive listening environment. These immersive environments can be obtained via both loudspeakers and headphones. Both come with their own sets of pros and cons. Although both mediums can be used to present the sound and historically delivery through loudspeakers has been the more researched topic, in the context of VR, with the growing market of individual consumption of content, headphone delivery is the faster-growing field, and it also presents greater advantages. Hence the article will review headphone-based reproduction in greater depth in section 2.1.2 as compared to loudspeaker-based reproduction which is reviewed in section 2.1.1.

2.1.1 Loudspeaker-based reproduction

Loudspeaker based spatialization of sound has existed since the first half of the twentieth century. The very first known attempt of spatial reproduction of sound through loudspeakers dates to 1911 when Edward Amet filed a patent on a device to pan a mono sound synchronized with a film projector around a series of loudspeakers such that the sound of an actor’s voice would follow his position on the screen [21]. It must be noted that, sound reproduction at this stage was still only mono in nature. In 1931, British Engineer Alan Blumlein filed a patent [22] that marked the birth of stereo sound. As we will see in section 2.2, Blumlein was not the first to record in two channels nor was he the first to broadcast audio in two channels, but he was the first to do both of these things together and present stereo to the world as we know it today. Blumlein's work was so ahead of its time, that even when he died in 1942, his work was largely unknown. It wasn't until about a decade later that stereo reproduction began to garner commercial viability.

While stereo reproduction of sound did not reach commercial viability until the 1950s, like Blumlein, there were many attempts made to push the boundaries of the average listening experience into more immersive ones. These mainly came from the cinema companies whose investment into expensive prototypes were with the idea to draw the audience and present them an experience that they couldn't afford to listen to at home. Notable among these are the Fox Movietone Follies in 1929, which used sound reproduction quite like the one described by Amet in 1911 and Stokowski's 3-channel reproduction of the Philadelphia Orchestra in Washington D.C in 1933 [23]. Stokowski's work also paved the way for the very first attempt at surround sound. This happened in the 1940 film, Fantasia. For Fantasia, a new audio system called Fantasound was designed to use three audio channels to transmit the sound similar to the 1933 attempt with the Philadelphia orchestra. Incidentally, Stokowski was also the conductor for this film. Apart from this, the system also used a separate control track that could pan the 3 audio tracks to a group of 10 loudspeakers, 9 of them placed around the audience horizontally (surround) and one on the ceiling (3D) [23]. As Malham and Anthony note [24], this was perhaps the most complete surround immersion yet achieved in a commercial setting.

The idea of using loudspeakers for immersion was making headway at the same time in the world of electronic music as well. Pierre Schaeffer used a four-loudspeaker playback system in his musique concrete. The earliest known performance using a four-loudspeaker setup was Schaeffer's Symphonie pour un homme seul in 1951. This was a five-track piece. One of the tracks was a tape machine controlled by the performer while the other 4 tracks fed to a single loudspeaker each. Thus, this performance had a mix of both live and preset sounds. The four loudspeakers were arranged with two loudspeakers in the front left and right, one at the rear in the centre and the fourth above the audience. This enabled the performer to place the sounds they controlled around the audience rather than just across the stage [25]. This effort of Schaeffer was followed by similar efforts in musique concrete by the likes of Pierre Henry, Jacques Poullin, and others. Another landmark performance was the five-loudspeaker piece, Gesang der Jünglinge by Karheinz Stockhausen in 1958. Stockhausen intended four loudspeakers to be positioned around the audience and one above them. However, at the world premiere in Cologne, this was not possible and the fifth ended up being positioned in the front centre. Stockhausen says of this piece, "..in this composition, for the first time, the direction and movement of the sounds in space is shaped by a musician, opening up a new dimension in musical experience." [26] Indeed, this did pave way for greater, immersive musical experiences to develop in the coming decades. Even as Schaffer, Stockhausen and others were experimenting with spatial sound in 4 and 5 loudspeaker setups and the commercially stereo was kicking off, Edgar Varèse pushed the boundaries of spatial sound music to heights that are perhaps extreme even to this day. In 1958, at the Phillips Pavilion in the World's Fair designed by Iannis Xenakis, Varèse presented his tape piece, Poeme Electronique over a whopping 425 loudspeakers! Varèse recorded this piece on four separate tape recorders which gradually desynchronized over time due to the different playing speeds. Varèse commented on this piece, "The loudspeakers were mounted in groups and in what is called "sound routes" (there were 9 predetermined routes) to achieve various effects such as that of the music running around the pavilion, was well as coming from different directions, reverberations, etc. For the first time, I heard my music literally projected into space." This path breaking installation is something that we can draw inspiration from to this very day.

As stereo reproduction of sound continued to shine through the 1950s, work progressed on pushing the boundaries of multichannel audio. The 1960s saw the birth of a 4-channel format known as quadrophonic sound. Initial attempts at producing quadrophonic sound involved extracting the out of phase signal components from the stereo sound and presenting them over a pair of rear loudspeakers. Pierre Scheiber developed a system in 1968 whereby he compressed 4 analog channels into just two for storage purposes and reconstructed them back to 4 channels during playback with certain constraints placed on channel separation and phase artifacts. There were other attempts of creating a variety of "quad" formats including trying 3 channels in the front and 1 surround channel in the rear similar to Stockhausen's speaker positioning during the Gesang der Jünglinge premiere in Cologne. In 1967, Pink Floyd performed the first ever surround sound rock concert at London's Queen Elizabeth Hall. The concert featured the band's own custom-made quadrophonic speaker system. During the concert, a sound mixing device called the "Azimuth Coordinator" was used to direct sounds to the multiple speakers and live music was supplemented with sounds from pre-recorded tapes [27]. Despite all of this, the quadrophonic sound format was a commercial failure. Dolby engineer Mark Davis [28] commenting on the failure notes that, the quad formats were possibly motivated by “commercial one-upmanship, [which] arguably result[ed] in products and systems being rushed into the marketplace prematurely.”

The commercial failure of quadrophonic sound did not deter research and attempts to further advance surround sound formats. Dolby laboratories made headway on this front in 1976 initially by introducing a 4-2-4 channel matrixing system that was used on the film A Star is Born. While the industry was still tentative about the prospects of surround sound, the big boost came from the sweeping success of the 1977 movie, Star Wars, where the swooping rear channel effects were a raging success. This paved the way for even further advancement on the surround sound front with the 1978 film, Superman being the first to feature famous 5.1 channel soundtrack. We repeatedly see the use of advance audio methods in cinema driving the development of consumer audio. Davis [28] rightly notes this, “The sustained establishment at last of surround sound in cinema would prove to be the gateway through which consumer audio evolved from stereo to surround.”

As stated earlier, 5.1 channel surround sound was first introduced in cinema in 1978 and this format attracted widespread acceptance. Within 13 years of its introduction, the format was standardized in June 1991 in Ottawa as ITU-R Rec. BS. 775-1 [26]. This was widely adopted for broadcasting, recording applications and the ‘home cinema’. The ITU standardization also helped avoid matrix encoding problems faced by the likes of quadrophonic sound. The ITU standard, however, did not define anything about how the sound signals should be represented or coded, it merely stated the ideal layout of the loudspeakers. The spatial encoding and sound field representation were left open in this format. It would not be too wrong to say that 5.1 surround sound is the most popular surround sound format to have come out.

Surround sound expansion continues to happen to this day and since 5.1, we have seen expansion with even larger number of channels being proposed, implemented, and seeing varying degrees of success. They include, 7.1, 10.2, and 22.2. None of these formats have gained as much popularity or acceptance as 5.1 and hence we shall limit our discussion here.

The section so far has concentrated on looking at loudspeaker-based reproduction of spatial audio more from the commercial side of things. While all this was happening, research on nuanced multichannel formats was happening slightly differently in academia. This included Ambisonics, Wave field Synthesis, Vector Base Amplitude Panning (VBAP) and Distance-Based Amplitude Panning (DBAP).

In 1973, during the time commercial engineers were attempting to encode quadrophonic sound, mathematician Michael Gerzon suggested an encoding scheme that came to be known as Ambisonics [29]. Ambisonics worked on the principle of encoding the sound signal in a sphere around the listener using spherical harmonic basis functions. The encoded signal does not contain the speaker signals, but rather a speaker independent sound representation known as B-format which is then decoded based on the listener's speaker setup. This offered considerably more flexibility for the sound engineer/composer to think in terms of sound source positions rather than loudspeaker positions. Mathematically, Gerzon generalized the system for an Ambisonic order "n", however, the sound field microphone he developed [30], was restricted to the first order. This heavily reduced the spatial precision that was possible during reproduction. First order Ambisonic signal meant a 4-channel recording (1 omnidirectional signal and 3 pairs of orthogonal signals in the X, Y and Z directions in a cartesian coordinate system). More about Ambisonics and mathematics behind it in Chapter 4, where Ambisonics was used as part of the experiment. Despite the flexibility and the solid mathematical foundation, Ambisonics did not garner much attention apart from academia and some recording enthusiasts. Since the turn of the 21st century, however, Ambisonics has seen a resurgence, especially with the development of higher order microphones (the 32-channel eigenmike by MH Acoustics or 64-channel microphone array described in [31]) as well as the boom in Virtual Reality (VR). It lends itself particularly well to VR applications as the B-format scene can be fairly easily rotated to match the rotations of the user's head orientation and then decoded using binaural stereo.

Another method to capture the entire sound field is called wave field synthesis. The foundational theory of wave field synthesis was given by William Snow [32] way back in 1955 as shown in the figure below. However, his theory couldn't be put to practice due to technical constraints of dealing with multichannel setups. In 1988, AJ Berkhout [33] laid the foundations for wavefield synthesis as we know today, but it wasn't until 1993 that Berkhout et. al. [34] coined the term wave field synthesis (WFS) for the theory proposed a few years prior. Much later, in 2008, Spors et. al. [35] presented a unified, generalized theory of WFS for two- and three-dimensional reproduction.

The basic theory of WFS revolves around three steps [33]:

i. Wave field acquisition by a microphone array

ii. Wave field extrapolation by a multichannel processor

iii. Wave field reconstruction by a loudspeaker array

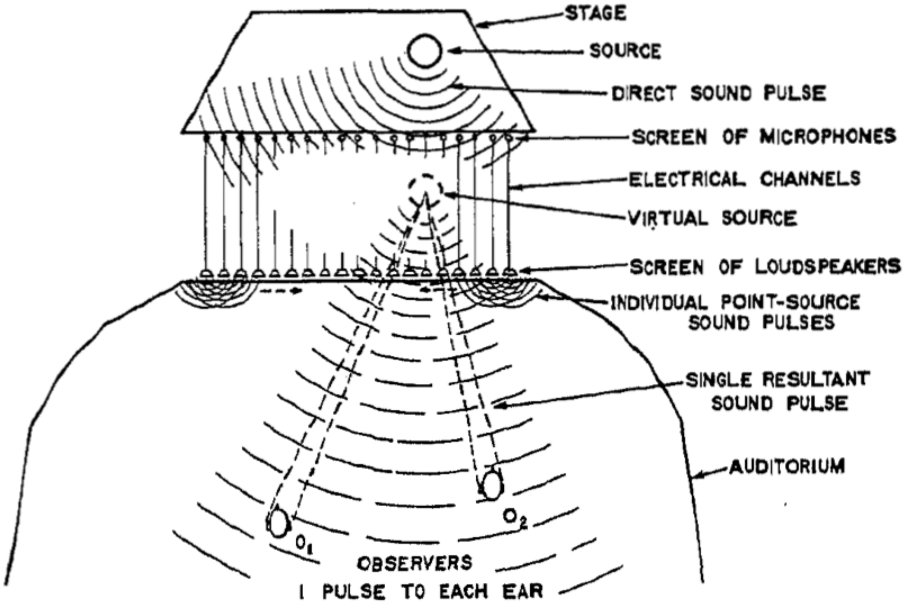

Figure 6: Foundational theory of wave field synthesis given by William Snow [32]

The microphone array is located near the sound source and each microphone captures the signal separately (no mixing). The process then applies spatial convolution. This is achieved through matrix multiplication. Finally, the loudspeaker array reemits the desired wavefronts into the hall towards the audience. The resulting sound field is ideally optimum for the entire enclosed space.

Ambisonics and WFS encoded the entire sound field within them. Their ideal reproduction was constricted to certain specific loudspeaker layouts such as orthogonal loudspeaker placements for Ambisonics. For arbitrary loudspeaker layouts, the multichannel formats Vector Base Amplitude Panning (VBAP) and Distance-Based Amplitude Panning (DBAP) were proposed in the 1990s and 2000s.

Introduced in 1996, VBAP was formulated as a scheme that would position virtual sound sources that would be independent of the loudspeaker arrangement thus being an improvement on algorithms like Ambisonics [36]. The basic theory of VBAP is based on Blumlein's amplitude panning but is reformulated with vectors and vector bases using a triad of loudspeakers closest to the virtual sound source position. This allows highly accurate spatial gestures and achieves convincing spatial effects producing sharp acoustic images in the process. The biggest limitation with VBAP being all loudspeakers are required to be nearly equidistant from the listener, thus creatin a narrow sweet spot of reproduction.

Distance-based amplitude panning or DBAP was introduced independently by Lossius et. al. [37] and Kostadinov et. al. [38] (where it was referred to as Vector Distance Panning) both in 2009. The primary goal of DBAP was to formulate a multichannel spatialization technique that is neither dependent on the speaker position nor the listener position thus making it very robust. This would make it useful in real world situations such as concerts, stage productions, installations, and museum sound design [37]. DBAP is an extension of the principle of equal intensity panning to a large array of loudspeakers. It assumes that all the loudspeakers in the array are active at all times and the gain applied to a particular loudspeaker at any instant of time is dependent on the distance of the loudspeaker from the virtual sound source. Given that DBAP works independent of loudspeaker position and listener position, it lends itself well to a large listening area or “sweet spot”.

VBAP and DBAP are explained in more detail in Chapter 4 where they were used as methods of spatialization in the experiment.

2.1.2 Headphone-based reproduction

Headphones can provide us with one of the most efficient ways of delivering audio to the consumer. It provides us complete control over what is being delivered in each ear. In audio terminology this means, the left channel to the left ear and the right channel to the right ear. This is called binaural audio. Therefore, one could argue that 3D audio technology begins and ends with binaural. However, before we begin a more elaborate discussion on the topic, it is first important how we define the term, binaural [39]. The earliest definition of "Binaural" was coined as early as 1861 by Alison to describe that two ears are involved in hearing [40]. "Binaural" literally means relating to or hearing with the two ears. However, in the field of 3D audio since about the 1970s or so, the term binaural is reserved for a two-channel sound that has been filtered or processed with all the spatial cues of time, intensity, and spectrum of the audio from the ears, head, and torso of the listener. This provides a complete acoustic image that resembles human localization as heard in a natural listening environment to the two ears. For the remainder of the article, the term binaural will refer to this latest definition of having a complete acoustic image.

As mentioned in the previous section, attempts at producing binaural audio occurred even before the realization of stereo sound by Alan Blumlein in 1931. Indeed, the very first experiments of two channel audio transmission and reception were done in 1880 by Alexander Graham Bell. Having invented the telephone 4 years prior, he did experiments with two telephone transmitters connected to a pair of receivers in what we could note as the first known binaural experiments. Only a year after these experiments, in 1881, engineer Clément Ader set up the first noted public demonstration of spatial audio with the transmission of the audio of the Paris Opera to headphones [23, 28]. Dubbed as the "Théâtrophone", this system of Ader's employed an array of 8 microphones across the stage. Each pair of microphones were connected to a pair of telephone receivers in four listening rooms. Listeners perceived a crude binaural rendering of the performance. The rendering also suffered from insufficient amplification and vibration damping. Yet, Ader's conception and execution were far ahead of its time, and it did garner traction and home subscriptions of the théâtrophone. These ran from the late 19th century well into the first part of the 20th century and it ceased to exist only around 1932 [41]. It might be interesting to note here that the nature of sound capture in Ader's théâtrophone is very similar of how wavefronts are captured in wave-field synthesis, so one could even say that the théâtrophone was also one of the first/primitive forms of wave field synthesis. This is, however, not done due to the nature of sound reception (interaural differences at the two ears of the listener) and hence we reserve the théâtrophone to describe an early example of binaural audio. It must be noted here that beyond interaural differences, the théâtrophone did not do any processing of the audio described above to fit in the definition of binaural sound. Indeed, Ader's demonstration fit in with the earlier definition of binaural which was used for techniques that recorded and reproduced signals for the two ears and not necessarily to describe signals that have been modified by the human body.

Binaural sounds with all the localization cues present can be obtained in two ways, either naturally, i.e., binaural recordings or through signal processing (binaural synthesis). Let us look at them one by one.

Binaural Recording

One of the fastest and most convincing ways to deliver 3D sound to listeners is through binaural recordings. These are usually done via dummy heads. Dummy heads are a physical representation of the human head, with anatomical features like that of an adult head. This includes the size, shape, and location of the ears. Some dummy heads even have shoulders and torso which help in accurate capture of body reflections from the upper body which aid us while listening to high frequency content. In these dummy heads, tiny microphones are placed at the location of the ears and thus they can capture the sound as it would appear at the ears of the listener.

The earliest dummy head recordings to be definitively noted are from 1927 when both Harvey Fletcher & Leon Sivian and Batlett Jones filed independent patents [42]. Both were very primitive dummy heads. They used spheroid objects as the “dummy head”. Fletcher’s efforts in creating more sophisticated dummy head recordings resulted in the very first human manikin, Oscar which came into existence in 1931 and was debuted at the Chicago World Fair in 1933. At the fair, Oscar was placed on a stage and a pair of headphones were connected to the manikin. A speaking person then walked around “Oscar” and the person waring the headsets could perceive that the speaking person was moving, a startling effect at the time [42]. Since, this was done in real time and the listener were also able to visually see the moving person added to the amazement of the audience. [43], however, reported that while Oscar gave very effective demonstrations of binaural recording, there were many localization errors, mainly front and back confusions, and distance errors. Though these issues had been identified as far back as 1933, we continue to deal with them to this day. As seen from the figure, Oscar did have the shape and looks of a human but was still a very primitive device. The reason were the large microphones. These was too large to fit as ears. Instead, they were placed as the cheekbones of the manikin.

It took only about a decade for Bell labs to improve upon Oscar as they came out with Oscar II in the 1940s. As seen from the figure, Oscar II has microphones fitted at the location of the ears instead of the cheekbones, but the microphones were fitted such that the membranes projected out of the ears and were at a distance of about 5mm from the concha. This meant that the recordings from Oscar II, while an improvement on Oscar, lacked in the effective capture of the coloration of the audio signals that happens due to the pinnae.

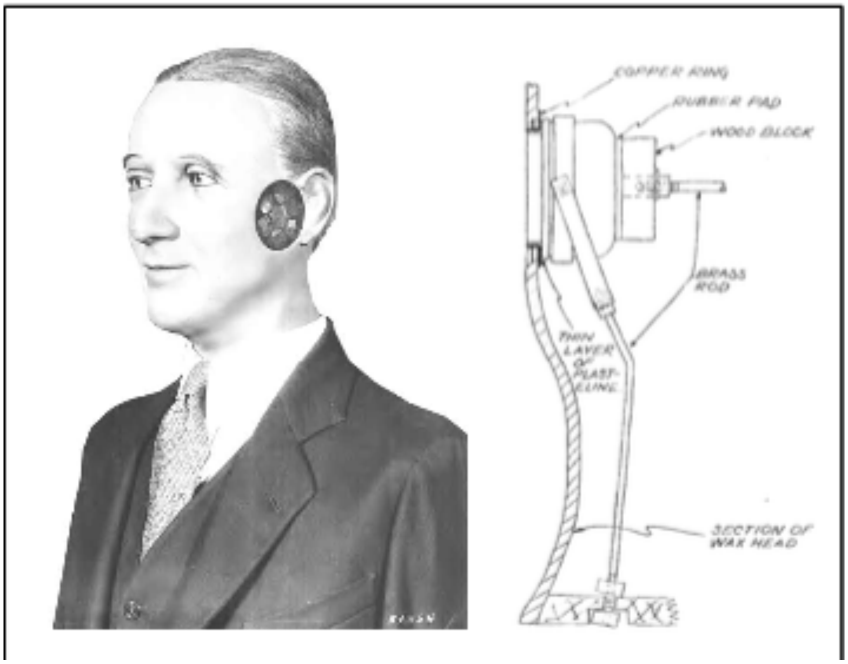

Figure 7: Oscar, the manikin made by Harvey Fletcher and his team. On the right, the lateral view of the microphone fitting [43].

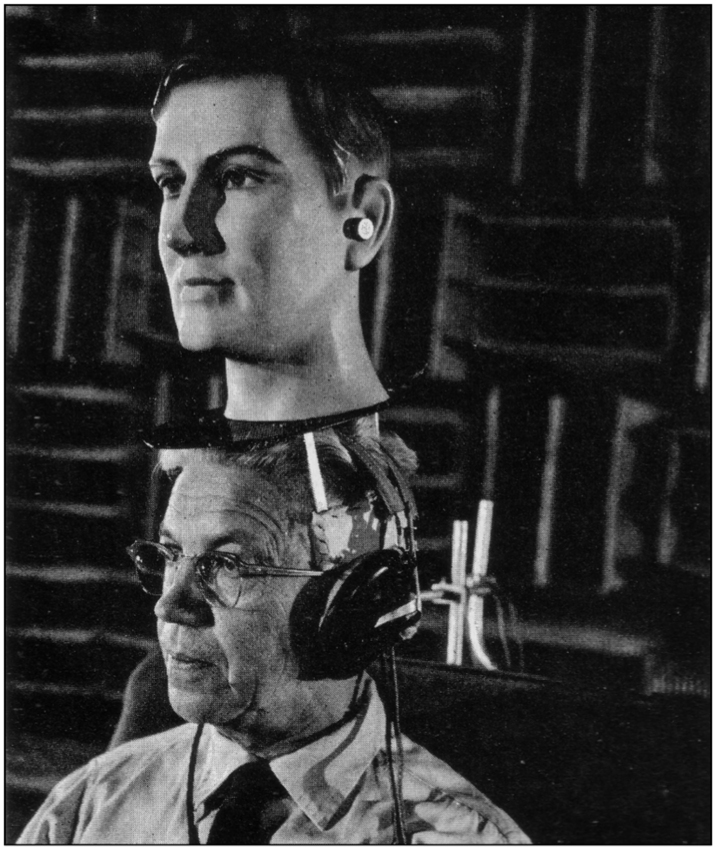

Figure 8: Oscar II being used by H. Fletcher in localization experiments [42]

In terms of design, present day dummy heads are not very different from these earlier models. The dummy heads today have smaller microphones that can be fitted inside the ear of the dummy head and therefore, they can be placed either at the entrance of the ear or anywhere along the ear canal right up to capturing sound at the location of the ear drum. This has allowed the flexibility in some dummy heads to have multiple pinnae designs that could be interchanged to capture binaural recordings with different pinna colorations in order to try and get closer to providing more effective listening experiences as each person's pinna is different.

Present day binaural recordings or binaural sound capture can be broadly classified into three kinds: (i) dummy head, (ii) dummy head with shoulder and torso and (iii) binaural microphones. These are depicted in the figure below:

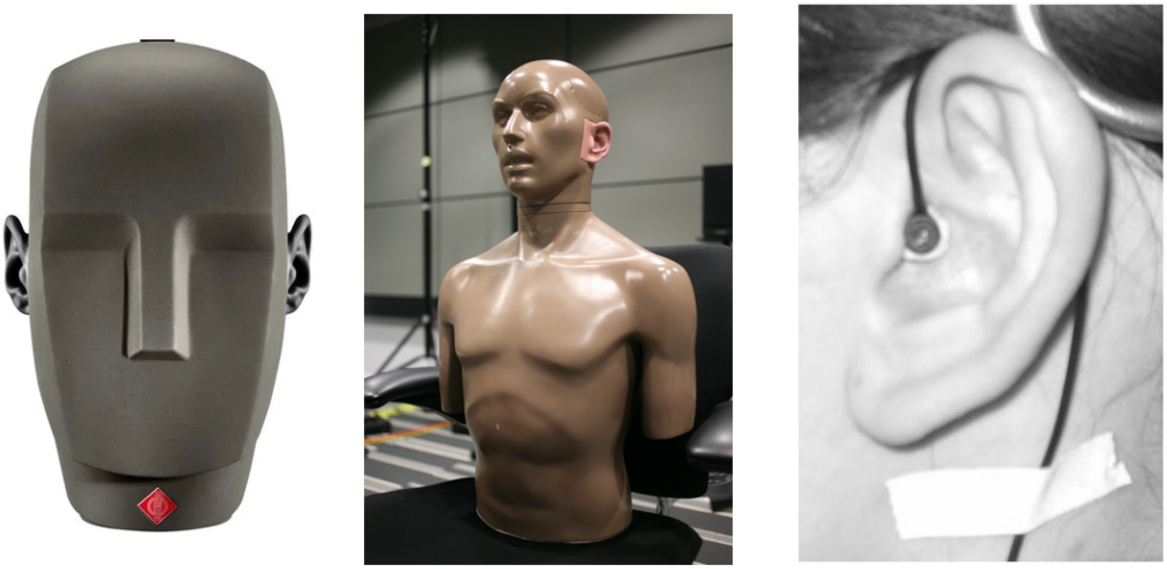

Figure 9: Left to right: Neumann KU-100, KEMAR, and in-ear binaural microphone

Dummy heads such as the Neumann KU-100 dummy head show above have all the anatomical features of a human head, but do not come with shoulder and torso. While we lose out some of the high frequency reflections that aid in elevation localization, these kinds of dummy heads are extremely portable devices, and it makes it easy create quality binaural content in multiple locations such as musical events. The KEMAR (Knowles Electronic Manikin for Acoustic Research) dummy head, shown in the center, does come with shoulders and torso and helps us capture complete binaural cues while being less portable. These are primarily used in acoustic measurement tasks in audio laboratories or to take binaural room impulse response measurements for acoustic characterization of rooms. Lastly, in-ear binaural microphones such as the one shown in the right side of the figure are excellent devices for recording personalized content. The recordings made with dummy heads are merely representations of average human heads, while inserting binaural microphone probes into your own ears and capturing the audio will give you the highest quality personalized content as the recording would capture all the reflections and coloring your body does at all times. Others listening to this content, however, would get an experience akin to listening to recordings made with any regular dummy head.

Binaural Synthesis

The signal processing way to obtain a binaural signal is called binaural synthesis. It is a process through which one can simulate binaural signals. In order to do this, we need what are known as HRTFs (Head Related Transfer Functions). By superimposing a sound signal with an HRTF, we are able to create a representation of the signal appearing at the two ears of the listener, thus synthesizing a "binaural signal", which would appear to originate at the location defined by the HRTF. The mathematical process of doing this involves convolving the audio signal with the HRIR (Head Related Impulse Response – time domain equivalent of the HRTF). This can be visualized in the figure below. All this clearly indicates that the most important element of binaural synthesis is the HRTF, which is described in the following section.

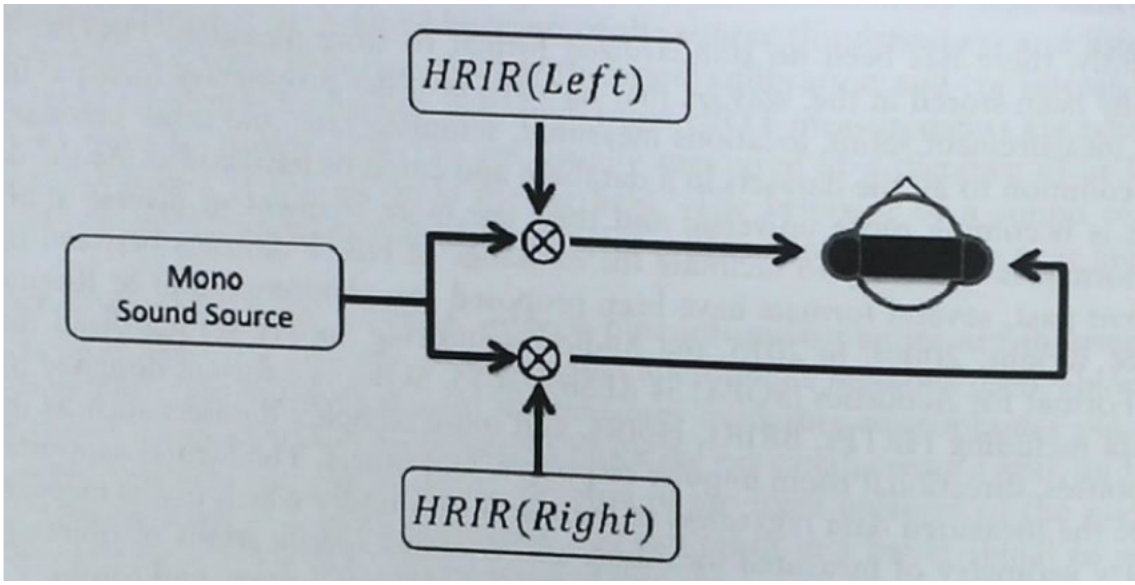

Figure 10: Binaural Synthesis - convolution of a monophonic sound source with the left and right HRIRs (taken from [39])

Head Related Transfer Function (HRTF)

The sound signal reaching the two ears of a listener from a sound source is different at each ear as well as different from the original sound signal. The transfer function from the sound source to the to the ear canals considering all the temporal and spectral changes that the signal undergoes in the process of reaching the two ear canals is known as the head-related transfer function (HRTF). Temporal changes are present due to the fact that the two ears are located on different sides of the skull and the sound signal has to traverse different distances to reach the two ears thus resulting in slightly different arrival times. Spectral differences arise due to the shape of the pinnae, the shadow effect caused by the skull on the contralateral (farther away from the sound source) ear and the reflections from the shoulders and torso.

The most effective way to capture an HRTF is through acoustic measurements. These can be done by inserting binaural microphones into a person’s ear and playing a test signal from a loudspeaker positioned at a certain location (azimuth, elevation, distance) relative to the subject. This test signal is recorded at both ears and the HRTF or its time domain equivalent, HRIR (Head-related impulse response) is extracted. Mathematically, this process can be given as:

$Y_L(\theta, \varphi, d) = H_L(\theta, \varphi, d)X$ and $Y_R(\theta, \varphi, d) = H_R(\theta, \varphi, d)X$

where,

θ = Azimuth

φ = Elevation

d = Distance

YL and YR = Spectra of the signals at the left and right ears respectively

HL and HR = HRTFs of the left and right ears respectively

X = Spectrum of the sound source

From this, the HRTF can be extracted as,

$$H_L(\theta, \varphi, d) = \frac{Y_L(\theta, \varphi, d)}{X} \; and \; H_R(\theta, \varphi, d) = \frac{Y_R(\theta, \varphi, d)}{X}$$

The extraction of θ φ and d would yield the localization of the sound source.

One important aspect of the capture of HRTF that is not mentioned above is the elimination of the unwanted signal characteristics that get superimposed on the recording, thereby negatively impacting the HRTF measurement. These are the spectral characteristics of all the equipment used in the recordings, i.e., microphones, loudspeakers, pre-amplifiers. Mathematically, this can be represented as,

Y(n) = X(n)S(n)M(n)H(n)

where,

Y(n) = Spectrum of the recorded signal

X(n) = Spectrum of the input signal

S(n) = Loudspeaker transfer function

M(n) = Transfer function of the microphone and pre-amplifier

H(n) = HRTF

Thus, HRTF after elimination of all these unwanted characteristics can be given as,

$$H(\mathrm{n}) = \frac{Y(\mathrm{n})}{X(\mathrm{n})S(\mathrm{n})M(\mathrm{n})}$$

HRTF measurements can be done in two ways. One way is to have a fixed subject, rotating speaker setup, and the second way is to have a fixed speaker, rotating subject setup. Since the measurement is directionally dependent and it's the relative location of the sound source (loudspeaker) with the subject, both methods would yield identical results with correct calibration. Typical HRTF data would have a density of 5°-15° in azimuth and elevation.

Until recently, there hadn't been any standard format for storage of HRTF data. In 2015, the Audio Engineering Society standardized the Spatially Oriented Format for Acoustics (SOFA) as AES69-2015. The standardization of HRTF storage eased the sharing of datasets, something very important in VR applications. As we saw above, the measurement of HRTF is a cumbersome, time-consuming process. Given that all of us have different body sizes and ear shapes, each person's HRTF will be very unique. That means, the best Binaural audio and the best immersive experiences in VR application with headphones would come from using individualized HRTF data. When that is not available, the best approach would be to find the "best fit". This was proposed in [44] and [45]. The graph below nicely categorizes the accuracy of the spatial audio image against the complexity of obtaining the HRTF. To be able to find the best fit, we would need to have access to a large dataset and therein the standardization of storage would be extremely essential.

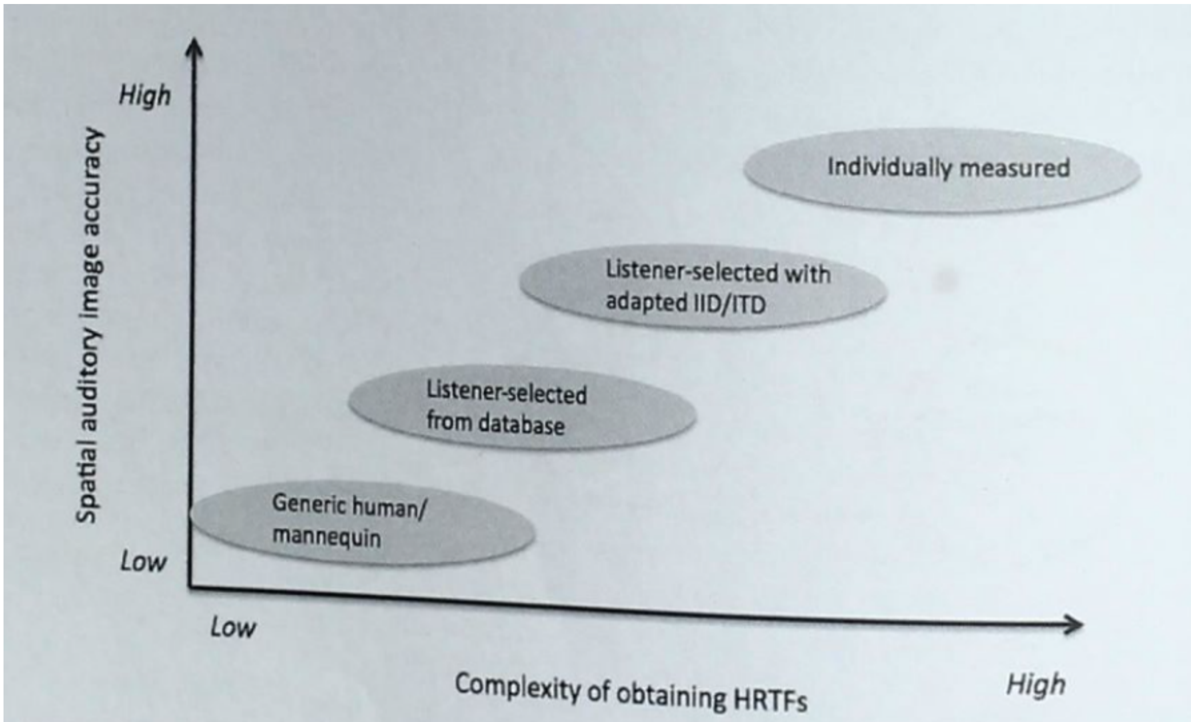

Figure 11: Spatial auditory image quality against the complexity of HRTF measurement (taken from [39])

2.2 Virtual environments

Having looked at the development of audio, it is now time to look at Virtual Environments to see how this part of the field developed before seeing the application of the two in tandem. But, before that, it is important to first define what exactly is a virtual environment. Different researchers have defined this term slightly differently according to the application/tool or use case they were restricting their research too. One such definition given by Carr and England [46] wherein they said, “A virtual environment is defined as a multisensory experience of location or set of locations through artificial electronic means.” Barfield and Furness [47] described a virtual environment as, “the observer can navigate in this environment and interactively manipulate it.” While Wickens and Baker, [48] defined it as, “A virtual environment is a combination of multiple features.” These features can include but is not restricted to a three-dimensional viewing, a dynamic interface, a closed-loop interaction, an ego-centered frame of reference, a multimodal interaction and a head-mounted display and tracking [49]. All these features increase a sense of immersion for the consumer. To that end, we can say that virtual environments are generally displays that perceptually surround the observer to create an immersive environment.

Once again, we come up with the term immersion. In the context of virtual environments, Bargar et. al [50] categorized immersion into two classes of experience – fictional constructs and feedback constructs. Fictional constructs they described as those that involve the observer's "willing suspension of disbelief." When this happens, it supports the narrative world created in the virtual environment much the same way it happens in literature, theater, and cinema. Feedback constructs on the other hand they described as "everyday experiences" that an observer constructs by taking actions and observing their consequences through multiple sensory modalities. Applying these to our topic of music in virtual environments, we could say that the task of music listening can be associated with the fictional construct whereas music performance would primarily be identified with the feedback construct. This is because, while the same piece of music might be playing in both scenarios, with music listening, the observer is free to use their imagination to create their own visual representation of the music while in the case of music performance, there is an element of restriction that is subconsciously imposed upon the observer as he/she takes into account the interaction between the performers as well as the interaction of the performers with the audience to create their the imagery. Thus, building off feedback of sorts.

Another important concept we come across when talking about virtual environments is the concept of presence. Researchers believe that presence is a measure of the utility of a virtual environment application. So, what exactly is presence? This term consistent with other concepts we've dealt with so far has different definitions and interpretations [51] depending upon the scope and functionalities of the virtual environment. The common view is that presence is the sense of being in a virtual environment rather than the place in which the participant's body is actually located [52]. Now, thinking about this view auditorily, it would seem to us that presence cannot be easily obtained without visual cues and conversely, just with convincing visuals, it would be difficult to obtain good presence without a convincing spatial image of the audio. Therefore, I would argue that for a successful virtual environment, multiple layers need to work in tandem.

Now, how do we evaluate whether the virtual environment created is a successful one or not? Sanchez et. al [52] say, "by a successful virtual environment is we mean that the person responds to the virtual stimuli as if they were real. Responses should be considered at every level, from unconscious physiological behaviours through automatic reactions, and from conscious volitional behaviours through to cognitive processing – including the reporting of the sense of 'being there'. Interestingly, this can take place despite every participant's absolute knowledge that the virtual environment is fake." Taking this definition, we could come up with two methods to evaluate the success – a questionnaire and a behavioural method. A questionnaire would simply ask the participant to rate on a scale, how much sense of "being there" did they feel while doing the task. The problem with this method is bias. If the person is well-versed in the field and the task at hand, there might be an internal bias created. The behavioural method would come in handy here. This would measure presence by testing the participants' response to a particular stimulus. If the participants behave in the virtual environment in a manner that is similar to how they would behave in an equivalent real environment, this would be a sign of presence. The limitation with this method, however, is that stimuli that would elicit such responses have to be created/presented in order to evaluate the environment. Something that may not always be possible or not particularly helpful in the goals of the virtual environment and huge technical overheads.

2.3 Landmark case studies

In this section we shall look at some landmark case studies that have shaped the virtual environment landscape and are continuing to inspire and progress this field of research. A lot of projects in this space are heavily based in the visual domain and the audio side of things is not at the forefront of the research. However, here, barring the CAVE (CAVE Automatic Virtual Environment) we shall focus on those case studies where audio and music have assumed as much importance if not more than the visual aspect.

Immersive virtual environments can be broadly classified into two types of systems: one is screen based virtual and augmented reality setups and two headset based virtual environment setups. Examples of screen-based systems include the CAVE, the Allosphere and the likes while examples of headset-based environments are DIVA Virtual Audio Reality System, Resillience, Immersive Orchestras and Calibrated Model of Notre-Dame Cathedral.

The CAVE

The CAVE virtual environment described by its creators as a virtual reality theater is perhaps one of the first screen-based virtual environments. It was built as a tool for scientific visualization. Designed in early 1991, it was implemented and first demonstrated to people in late 1991 and presented for the first time at the SIGGRAPH conference in 1992. The primary goals of the CAVE were to achieve good surround vision, be less sensitive to head rotation errors, mix VR imagery with real devices, guide and teach others in artificial worlds [53].

The CAVE was not without shortcomings though. Amongst them were that CAVE did not project to all 6 sides, which limited the immersiveness of the environment. The other major issue was dealing with the audio. The authors noted that with speakers surround the listening, it should have in theory been possible to achieve good directional sound, but sound localization was affected by the reflections off the screens. An issue that could not be resolved.

Despite these shortcomings, the CAVE did prove to be an effective and convincing virtual reality display and provided a great virtual experience to the visitors. In addition, the biggest contribution was that it spurred other researchers to think creatively and solve the shortcoming of the CAVE and push the boundaries of virtual experiences as we will see with the Allosphere.

The Allosphere

While the CAVE environment got the ball rolling, the Allosphere really set the standards for screen-based immersive environments. The Allosphere was conceptualized as an immersive environment that could be used as a tool for data discovery and data exploration. Unlike the CAVE counterpart, Allosphere ensured its focus was not just on data visualization and scientific visualization, but also on sonification and spatial audio. Early implementation and presentation of the Allosphere had 4 stereo screen projectors and 16.1 loudspeakers [54, 55]. This was increased over the years to its current setup of 26 stereo projectors and a 54.1 audio system [56].

With its novel spherical shape, its immersive multi modal capabilities, the Allosphere brings a unique immersive experience to the visitors. The central bridge allows for a large number of viewers (up to 30) to enjoy the experience at once. This is much larger than what CAVE like environments can offer.

The audio system of the Allosphere was carefully designed, with 3 rings of loudspeakers at different elevations. This allows for great 3D-audio imagery. The Allosphere audio framework supports for spatial sound reproduction via VBAP, Ambisonics and wave-field synthesis making it one of the most robust multimedia environments present.

VESPERS System

Built in 2016, VESPERS is a unique tool that can be thought of as a bridge from the screen-based virtual experiences such as the CAVE and the Allosphere to headset-based (and headphone) based setups in immersive orchestras later. This is because, the visual component of VESPERS is accessed via a headset and HTC Vive controllers for interaction in the virtual environment. The audio, however, is delivered via a 24.2 loudspeaker array. The aim of this project was to develop a hybrid system for the development and presentation of immersive virtual spaces [57].

VESPERS is a key project to discuss here, as audio and music composition from the core of this project. The goal of VESPERS was to visualize audio data as an object that could be moved around within the virtual environment using the vive controller and thereby create a sphere of musical information that when manipulated within the virtual space sculpts the spectra of each sound object within that virtual space. All this culminated in the creation of the VR spatial audio installation, “To Notice and Remember” using VESPERS.

Resilience

Resilience is the name of a collaborative performance between a laptop orchestra and a VR performer. As described in [58], “Resilience is a work created for and performed by a laptop orchestra, led by a VR performer." This piece also falls in the realm of the intersection between screen-based VR and headset-based VR.

This piece was very unique. There was one performer conductor in virtual reality. This person was wearing a headset and sat in the center of the stage facing away from the audience so that the left to right motions of the performer coincided with that of the audience. Since the conductor was wearing a headset, they conducted the laptop orchestra without actually seeing them. A projection screen was hung behind the ensemble so that the audience got a two-dimensional view of what the VR performer was looking at.

The design choices are extensively detailed in [58] and it shows how the connection between the real world and the virtual world can be formed in a VR performance such as resilience. The authors also describe and detail principles for musician/performer's immersion vs audience immersion. This is a key concept to keep in mind when virtual environment performance systems are being developed.

Immersive Orchestras

Immersive Orchestras is a novel immersive experience for orchestral music. In the words of the authors, "Our goal is to blend in the presentation art (music), technology (VR), past (classical repertoire) and future (new user applications). We believe music learning both for education or personal enrichment can benefit from advances in the area of Virtual and Serious Games." [59].

In this project, the authors create VR content of orchestral music that goes beyond just 360 visualization of the orchestra. They use audio separation to separate multitrack orchestral recording to obtain separate instrument tracks. Once this is done, any particular instrument in the orchestra can be emphasized over the full orchestra to get an acoustic zoom effect. They call this instrument emphasis and this entire process creates a unique immersive multimodal experience for the listener.

This tool was motivated by music education and new ways to experience music. By using the acoustic zoom, students could listen to specific instruments in greater detail to understand the nuances of the music from that instrument's perspective and how it sits in the whole mix. This is not possible in a regular mix as the rest of the orchestra would mask the instruments finer performance nuances.

Calibrated Model of Notre-Dame Cathedral

The last case study to be looked at is the virtual reality performance auralization in a calibrated model of the Notre-Dame Cathedral. In this project, a visual model of the Notre-Dame Cathedral was built using 3ds Max software based on architectural drawings, photos, videos [60]. An audio recording of the "La Vierge" concert was done using 44 microphones. This concert had a symphonic orchestra, 2 choirs and 7 soloists. The recordings were convolved with 3rd order Ambisonic room impulse responses (obtained separately) for auralization. The auralization and the visual model were integrated using BlenderVR to create a unique VR experience that allowed visitors to experience a unique immersive, interactive audio-visual VR application and enjoy the "La Vierge" concert.

This VR application was a step towards new experiences of live events. Acoustic simulations of complex places like the Notre-Dame Cathedral are a challenging task. Accomplishing this, only opens doors for other such simulations, which could lead to new ways of concert consumption. The barrier of physical presence of grand celebrations such as the 850th anniversary of the Cathedral can then be experienced long after the event takes place and in a multitude of locations while still getting the feeling of being "in the best seat in the house."

Concluding Remarks

As mentioned in the opening lines of this section, these case studies continue to inspire and progress the field to this day. Since the onset of the global pandemic in 2020, researchers and artists were forced to stay at home and think in new ways. This led to many innovative virtual collaborations. It would not be a far stretch to say that at least a handful of them would have been influenced by the case studies looked at above. As the world moves into this new normal, with the growing number of virtual reality applications, the impact of these projects is far beyond the technology and artistry it portrays. It has cultural implications too that are potentially significant for the way multimedia is consumed and looked at.

In summary, this chapter has provided us an extensive understanding of the field and laid the foundation for the upcoming chapters that deal with specific sub-topics/questions that are within the areas looked at within this chapter, yet were unanswered, thus shaping this dissertation.

3. Estimation of Sound Source Distance

The estimation of sound source distance has been a topic of research interest for a number of decades now. Humans are known to be good at localizing sound in the azimuth and elevation but are poor at estimating the sound source distance. This project looks at examining the effect of a known environment on the estimation of sound source distance. The project aims at initially testing the subject's perception of sound source in an unknown environment and then examining the effect of training the subject to the environment to see if training/learning the acoustics of the environment improves the estimation of the source distance.

3.1 Introduction

The human auditory system can provide listeners with critical information about the spatial layout of their environment [61]. This information enables humans to turn and face the sound source, use the audio cues when vision is ineffective, such as the dark and so on. For many decades, scientists have tried to understand how the auditory system processes the spatial information and how well we are able to localize sound. While sound localization studies have been a subject of research for many decades, the sound distance perception has received comparatively less attention of researchers [62]. Here we try and understand how good we are at estimating the sound source distance and does this perception change as we learn the acoustics of the environment.

The rest of the chapter is organized in the following way. Section 3.2 presents a brief overview of the prior work done in the field of auditory distance perception, and what factors change the distance perception in what way. In section 3.3, the methodology of the current research is described including the pilot study, the design of the experiment, and the participants. Results of the experiment and a brief discussion are covered in section 3.4 and section 3.5 has a summer of the chapter.

3.2 Background

The estimation of sound source distance has been a popular research topic for a number of years. Studies in the past have shown that we are good at estimating the relative distance of sources [63, 64]. However, estimating the absolute source distance has always been poor. Zahorik [61, 62, 63] showed that humans tend to overestimate source distance that is less than 1m away and underestimate source distance for sources greater than 1m. This behavior has been well captured and explained in the figure below.

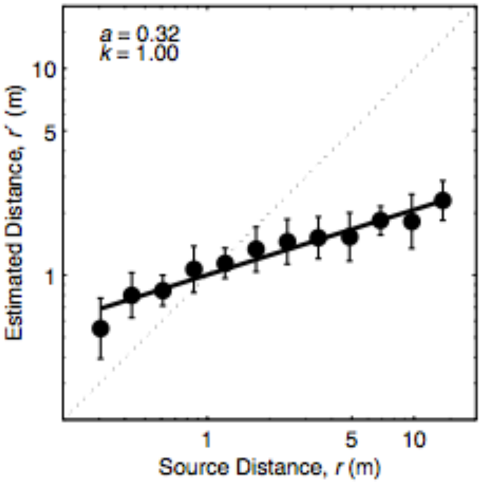

Figure 12: Plot showing source distance and estimated source distance. Figure from [61]

As seen from the figure, the underestimation of sound source distance gets worse with increase in distance. [61, 63] showed the effects of vision and loudness on the source distance estimation. Vision as with other spatial cues, has a positive effect on the source distance estimation. Perceived loudness, however, does not change with change in source distance. This is thought off as the primary reason for poor estimation of source distance.

However, it has also been shown that the accuracy of distance estimation improves when the experimentee is presented with known stimuli [62, 64], e.g., a friend's voice or a sound heard in daily life as opposed to stimuli usually used in localization experiments, e.g., sine sweep or pink noise bursts.

Given the above observations, the author hypothesizes that, "If there is a 'known' factor associated with the accuracy of estimation of sound source distance, then a 'known' acoustic environment should also improve our sound source distance estimates."

The goals of this project are twofold: a) To design an experiment to examine the effect of environment on sound source distance and b) To conduct a subjective analysis and assess the effects in an empirical manner.

3.3 Methodology

3.3.1 Experiment design

To test the above hypothesis, a subjective experiment was designed. A large room was selected, and loudspeakers were lined up one behind the other to play the stimulus. The experiment setup is explained in detail in section 3.2.

As the stimulus, a sine sweep from 100Hz to 5000Hz was created using MATLAB. The design choice was between a sine sweep and pink noise bursts. Eventually, the authors decided to use sine sweeps as it was felt they covered a large frequency range and stayed in the environment for a little longer than noise bursts, thereby giving the subject a better opportunity to estimate the distance from which it was played from.

There were four testing phases in the experiment. All testing phases, hereon referred to as distance localization test, involved playing the sine sweep randomly from one of the loudspeakers in the configuration and asking the subject to name the loudspeaker number from which the sound was being played. The four testing phases were:

- Pre-training phase: This was the first phase of the experiment. The participant took a round of distance localization test soon after entering the room and completing the consent form and the background questionnaire.
- Self-training phase: After the pre-training test, the subject was asked to explore the environment and figure out the acoustics of the room in whatever manner he/she deemed appropriate. For subjects new to sound localization and surround sound concepts, the concepts of impulse response, reverberation, early and late reflections, Interaural time difference (ITD), Interaural intensity difference (IID) and methods capturing impulse responses were briefly explained. The experimenter assisted the subject as requested (clapping from various parts of the room, etc.). Once the subject felt he/she had trained himself/herself sufficiently, a round of distance localization test was performed.
- Guided training phase 1: In this phase, training took place in the form of sine sweeps being played from the loudspeakers, much like the testing phase. Five rounds of training were conducted with sounds being played from the loudspeakers in a random order. At the end of each sound, the subject was informed of the loudspeaker it was played from. Thus, with this, the subject got an idea of how the sweep sounded from different places in the room. At the end of five rounds, another distance localization test was conducted.

- Guided training phase 2: This was a repeat of the guided training phase 1. This was to see if there is any further training has any impact on localization accuracy or would there be a plateau effect.

In all the testing rounds, the sound was played only once from the loudspeaker. The subject could take their time to decide on their answer.

3.3.2 Experiment setup

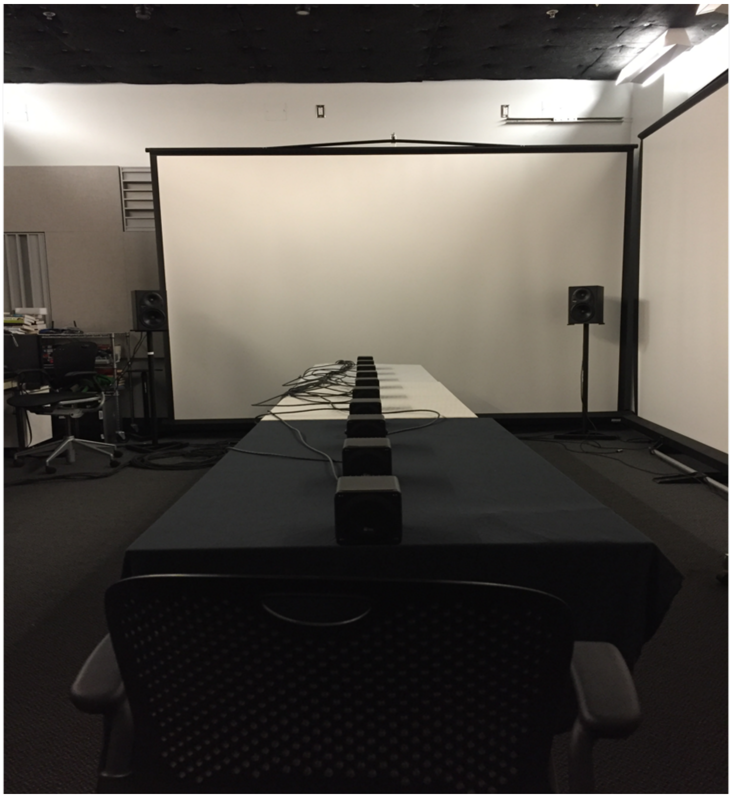

Figure 13: Experiment setup of main study

The experiment was conducted in a room in Elings Hall in University of California, Santa Barbara. A picture of the room with the speaker setup is shown below. The room is rectangular and has the dimensions – 25ft x19ft x 11ft (LxBxH). The walls are partially covered with sound absorption panels. The room is not very live, but it isn't a totally dead space either[1].

[1] No measurements were done to measure the room acoustics.

Ten identical Meyer loudspeakers were used for the experiment. They were placed in an array fashion, one behind the other, and 1.5ft apart. The first speaker was 1ft away from the listener. This configuration allowed the first 2 loudspeakers to be within 1m from the listener, the third ~1m from the listener and the rest farther away. Of the ten speakers, speaker 4 and speaker 7 were dummy speakers and sounds were not played from these speakers. This was done to obtain some a greater resolution of localization accuracy. The participants were not informed that speakers 4 and 7 were dummy.

3.3.3 Procedure

Twenty-four people (15 male and 9 female), mainly graduate students from University of California, Santa Barbara with varied experience with concepts of spatial sound and sound localization took part in the experiment. Their mean age was 24.12 (SD = 2.25). Although background information on their familiarity with spatial sound concepts and their prior participation in sound localization experiments were collected, this factor was not eventually used in analysis. All participants fell into the same pool. No participant had been to the room before this experiment.

3.3.4 Pilot Study

Prior to the experiment described above was conducted; a pilot study very similar experiment was conducted as a pilot study. The intention was to get a sense of the audio settings and to verify the null hypothesis before proceeding further with the study. The pilot study was conducted in a room in the Bobst Library at New York University. A picture of the experiment room is shown above. The room was rectangular and had the dimensions – 17.5ft x 13.5ft x

10ft. This space was fairly reverberant[2]. Eight Yamaha loudspeakers were used in the experiment. There were no dummy speakers used in the pilot study. The rest of the setup was the same.

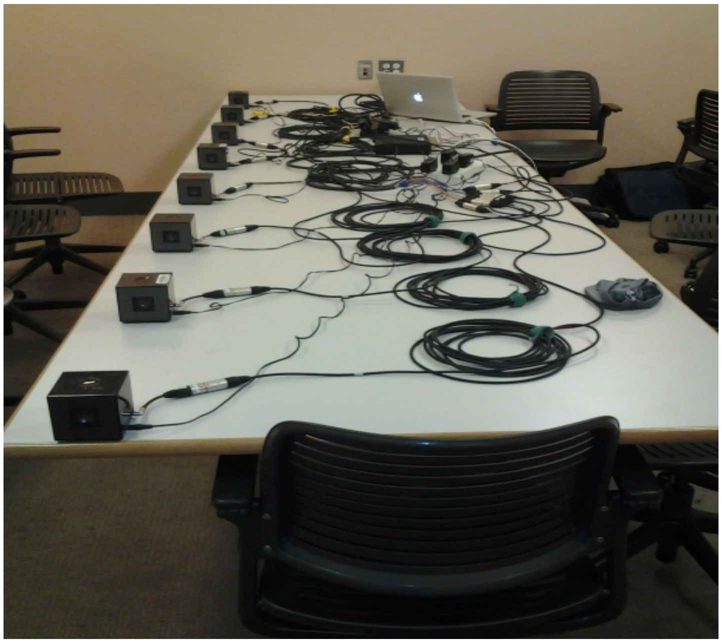

Figure 14: Experiment setup of pilot study

Thirteen graduate students (9 male and 4 female) of New York University with varied experience with concepts of spatial sound and sound localization took part in the experiment. Their mean age was 28.07 (SD = 3.48). Although background information on their familiarity with spatial sound concepts and their prior participation in sound localization experiments were collected, this factor was not eventually used in analysis. All participants fell into the same pool. No participant had been to the room before this experiment.

[2] No measurements were done to measure room acoustics.

3.4 Results and Discussion

Data collected was analyzed using SPSS software. A repeated measures ANOVA was performed to measure the mean error of the participants across the testing phases. Thus, the measure of error in performance became the dependent variable and three/four testing phases became the independent variables. This gave a measure on how accurate/inaccurate they were across the phases. If the hypothesis were to be satisfied, there should be a reduction in the mean error in performance across the testing phases.

The tables below show the mean error and standard deviation across all participants in the pilot and main studies. Corresponding plots visualize the data in a simple manner.

Table 1: Descriptive statistics from the pilot study

Session	Mean Error (ft.)	Standard Deviation
T1 (Pre Training)	2.043	2.002
T2 (Self Training)	1.966	1.912
T3 (Guided Training 1)	1.280	1.761

p = 0.003 (Wilks' Lambda multivariate test)

Table 2: Descriptive statistics of main study

Session	Mean Error (ft.)	Standard Deviation
T1 (Pre Training)	3.671	2.981
T2 (Self Training)	2.824	2.452
T3 (Guided Training 1)	1.295	1.418
T4 (Guided Training 2)	1.280	1.415

p < 0.001(Wilks' Lambda multivariate test)

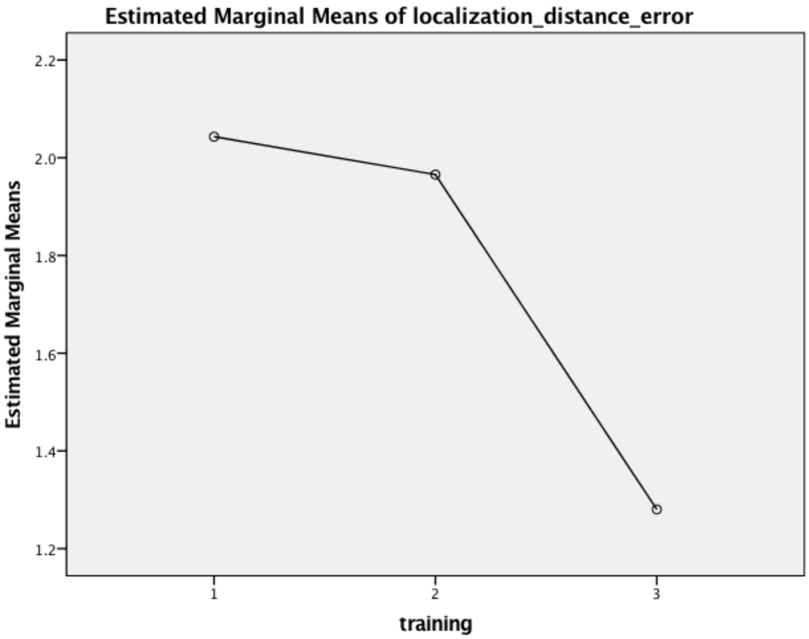

Figure 15: Plot showing the mean error across testing phases (pilot study)

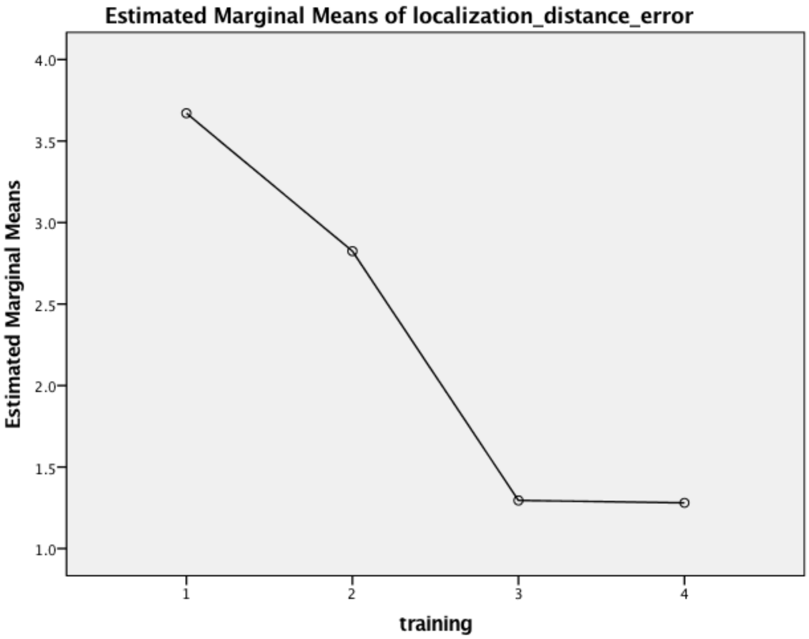

Figure 16: Plot showing the mean error across testing phases (main study)

Discussion

Initially, for the pilot study, a one-way repeated measures analysis of variance (ANOVA) was conducted to evaluate the change in distance localization of the participant across the testing phases with a subject pool (N=13). The results showed a drop in the mean error from 2.043ft in the first testing phase to 1.28ft in the last testing phase. The was also found to be statistically significant, Wilks' Lambda = 0.901, $p<0.005$, $\eta^2=0.099$.

Then a pairwise comparison revealed that the pairwise difference was not statistically significant between testing phases 1 and 2. This along with $\eta^2=0.099$, showed that the effects were small. This could be draw to the subject pool being small to draw a definite conclusion and a bigger study with a greater subject pool would be necessary.

With the main study, once again a one-way repeated measures ANOVA was conducted to evaluate the null hypothesis that there is no change in the participants' estimation of sound source distance when measured with no training, self training and guided training with a subject pool (N=26). The results of the ANOVA indicated a significant effect, Wilks' Lambda = 0.563, $p<0.001$, $\eta^2=0.44$. Thus there is significant evidence to reject the null hypothesis.

In addition, pairwise comparisons indicated that each pairwise difference, barring the guided training 1 and guided training 2 was significant, $p<0.01$. This seemed to suggest that the effect of the environment and familiarity with the acoustics of the room had a significant effect in the improvement in estimation of the sound source distance. However, the plateau effect noticed at the end is not statistically significant.

3.5 Summary

An experiment was designed to test the effect of training on the estimation of sound source distance and successfully conducted. The results and subsequent analysis shows that there is a drop in the mean error in performance of the participant with training. This drop in error is significant from test 1 to test 2 and test 2 to test 3, but not as significant from test 3 to test 4. This seems to suggest that the training, even self-training definitely had an effect on the performance of the participant. He/she became more acquainted with the acoustics of the space and could localize better along the distance axis. Mean error fell from 3.671 in test session 1 to 2.824 in session 2. That is a 23.07% improvement in performance with just self-training indicating that the hypothesis is supported. The p-test showing statistical significance of the data further supports this hypothesis. This gives an insight into how the acoustic distance perception of an acoustic space changes over time.

This study presented us with an understanding on how one's perception of sound source distance changes as they learn the acoustics of the space. Follow up studies can see, how much retention of this is there. Localization studies along azimuth and elevation in a similar manner would also help us draw how the total acoustic image of a space changes in humans over time as they become more familiar with the acoustics of the room.

4. Spatialization Techniques & Music Genres – Loudspeaker Study

Sound spatialization has interested artists and researchers alike for a number of decades. While the researchers have focused on the ability to effectively spatialize sound in different loudspeaker configurations as well as bring in an immersive environment into headphone listening, artists have been focusing on trying to create various compositions and installations, which utilize these loudspeaker configurations to create interesting effects and pieces. This study aims to understand if there is any preferences or biases for certain spatialization techniques with certain genres of music and what the underlying reason could be for any such preferences. This is done with a user study in a listening room over loudspeakers. Results showed that subjects preferred the Vector Base Amplitude Panning (VBAP) configuration for all genres except Rock music where they preferred Ambisonics method of delivery.

4.1 Introduction

An efficient workflow for sound spatialization requires structure, flexibility, and interoperability across all involved components [76]. In addition, one requires a solid understanding of the spatialization systems, their capabilities and limitations. The two most common techniques for spatial audio reproduction are Vector Base Amplitude Panning (VBAP) and Ambisonics. These have the ability to place the sound sources anywhere on a surface represented by a loudspeaker array [77]. In recent times, another approach, Distance-Based Amplitude Panning (DBAP) attempts to pan sounds in two and three-dimensional spaces [77].

The majority of studies evaluating the performance of the aforementioned spatial audio reproduction techniques have involved using psychoacoustic test signals such as broadband noise or sine sweeps and have mainly focused on 2D and 3D localization of the sound [78,79,80,81] and on rare occasions, sound coloration [82, 83].

Learning from these studies and putting the theories into practice has resulted in stand-alone applications. These applications could be direction-based frameworks such as SSR [84] and Zirkonium [85] or diffusion approaches like Scatter [86].

Significantly less research has happened with respect to evaluation of these reproduction techniques with more ecologically valid stimuli such as music. This study, therefore, aims to bridge that gap by comparing these different spatial audio reproduction techniques with different genres of mainstream music.

As part of this project, a user listening study was designed and conducted which involved subjects listening to different musical genres played back using the aforementioned different reproduction techniques. The idea behind this study was to see if there is a preference for a particular reproduction technique for a specific genre of music.

This research has both scientific and artistic significance. On the science front, we could learn in greater depth the human perception of sound spatialization and through the quantitative analysis of the binaural audio, we could get insights as to what musical parameters might be influencing decisions on the various sound reproduction techniques under investigation. While on the artistic front, we could gain knowledge of the effective use of each of the spatialization techniques with different audio input. Given that the study is being conducted from a human perception standpoint as opposed to simulation, the hope is that this paves way for the creation of more engaging spatial audio installations, auditory displays and compositions.

The chapter is organized in the following manner: Section 4.2 gives a brief overview of the background of the field with brief descriptions of the spatial audio reproduction techniques under comparison. Section 4.3 describes the methodology and the design of the experiment. Results and discussion are covered in section 4.4 while section 4.5 concludes the chapter with a succinct summary.

4.2 Background

Early efforts of creating phantom images in spatial audio came from the Bell labs in the US and simultaneously from Blumlein in the UK. From these early developments, came what is known as equal power panning for loudspeakers. This is an amplitude panning technique using a panpot that controls the gains on each pair of loudspeakers, wherein equal gain on both the left and right loudspeaker would give rise to a phantom image at the center. This is an effective panning law that is used even in the present day in stereophonic reproduction.

The basic idea behind amplitude panning is to steer the perception of an individual sound object/event in a particular direction by varying the gains of the signals applied to each loudspeaker. In the general stereophonic setup, this can be formulated as

$$s_i(t) = g_i s(t) \qquad (1)$$

where $s_i(t)$ is the signal applied to loudspeaker i, g_i is the gain factor applied to the corresponding loudspeaker and t is the time parameter.

4.2.1 Ambisonics

Ambisonics is a powerful spatial sound rendering technique based on the decomposition of the soundfield using spherical harmonics. Gerzon described this process initially in [29]. In his paper, Gerzon describes the approach to Ambisonics as “For each possible position of a sound in space, for each possible direction and for each possible distance away from the listener, assign a particular way of storing the sound on the available channels. Different sound positions correspond to the stored sound having different relative phases and amplitudes on the various channels. To reproduce the sound, first decide on a layout of loudspeakers around the listener, and then choose what combinations of the recorded information channels, with what phases and amplitudes, are to be fed to each speaker. The apparatus that converts the information channels to speaker feed signals is called a “decoder” and must be designed to ensure the best subjective approximation to the effect of the original sound field” [87]. This is different from the Blumlein approach touched upon earlier in a number of ways. Firstly, the initial encoding stage is removed from the eventual playback system, its sole aim being to capture as much information about the sound scene as possible using a certain number of channels. The decoding stage can now use the recorded spatial information to determine the appropriate loudspeaker signals that will recreate this spatial scene.

In the encoding stage, we can encode a monophonic signal s by multiplying it with a vector **Y** which has spherical harmonic coefficients as $Y_n^m(\theta, \emptyset)$ such that:

$$\mathbf{b} = \mathrm{s}\mathbf{Y} \qquad (2)$$

The spherical harmonic function is computed as:

$$Y_n^m(\theta, \emptyset) = N_n^{|m|} P_n^{|m|}(\sin(\theta)) \begin{cases} \cos(|m|\theta) & if\ m \geq 0 \\ \sin(|m|\theta) & if\ m < 0 \end{cases} \qquad (3)$$

where ϕ is the azimuth angle, θ is the elevation angle, n is the Ambisonics order, m is the degree of the spherical harmonic, $P_n^{|m|}$ is the Legendre function and $N_n^{|m|}$ is the SN3D normalization term.

The spherical harmonics described above can be visualized in the figure below. From this, using Ambisonic order, n = 1 as an example, we can realize that capture of the soundfield can be done using 4 microphone capsules in a tetrahedral configuration that would capture the entire 3-dimensional field.

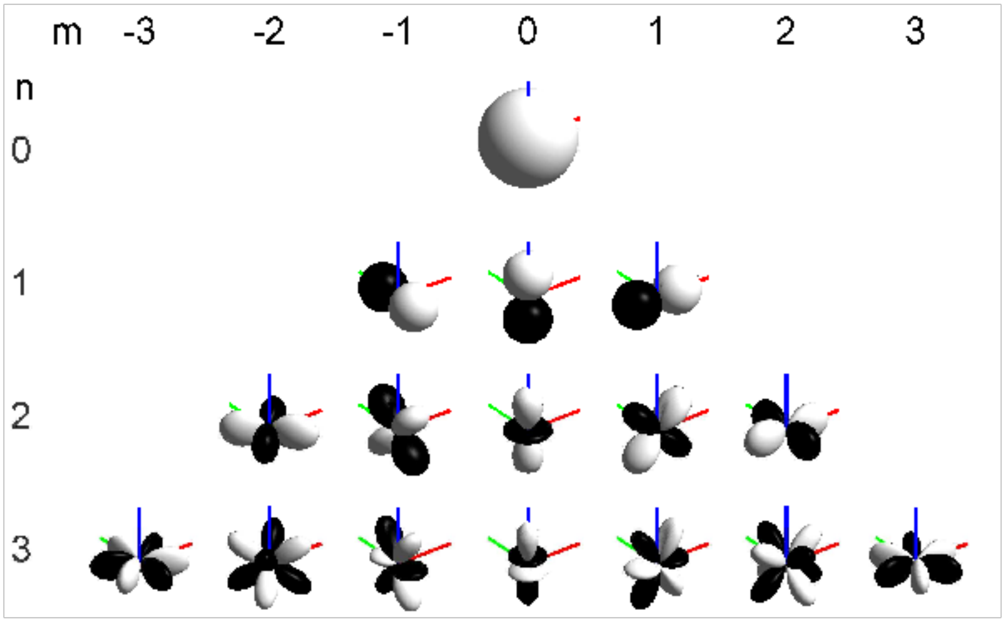

Figure 17: Pictorial representation of Ambisonic B-format up to 3rd order

In the decoding stage, we multiply **b** with the inverse of the loudspeaker encoding matrix such that,

$$\mathbf{g} = \mathbf{D}\mathbf{b} \qquad (4)$$

where **g** is the gain of the loudpseaker, and **D** is the inverse of the loudspeaker encoding matrix. In this project, ambisonic decoding was done according the MaxRe weighting as described by Gerzon in [88].

4.2.2 Vector Base Amplitude Panning (VBAP)

Ville Pulkki's vector base amplitude panning (VBAP) is one of the most robust and generic amplitude panning algorithm that works on any arbitrary speaker layout. With this, it is possible to create two- or three-dimensional sound fields.

In VBAP, the amplitude panning method is reformulated with vector and vector bases. The reformulation leads to simple equations for amplitude panning, and the use of vectors makes the panning methods computationally efficient [36].

Three-dimensional VBAP (used in the current experiment) generally involves the typical two-channel stereophonic setup and a third loudspeaker that is placed in a position roughly the same distance from the listener as the other two loudspeakers and in a plane that is different from the plane of the listener and the two loudspeakers. The virtual source can now appear within a triangle formed by the loudspeakers when viewed from the listener's position as shown in the figure below. This is known as the active triangle and no matter the listener's position; the virtual sound source cannot be positioned outside of the active triangle. This limits the maximum localization error of the virtual source to the dimensions of the active triangle. This also means that the denser the loudspeaker setup, the smaller the triangles are and the lesser the localization error possible. The virtual source position perceived by the listener is governed by the gain factors applied on each loudspeaker in this setup and can then be realized as,

$$g_1^2 + g_2^2 + g_3^2 = C \qquad (5)$$

where g_1, g_2, g_3 are the gain factors of loudspeakers 1, 2 and 3 respectively and C is a constant value of sound power. The sample configuration of this setup is shown in the diagram below.

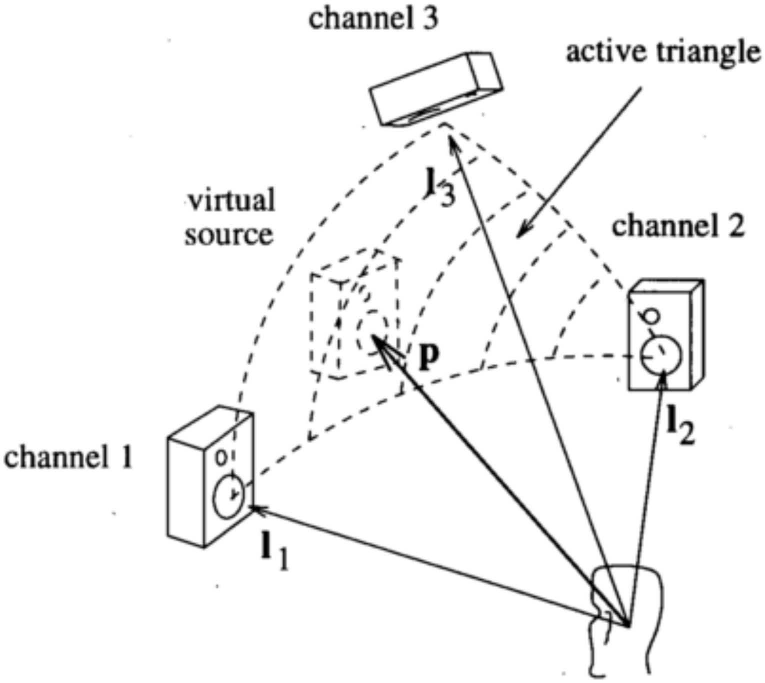

Figure 18: Sample configuration of three-dimensional VBAP. Loudspeakers form a triangle into which the virtual source is placed. Figure taken from [36]

4.2.3 Distance-Based Amplitude Panning (DBAP)

Similar to VBAP, distance-based amplitude panning (DBAP) works on arbitrary speaker positions. In addition to that, the listener positions are also not taken into account, thereby the algorithm rendering itself well to a larger listening area, not confining it to a narrow “sweet spot” as in the case of Ambisonics and VBAP. Thus, DBAP is seemingly more suited for real-world applications like concerts and sound installations.

DBAP is defined as a matrix-based spatialization technique that takes the actual positions of the speakers in space as the point of departure, while making no assumptions as to where the listeners are situated [37].

DBAP works on the principle of equal intensity panning. It extends this from a pair of loudspeakers (in a standard stereo setup) to multiple channels in a loudspeaker array of any size, with no assumptions on the loudspeaker configuration in terms of their position in the cartesian space or relative to one another. Assuming the cartesian coordinates of the virtual source are (x_s, y_s), then the distance d_i from the source to each of the loudspeakers can be given as,

$$d_i = \sqrt{(x_i - x_s)^2 + (y_i - y_s)^2} \qquad (6)$$

where "i" represents the ith loudspeaker.

Equal intensity of panning is controlled by making the assumption that the intensity is constant regardless of the position of the virtual source. Mathematically, this is denoted as,

$$I = \sum_{i=1}^{N} v_i^2 = 1 \qquad (7)$$

where v_i is the amplitude of the ith loudspeaker.

4.3 Methodology

Here the methodology of the experiment to compare different spatial audio reproduction techniques with different musical genres is outlined. The null hypothesis being that there is no preference for any reproduction technique for a particular genre. In other words, the method of playback did not have an impact on the musical genre.

4.3.1 Apparatus

The experiment took place in the audio research lab at the Hochschule für Gestaltung (HfG) Karlsruhe. The lab has 29 loudspeakers arranged in a hemispherical configuration. Measured from the sweet spot, the loudspeakers are located at a distance of 2.4m. 16 of the 29 loudspeakers are placed in the horizontal plane, 8 of them at an approximate elevation of 30°, 4 at approximately 60° elevation and 1 directly above the listener (0°, 90°). Figure 2 shows the experiment in progress.

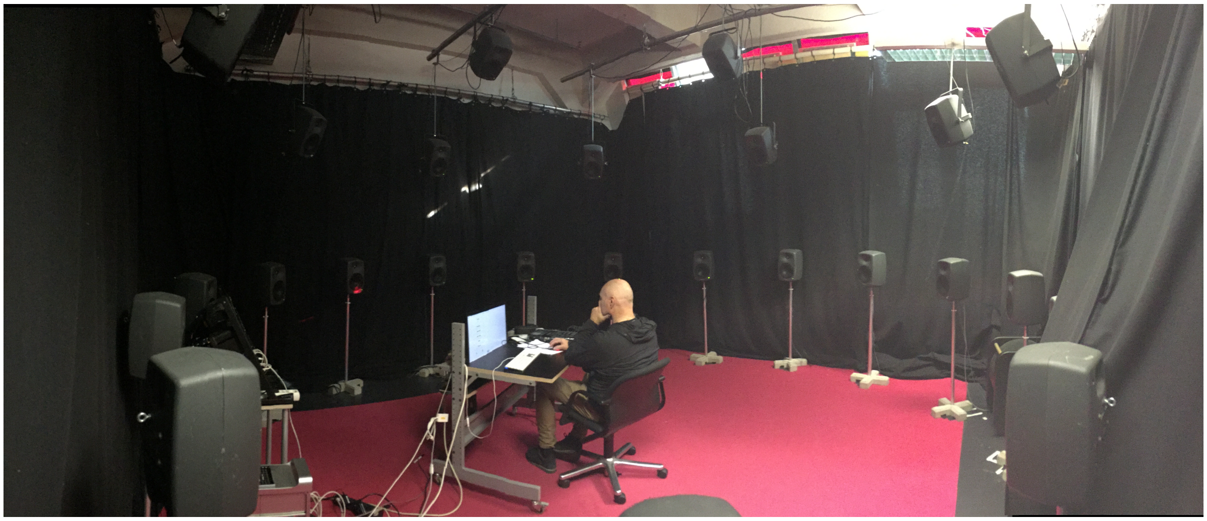

Figure 19: The experiment in progress

A Max MSP patch was designed to handle playback across all three spatial audio reproduction techniques. The IRCAM SPAT [89] patches were used for the encoding and decoding of the Ambisonics signals as well as the implementation of VBAP and DBAP algorithms.

As seen from the interface, it contained 5 attributes on which the participants made selections. The scales were quasi-continuous that could obtain values between 1 and 128 with a step interval of 1. The two buttons A and B played back the two musical excerpts under comparison. The button, “next trial” proceeded to the next test pair with no option to return to the previous trial. The trial number was also displayed at top of the interface.

4.3.2 Stimuli

As stated earlier, four different music genres were selected for the purpose of this study. 40-50s excerpts were selected from each song. All four stimuli were mixed and mastered in 5.1 by the same engineer. This prevented the tracks from having any mixing biases different engineers might have. The stimuli were: 1. A 42s excerpt of Dvořák: Symphony No. 9 in E Minor, “From the New World”: IV Allegro Con Fuoco. 2. A 41s Electronic excerpt by an unknown composer. 3. 49s excerpt of the soft rock song Hey 19 by Steely Dan and 4. A 39s long excerpt of an Afrobeat song Aramile by Babatunde Olatunji. All the stimuli used in this experiment were mixed and mastered by audio engineer, Paul Geluso. This ensured that the mixing & mastering strategies and decisions had the same thought process and did not create any engineer biases that could have altered the dynamics of the mixes and added another variable to the experiment. I interviewed Paul which is detailed in the following section where he talk about his journey and how that impacted the decisions he takes as an audio engineer, his thoughts on spatial audio, the progression of the field and even Dolby Atmos.

4.3.3 Interview with Paul Geluso

Paul Geluso a music assistant professor and director of the music technology program at New York University (NYU). Paul Geluso’s work focuses on the theoretical, practical, and artistic aspects of sound recording and reproduction. He has been credited as producer, recording engineer, mastering engineer, and/or musician on hundreds of commercially released recordings, including Grammy nominated and Latin Grammy nominated titles and awardwinning films. His research focusses on new ways to capture, mix, and process immersive audio for playback on multi-channel sound systems, recently co-editing "Immersive Sound: The Art and Science of Binaural and Multi-channel Audio" published by Focal Press-Routledge [90].

Shashank Aswathanarayana (SA): To begin with, can you talk a little bit about journey as an audio engineer? Did you start with live sound and then move to the studio or the other way around?

Paul Geluso (PG): My journey started very young at home. My dad built a studio at home for my older brother when I was 5. We had a 4 track Tascam 3340 and my brother built a console at home with parts from radio shack and a piece of plywood. I grew up wearing headphones, playing drums and recording. I cannot remember a life without a recording studio around me. So, in summary, my journey began in the studio for sure.

SA: When and how did you come across multichannel sound?

PG: Through working at Harvestworks Digital Media Arts in the 1990s, where I worked as the chief sound engineer, I got exposed to the arts scene and multichannel sound. Here I designed systems that didn't exist. These were site specific systems for art/sound installations and performances. I quickly got away from stereo and moved into the immersive sound/3D audio realm, but it probably didn't even have that name at that time. Since 90s, I've been working off the map when it came to speaker systems.

SA: When did you first work with Ambisonics?

PG: This happened in the late 90s. We at NYU were working with composer/sound artist Charlie Morrow who was interested in live Ambisonics. We had a cube, and I was mixing for the cube in Ambisonics, but Ambisonics was not very popular then as the professional audio world had rejected it saying it was too phasy and didn't have much control or the punch needed. My phone started ringing about Ambisonics only when VR started coming out. At this time, all of a sudden, they had a reason to use more immersive mixes, mixes that were half baked, and mixes that captured the entire soundfield around the listener.

SA: Was this also the time when you were thinking about new recording techniques for multichannel mixes?

PG: Yes, through my research at NYU, I started working in developing new ways to capture 3D audio working with Wilfried van Baelen (inventor of Auro3D), Gregor Zielinski (Sennheiser), Tom Ammerman (New Audio Technology GmbH) and Agnieszka Roginska (NYU). We called ourselves the 3D gang and we would meet discuss ideas. We were an eclectic group of people. Later David Bowles (Swineshead Productions, LLC) also joined the group, and we would meet off-campus, rent a studio, and do our 3D audio thing. This was even during AES (Audio Engineering Society) Conventions when we would have 3D audio events off venue. AES at that time didn't have immersive sound. Finally, more recently AES has started to adopt immersive rooms driven by the speaker manufacturers, Genelec and PMC.

SA: You speak a lot about designing new systems and ways to capture, what's your take on Dolby Atmos, which is being projected as the solution for everything immersive sound?

PG: Well, Atmos is great for certain situations. For instance, sitting on a chair in the movie theater, Atmos works great. With VR over headphones, not so much. I don't think it's compatible and I'm not convinced with the mix I get. Another instance where Dolby Atmos fails is when you're panning around the system, I don't know what they're using under the hood, but it looks like a focus thing. The loudest sound in 1 speaker and lot of the others also carry it faded. This is unlike say in VBAP where is only 3 speakers which carry the sound.

When you have a lot of speakers carrying the same sound as that, you get phasing problems as I’ve observed with Atmos. Therefore, even with Atmos, I try to use a hybrid approach. It’s got sound objects, which are great, and I use them, but I also want to use standard pop/film mixing ideas within it. You cannot have one system and method for all purposes.

SA: Can you talk a little more about this hybrid approach you mentioned?

PG: A very basic example would be what I call stereo+. This is basically mixing with stereo and adding some binaural and reverb effects to the mix so that it’s compatible with speakers and on headphones, the binaural wakes up. So, you have the main vocals, or the most important thing locked and other atmospheric sounds that give you the sense of immersion. I think this is entertaining. I’m experimenting for VR with unity mixes actually. I’m attempting to create virtual environments and having multiple camera views giving different perspectives, but certain omni sounds remaining the same. *(On a side note: I have a prototype described in chapter 6 is of a virtual environment. This is not what Paul Geluso describes above, but the creation of the prototype comes from the same line of thinking)*.

SA: What is your mix philosophy?

PG: My mix philosophy is quite simple. It is to convey the artist’s idea to the audience in the clearest possible way. The next thing to keep in mind is listener fatigue. You don’t want them to find it hard/squint to listen. If they squint to listen, they’ll turn it off. This is very important. People talk about balance in the mix, but sometimes forget this basic thing. You need to give a good listening experience to the consumer and hand it in a plate and not make them work for it. With that as the starting point, I look for what is the most important aspect of the music. For instance, in pop, it’s the vocal and the groove. So, I start with that and then bring the other elements in. That’s not to say the other elements like the bass is not important. They absolutely are. However, if the vocal and the groove are not clear and upfront, then the listener is squinting, which is never a good thing. On the other hand, in Latin music, the vocal sits much lower than the percussion, so it’s important to understand the music as well. When immersive sound comes into the picture, we add another layer to the whole thing. A perceptual concept called concert realism comes into the picture. It’s not just about transporting the consumer to the concert hall. They want more than that. They want the best experience. They want to be on their couch but feels like he/she is in the best seat in the house. For this, purely capturing the soundfield and reproducing it is not sufficient. You need to go beyond that. That’s what goes into the surround part of the mix to create the impression of being transported into a new world. This cannot be done only with reverb.

4.3.4 Procedure

Participants were all given the same information prior to the start of the experiment and consent of participation was taken. They were then asked to fill out a brief background questionnaire on their musical background before the tasks of the experiment were explained to them.

This experiment was designed as an A/B user study using four different music genres (electronic, soft rock, western classical and Afrobeat percussion) and the three different reproduction techniques (VBAP, DBAP and Ambisonics). The subjects listened to all genres reproduced by all techniques in a random order and rated their preferences given a number of criteria such as immersion, spatial clarity, sound quality among others. Each criterion was clearly explained to the subject beforehand and a definition of each of them was also included in the user interface so as to not ensure uniformity in interpretations and expectation across subjects. The participants were told that they could playback the excerpts any number of times before rating their preferences. Playback was randomized for all sets so as to not create any order biases.

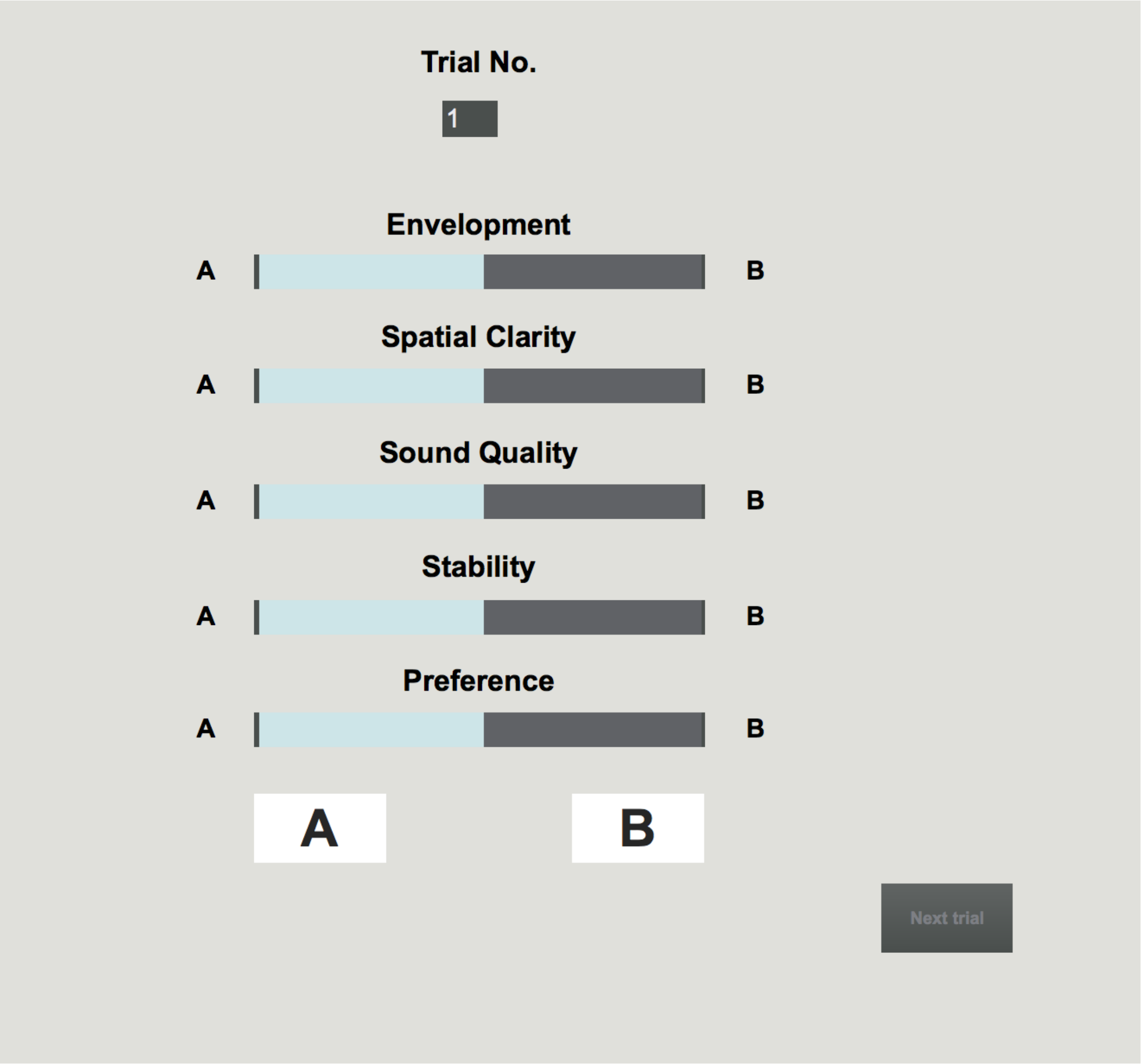

Figure 20: The Max/MSP Patch of the graphical user interface

Each criterion was defined as follows. These definitions were inspired from previous studies as in [17] and [18]:

- Envelopment – The degree to which the sound envelops the listener, giving a sense of immersion.
- Spatial Clarity – The degree to which each of the elements of the music could be clearly decipherable and localized.

- Sound Quality – The degree to which the music rendered was without any noticeable audio colouration or distortion.
- Stability – This was defined as how stable/sensitive the sound image was to head movements. Preference – This was just a straight up choice of which of the two pieces in the trial provided a better overall listening experience and to what degree.

Typically, participants listened to each pair twice to come up with judgements to each attribute. Therefore, with the number of trials in question, the participants spent about 80mins doing the experiment. No participant, however, complained of experimental fatigue. Following the completion of all the trials, the author conducted a short informal interview of the participants take on the experiment and any difficulties they had either understanding the experiment of doing the experiment.

4.4 Results and Discussion

A total of eleven people (6 male and 5 female) took part in the experiment. Participants were mainly students, faculty and researchers from Hochschule für Gestaltung (HfG) and the adjacent Zentrum für Kunst und Medien (ZKM) | Center for Art and Media, Karlsruhe. Their mean age was 38.72 (SD = 12.11). Other background information collected was on their experience with spatial audio and general musical experience although none of this information was eventually used in the analysis.

Pairwise Comparison

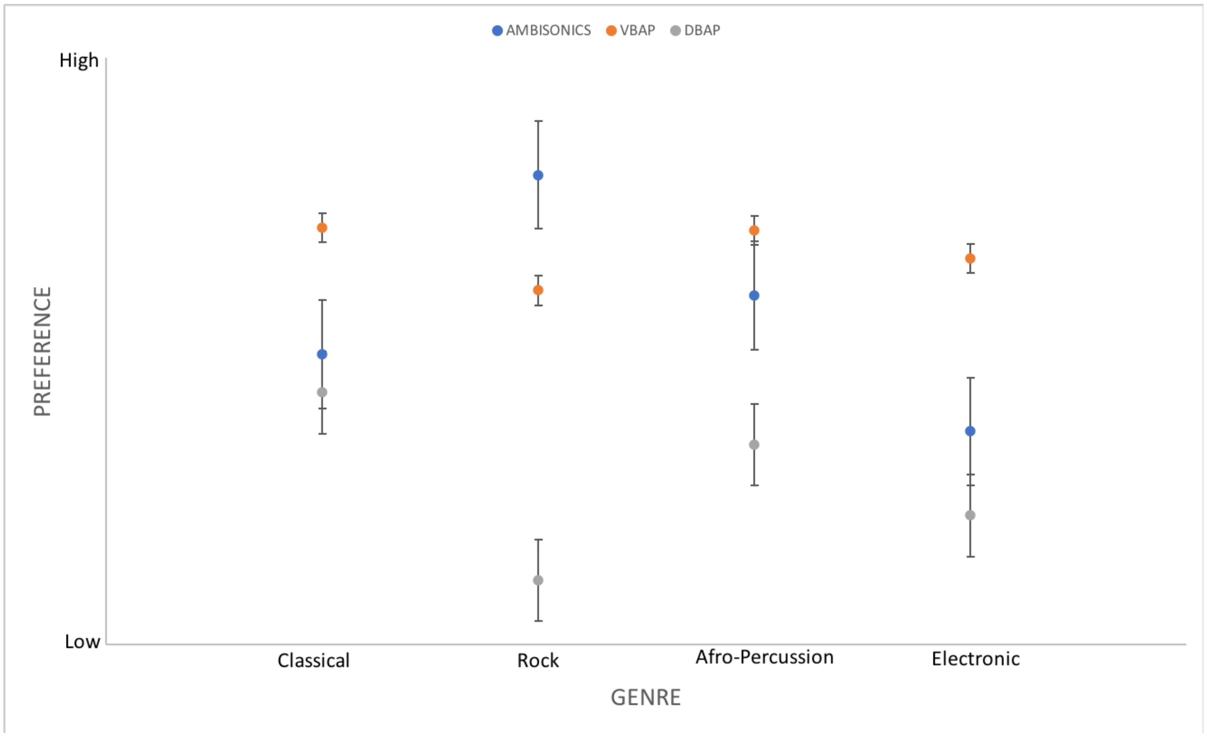

Figure 21: Preference scales and standard errors for the different spatial audio reproduction techniques and the four different music genres.

The analysis of the pairwise comparison data was done individually for each participant and the scales for each piece and each sound reproduction technique was separated. Normalized ratings (between 0 and 1) were used in each of the scales and was used as the frequency with which each participant selected the corresponding algorithm used in spatializing the sound as is typically done in a scaling experiment of A/B comparison data. Thurstone Case V procedure [93] was then computed for each musical piece and spatial reproduction algorithm. This is a useful scaling method typically used in A/B comparisons as it allows us to make the assumption of equal variances and uncorrelated distributions as would be required when scaling participants individually.

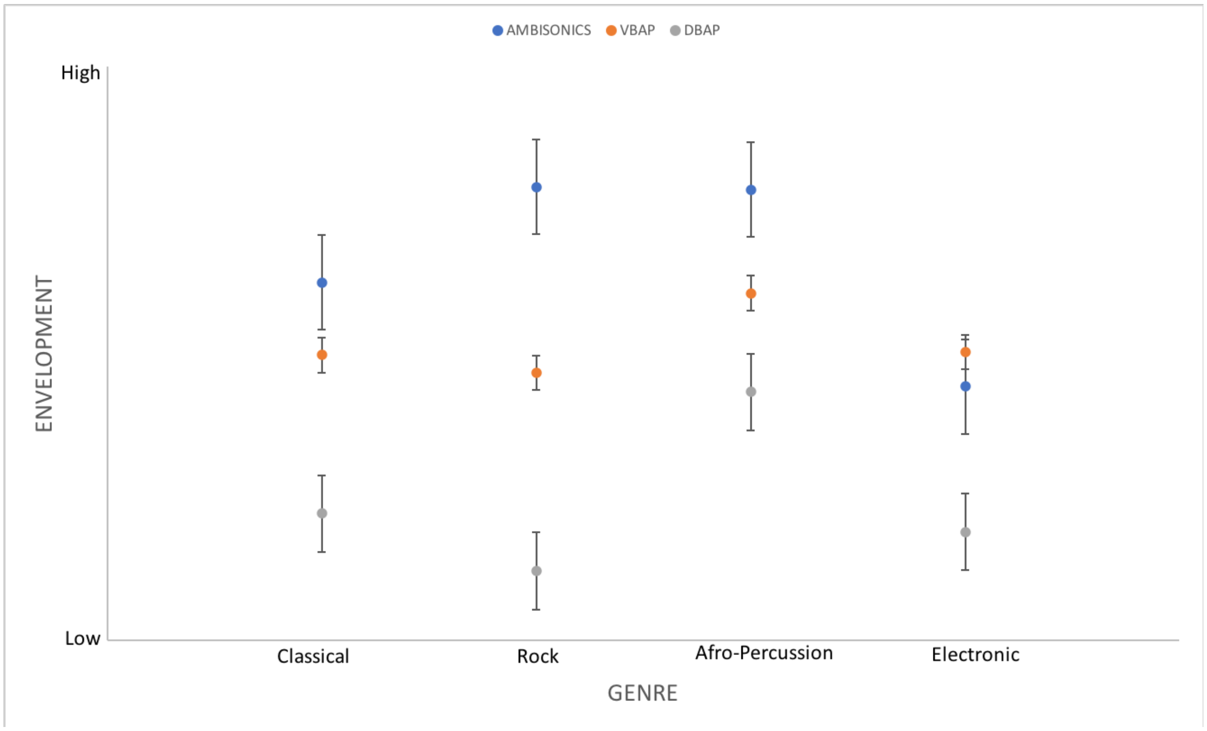

Figure 22: Envelopment scales and standard errors for the different spatial audio reproduction techniques and the four different music genres.

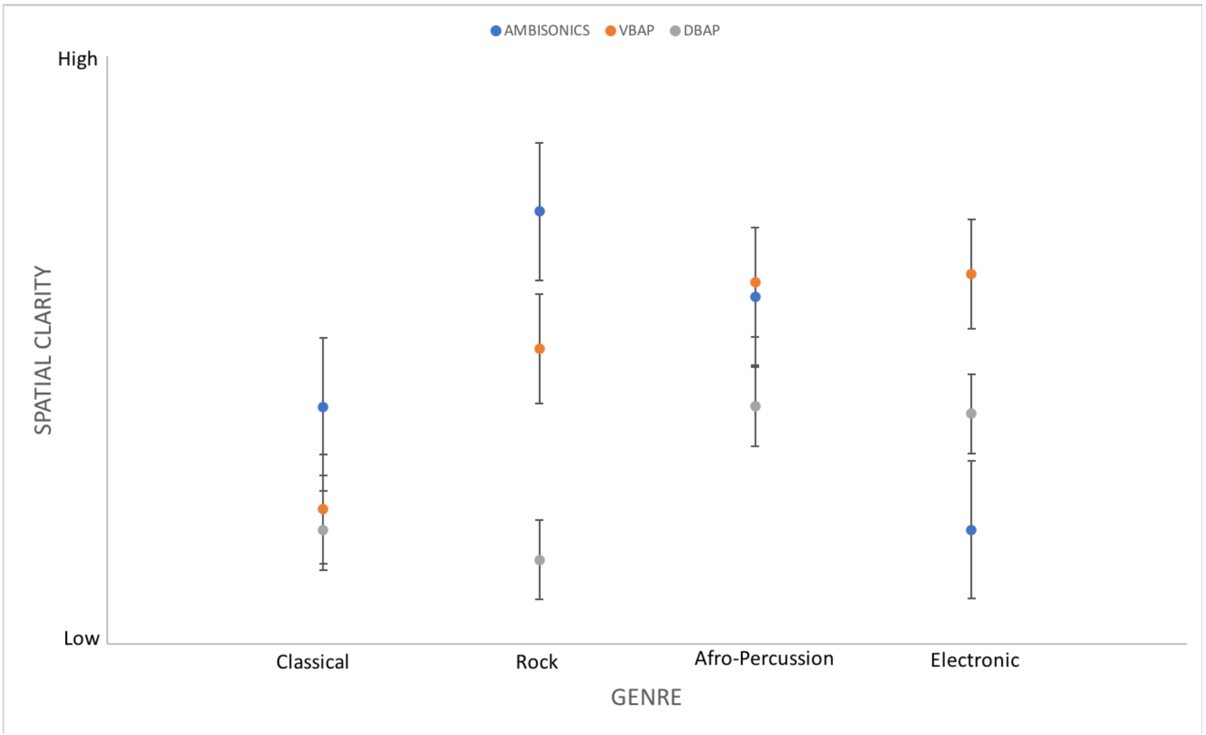

Figure 23: Spatial Clarity scales and standard errors for the different spatial audio reproduction techniques and the four different music genres.

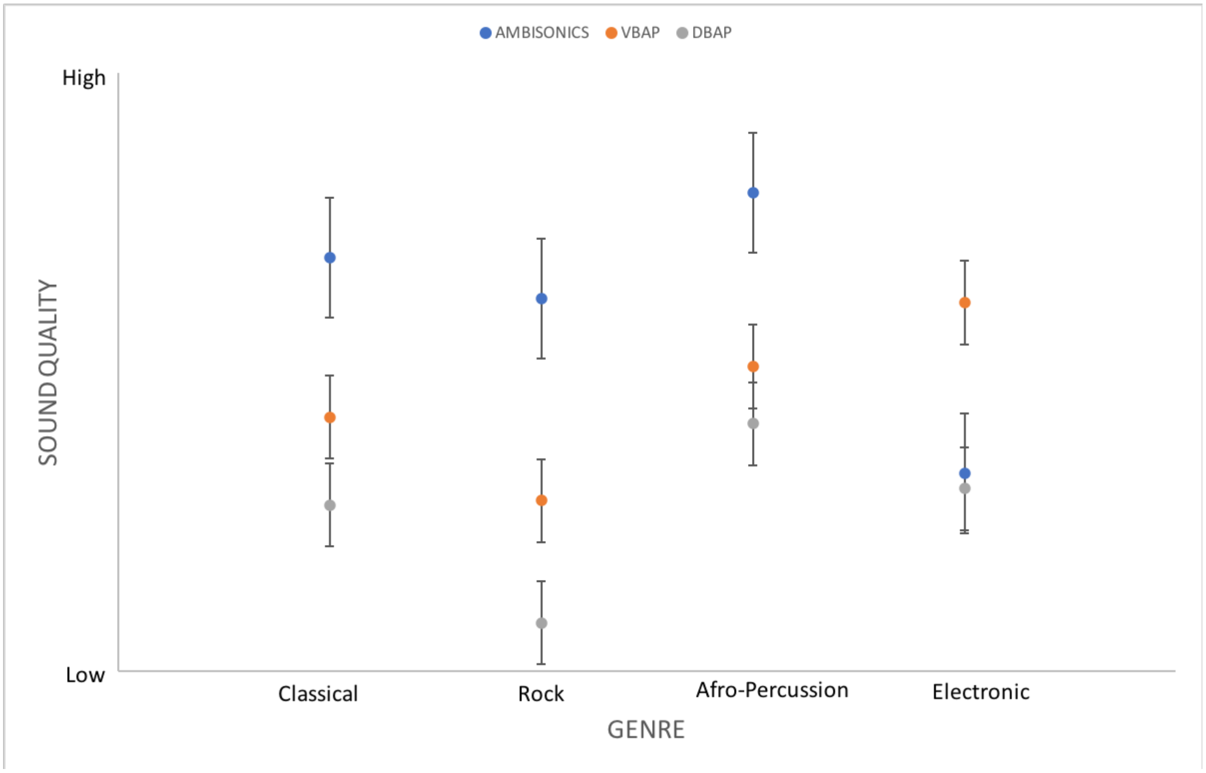

Figure 24: Sound Quality scales and standard errors for the different spatial audio reproduction techniques and the four different music genres.

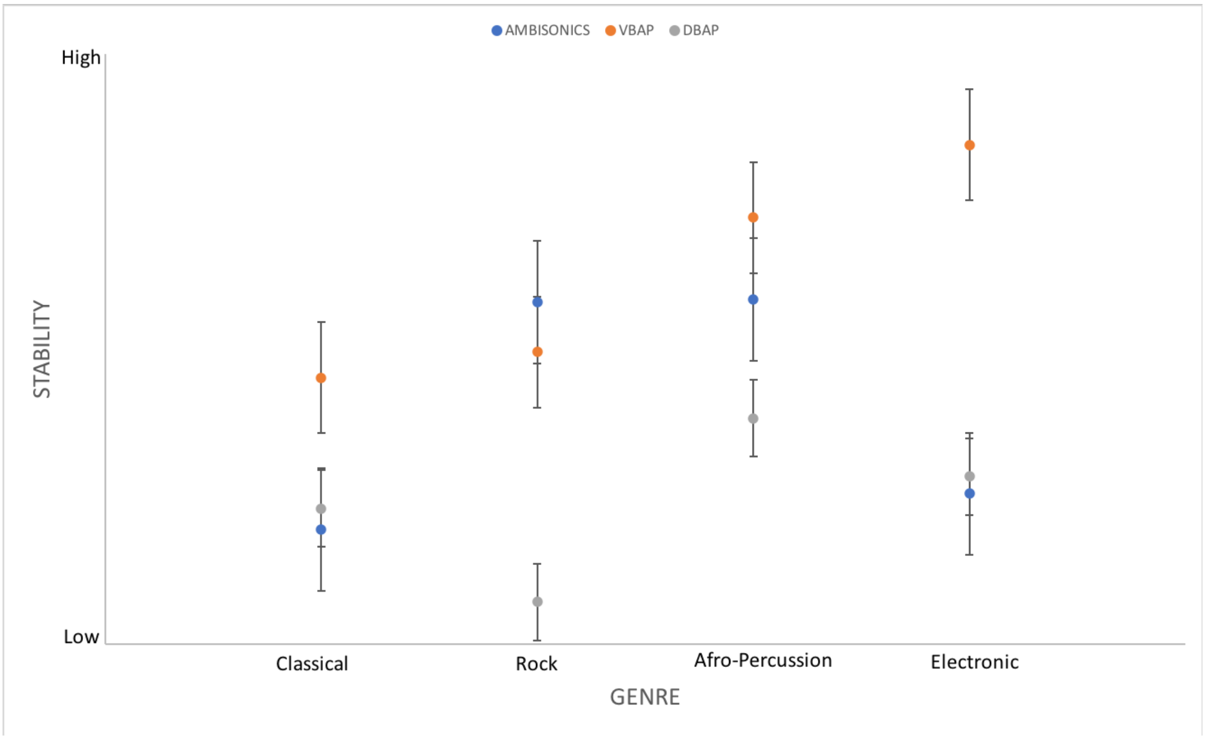

Figure 25: Stability scales and standard errors for the different spatial audio reproduction techniques and the four different music genres.

Mean scales with their standard errors for all the attributes are plotted from Figures 4-8. The means are unnormalized to depict the magnitude of the perceived difference between the algorithms and the standard errors give us the variability in the dataset.

From the plots, we can infer that certain algorithms did well for certain attributes, and this also seems to indicate the preference of a particular algorithm for a particular genre.

From figure 4, it is evident that VBAP was the preferred reproduction technique for the all the music genres except Rock music where Ambisonics was the preferred choice.

For envelopment, Ambisonics was selected as the most preferred method across all the music genres except Electronic music for which VBAP was selected.

Figure 6 suggested that both Ambisonics and VBAP had better spatial clarity across all musical genres than DBAP with VBAP being the preferred one for Afro-Percussion and Electronic music, while Ambisonics with highest spatial clarity for Classical and Rock music.

Ambisonics seemed to have the best sound quality for Classical, Rock and Afro-Percussion music and VBAP having the best sound quality for Electronic music.

Lastly, when it came to Stability, VBAP seemed to have the most stable audio for all genres except Rock music where Ambisonics has a slightly higher rating.

Analysis of Variance (ANOVA)

Using the scale values mentioned above, a more detailed analysis of the ratings was done. A one-way repeated measures Analysis of Variance (ANOVA) was performed independently for the different attributes and the pieces used in the experiment. The reproduction technique was used as the independent variable and the scale values for each participant and the corresponding reproduction technique was used as the dependent variable. These results are summarized in Table 1.

With a small sample size, as in this case, it was also important to perform the Mauchly's test for Sphericity. This showed the homogeneity of the data. Where the assumption of sphericity was not met, sphericity corrections were applied using the Huynh-Fieldt (HF) correction ANOVA results. This happened only on 3 occasions (Envelopment & Rock, Spatial Clarity & Electronic, Stability & Afro Percussion) and in each of the three occasions, the corrected results showed statistical significance ($p<0.05$).

For the remaining attributes and pieces, it was observed that significant main effects ($p<0.05$) were achieved for a few pieces and attributes. No noticeable pattern emerged the cases where significant effects were seen or not seen. These results are interpreted in the discussion section.

Table 3: Summary of the statistical analysis for the four pieces in the experiment. One-way repeated measures ANOVA for each attribute and piece.

Attribute	**Genre**	**F-ratio ($F_{2, 18}$)**	**p-value**
Preference	Classical	1.74	0.16
	Rock	1.08	0.11
	Afro Percussion	2.56	0.1
	Electronic	5.9	0.01
Envelopment	Classical	0.57	0.61
	Rock	0.27	0.02
	Afro Percussion	0.68	0.62
	Electronic	3.91	0.03
Spatial Clarity	Classical	0.72	0.5
	Rock	0.81	0.46
	Afro Percussion	2.62	0.1
	Electronic	0.65	0.04
Sound Quality	Classical	0.5	0.6
	Rock	0.73	0.49
	Afro Percussion	0.05	0.94
	Electronic	2.65	0.09

Stability	Classical	3.42	0.05
	Rock	0.64	0.54
	Afro Percussion	3.9	0.04
	Electronic	0.56	0.31

Discussion

Many studies have looked at and evaluated spatial audio reproduction techniques with non-musical test stimuli, but few studies have taken place with musical material. This study aimed at bridging that gap and to evaluate the listening experience and preference of listeners to musical genre with mainstream music material.

It was interesting to note that VBAP seemed to be the preferred option among the listeners for Classical, Afro-Percussion and Electronic music while Ambisonics was preferred for Rock music. One can clearly identify this trend by looking at the other attributes as well, where DBAP consistently scored lower, leading to its preference also being the least.

While the use of non-musical stimuli to conduct comparative studies has its merit, this study has shown that the comparisons using mainstream musical material can also be extremely useful and can provide interesting results.

It has been shown in the past [91, 92, 97] that there is variation in listener's judgments across musical material. Such differences were expected and seen in this study as well. While not all the results displayed significant main effects, given the observations seen in [91, 95] along with the aforementioned expectations, one can safely assume that the lack of significant main effects were due to the subject pool being small. The fact that the sphericity test was met, showing homogeneity of the variables also points towards the same conclusion. The lack of familiarity with the sound material could have also contributed towards this as it has previously been seen [91] that lack of familiarity with the material can obscure the results.

4.5 Summary

In conclusion, a comparative study of Ambisonics, VBAP and DBAP was performed using musical material of 4 different genres. It was observed how each of the techniques influence attributes such as Preference, Envelopment, Spatial Clarity, Sound Quality and Stability. It provided a glimpse at a potential correlation between the musical genre being played and a particular spatial audio reproduction technique providing the most ideal listening experience. This can be concluded based, not only on the preference attribute rated by the listeners based on the pairwise comparison test described above, but also on the other attribute ratings which seem to align with the preference judgment made by the listeners. These were also verified in the one-way ANOVA analysis carried out.

5. Spatialization Techniques & Music Genres – Binaural Study

The main characteristics of a good spatialization algorithm are that the resulting sound image is enveloping, stable, clear. Such a sound image would then not only give a good sense of immersion to the listener but will also enable the listener to clearly localize the different sound elements, and ideally move around or at least have freedom to move the head and not have any major distortions in the listening experience. Part 1 (chapter 4) of this study compared three algorithms, Ambisonics, VBAP and DBAP with four different music genres. In part 2, a more in-depth analysis is done of the results found in part 1. Binaural recordings done using a Neumann KU100 dummy head are used to compare the results found in the loudspeaker study with attributes found in the recordings. It was seen how more bright, next spectrally complex and more spectrally flat signals could have resulted in the user ratings.

5.1 Introduction

Chapter 4 focused on a subjective study with users listening to 4 different genres being played back over a 29-loudspeaker setup comparing the aforementioned three spatialization techniques. Comparison was done using a scaling approach similar to the one described in [91]. Perceptual attributes, envelopment, spatial clarity, sound quality, and stability were rated along with Preference. These attributes have previously been seen in literature as forming an integral part of what listeners consider important in spatial audio listening [92, 96, 97]. These attributes have been used in procedures involving 2D multi-channel setups and less so in 3D rendering as seen here. The understanding, however, is that such attributes are easily transposable. Primary results from part 1 indicated that VBAP was the preferred choice of reproduction technique for three of the genres while Ambisonics was the preferred choice for the fourth.

This part of the study focuses on further analysis of the work done in Chapter 4. The goal is to make inferences on the feasibility of these spatial audio reproduction techniques, the extent of the judgments being influenced by the music material as opposed to the algorithms, the way preference is looked at on the basis of the ratings of the other attributes and finally to understand if any structural differences in the audio signals could have possibly impacted the any of the user ratings. In order to do such an analysis, Binaural recordings of the material were done in the space (described further in the methodology section). Feature extraction and analysis of these recordings is a novel approach to understand the structure of the audio signal impacting the ratings obtained in the loudspeaker study.

The rest of the chapter has been organized as follows: A succinct review of binaural recording using dummy heads (a more detailed review of Binaural recording is present in chapter 2, section 2.1.2). The experiment methodology is then presented, following which are the results and discussion of the work. Finally, conclusions and possible future research directions are drawn.

5.2 Background

The fundamental idea of binaural recording is that in a natural listening environment, we create the auditory impression based only on two inputs, i.e., the sound pressure at our two eardrums. [98, 99, 100]. If these are recorded in the ears of the listener and reproduced exactly, then we can say that we have exactly recreated the complete auditive experience including the timbral and spatial aspects of the sound.

The process/idea of binaural recordings is not new and dates back more than a century. Initial binaural recording devices were developed in the period, 1880-1930 [42]. There are several ways of capturing binaural recordings described in the literature. These include recording at the entrance of the ear canal, recording at the eardrums, or at multiple points along the ear canal between the entrance and the eardrum. Each of the methods has its own set of pros and cons and detailed review can be found in [98].

Another distinct way of recording binaural signals is the use of artificial heads or dummy heads. These are models of an average human head including nose, pinnae, ear canal and sometimes even a model of the torso is included. Common available dummy heads are from manufacturers like Neumann, Head Acoustics GmbH, Brüel & Kjær. In this experiment, recordings were done using the KU100 dummy head by Neumann.

Dummy head recordings and reproduction has several applications, which include but not limited to comparison of concert halls, assessment of speech in rooms, impact of room acoustics on loudspeaker sound. Keeping these in mind, this paper tries to use Binaural recordings and qualitatively assess the recordings to make predictions and comparisons on the results from the subjective evaluation of the study using loudspeakers.

5.3 Methodology

5.3.1 Apparatus and Stimuli

The experiment took place in the audio research lab at the Hochschule für Gestaltung (HfG), Karlsruhe. The lab has 29 loudspeakers arranged in a hemispherical configuration. Measured from the sweet spot at the center of the room, the loudspeakers are located at a distance of 2.4m. 16 of the 29 loudspeakers are placed in the horizontal plane, 8 of them at an approximate elevation of 30°, 4 at approximately 60° elevation and 1 directly above the listener (0°, 90°). Each speaker was calibrated to 79dBc (broadcast standard) using -20dBFS pink noise as described in [101]. Figure 1 shows the loudspeaker subject study in progress. The experiment procedure is described in detail in [95], the first part of this paper.

Stimuli

Four different music genres were selected for the purpose of this study. 40-50s excerpts were selected from each song. All four stimuli were mixed and mastered in 5.1 by the same engineer. This prevented the tracks from having any mixing biases different engineers might have. The stimuli were: 1. A 42s excerpt of Dvořák: Symphony No. 9 in E Minor, "From the New World": IV Allegro Con Fuoco. 2. A 41s Electronic excerpt by an unknown composer. 3. 49s excerpt of the soft rock song Hey 19 by Steely Dan and 4. A 39s long excerpt of an Afrobeat song Aramile by Babatunde Olatunji.

5.3.2 Procedure

Binaural recordings were done using the Neumann KU 100 dummy head. The dummy head was mounted on to a mic stand and the height of was adjusted such that the loudspeakers on the horizontal plane were at ear level to the dummy head, similar to the subject experiment. The dummy head was placed in the sweet spot in the center of the room in order to capture the recording as close to the subjective study as possible. Following this, levels were adjusted on both the binaural mic inputs by playing a constant signal from the front loudspeaker only. This ensured that both the left and right channel inputs were calibrated correctly. The binaural recording process is depicted in figure 2.

All the stimuli were then played back using the same user interface as the subjective study. Mic inputs were recorded on a laptop using Audacity software and the files were stored as Audacity sessions and in uncompressed mono files (one for each channel) in the wav format.

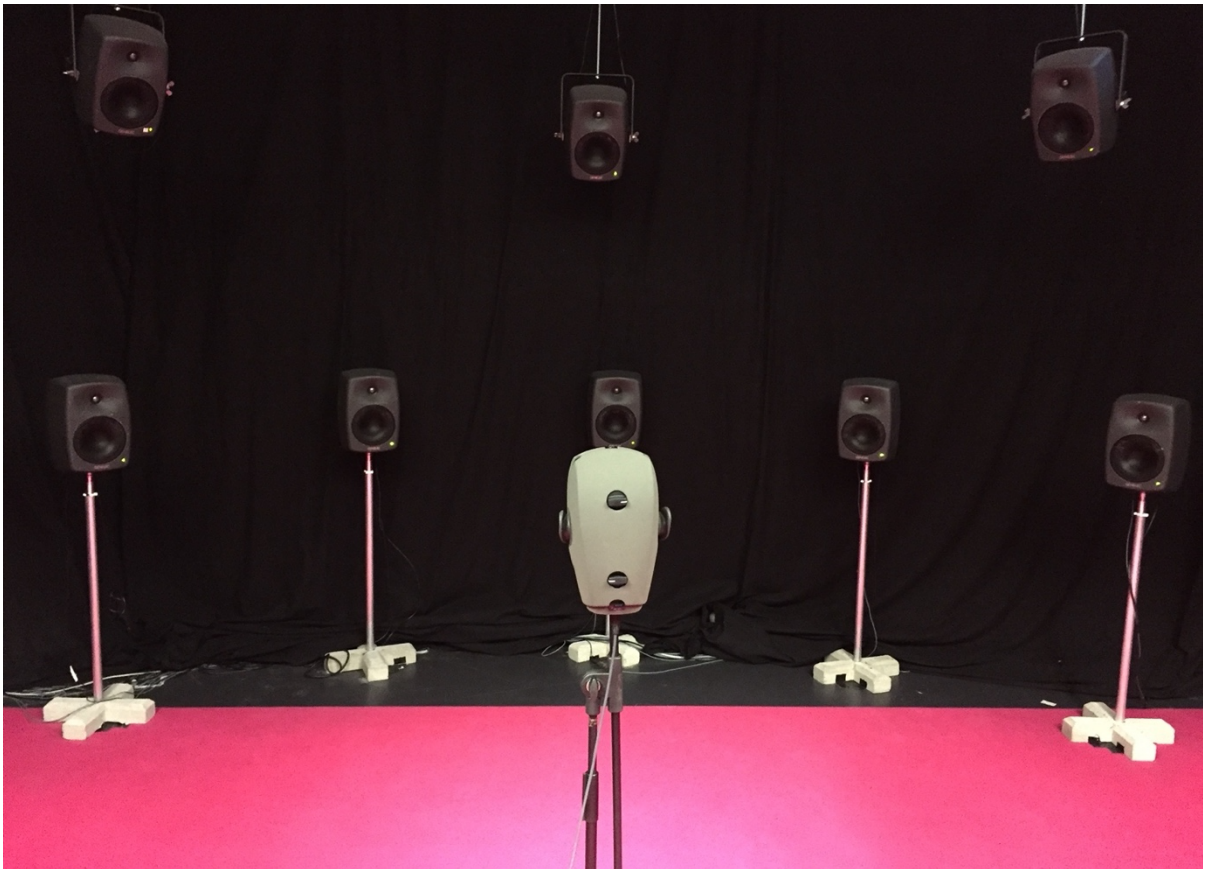

Figure 26: Binaural recording in process

5.4 Results and Discussion

5.4.1 Results

Preliminary analysis of the experiment can be found in [95] (Chapter 4). To summarize, 11 participants (6 male and 5 female) with varying experience in the field of spatial and 3D audio took part in the experiment. Analysis of the A/B comparison experiment was done by normalizing the ratings (0 to 1) and the scales were used as the frequency with which each participant selected the corresponding algorithm as typically done in a scaling experiment of A/B comparison data using these scale values, a one-way repeated measures Analysis of Variance (ANOVA) was performed independently for the different attributes and the pieces used in the experiment. Further analysis is described below.

Binaural Recording Analysis

Analysis of the Binaural recordings were done using Essentia [23], an open-source library for audio analysis and audio-based music information retrieval. Many spectral features from the audio were extracted and studied. Comparisons were made across each piece and the three reproduction techniques to see at the listener, how the spectral features interacted across genres and if that could potentially have impacted the ratings seen in [17] and in the ANOVA analysis above. The results are summarized in Table 2 and plots drawn from the recordings are shown in figures 3-10.

Table 4: Summary of the features extracted for the four pieces and the corresponding reproduction technique in the experiment.

Rock			
Feature	**Ambisonics**	**VBAP**	**DBAP**
Spectral Centroid (Hz)	581.81	590.58	491.25
Spectral Spread	5.53	41.3	13.22
Spectral Kurtosis	6.18	6.1	6.1
Dynamic Complexity (dB)	4.37	4.13	5.22
Classical			
Feature	**Ambisonics**	**VBAP**	**DBAP**
Spectral Centroid (Hz)	594.72	571.64	464.06

Spectral Spread	0.13	21.33	29.84
Spectral Kurtosis	14.56	6.15	6.04
Dynamic Complexity (dB)	6.38	5.77	6.9
Electronic			
Feature	**Ambisonics**	**VBAP**	**DBAP**
Spectral Centroid (Hz)	236.95	238.38	291.3
Spectral Spread	0.21	0.2	0.09
Spectral Kurtosis	1.18	1.04	6.05
Dynamic Complexity (dB)	5.08	4.88	6.81
Afro Percussion			
Feature	**Ambisonics**	**VBAP**	**DBAP**
Spectral Centroid (Hz)	299	271.61	319.65
Spectral Spread	0.14	0.02	0.14
Spectral Kurtosis	13	0.32	1.44
Dynamic Complexity (dB)	4.11	4.29	5.63

Spectral Centroid of the signal, which describes the brightness of a signal, is computed by applying a first difference filter to the signal then dividing the norm of the resulting signal by the norm of the input audio signal. The result is given in hertz and characterizes the brightness of a signal.

Spectral Spread and Kurtosis both describe the distribution shape of the signal. The algorithms used in Essentia for this are described in [103]. Spectral spread a measure of the variance of the signal while kurtosis is a measure of the "tailedness" of a distribution as is defined as the ratio of the 4th central moment of the signal and its standard deviation to the power of 4.

Dynamic complexity expressed in dB is the average absolute deviation of the signal from its global loudness and it can be related to the dynamic range and fluctuation of loudness in the music signal. This algorithm was written as described in [104].

Spectral complexity was computed in a framewise manner based on the number of spectral peaks in the given spectrum of the frame, while spectral flatness is the ratio of the geometric mean and the arithmetic mean of the input. Further descriptions of these can be found in [103] and [105]. The plots of the spectral complexity and flatness are shown for each piece individually with all the reproduction techniques plotted against one another in the same plot as seen in figures 3-10. These results along with the results shown in table 2 are further discussed in the discussion section.

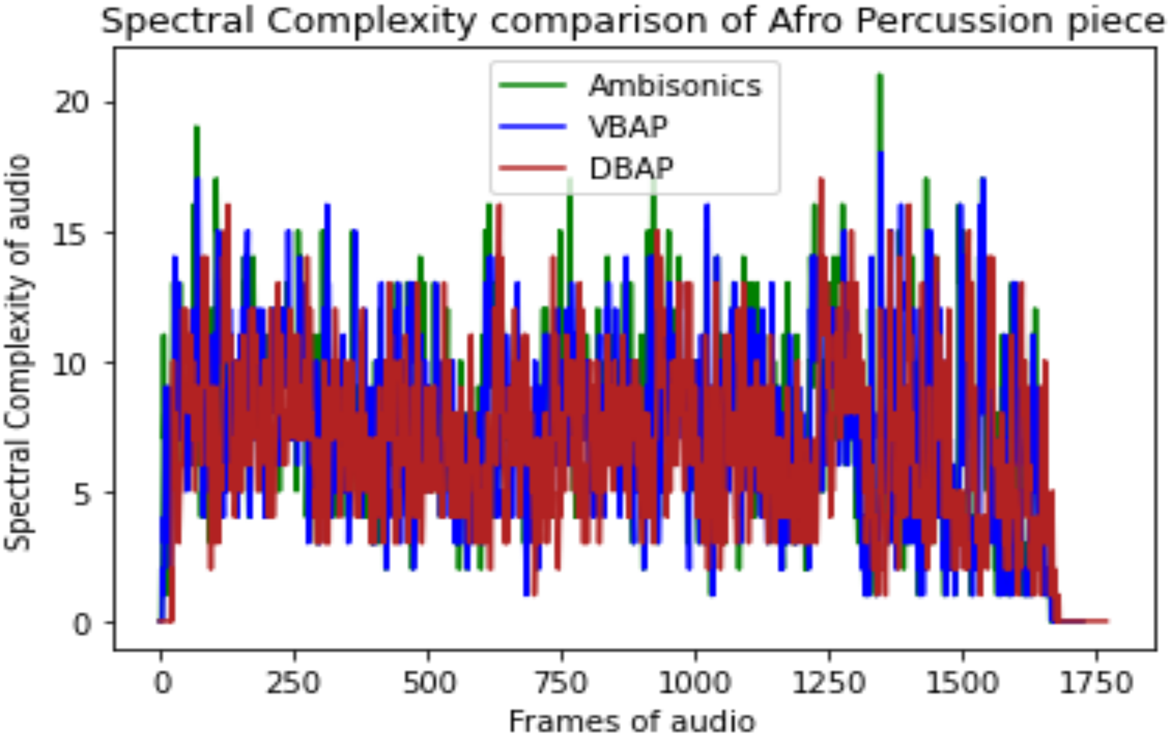

Figure 27: Framewise Spectral Complexity comparison of the Afro Percussion piece across the three reproduction techniques

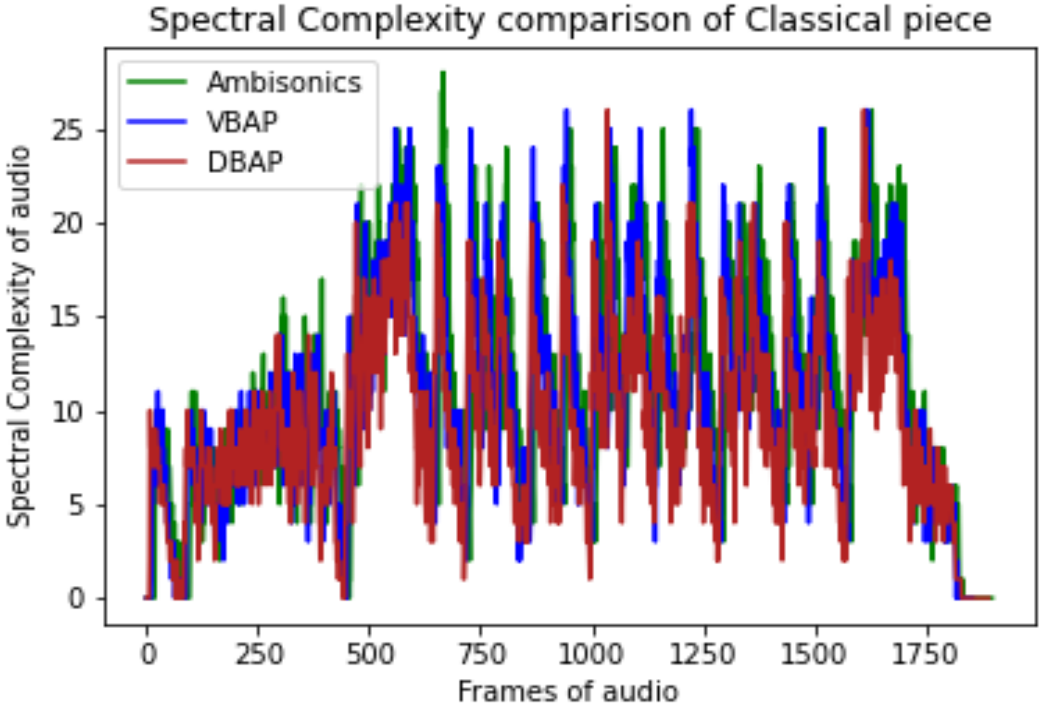

Figure 28: Framewise Spectral Complexity comparison of the Classical piece across the three reproduction techniques

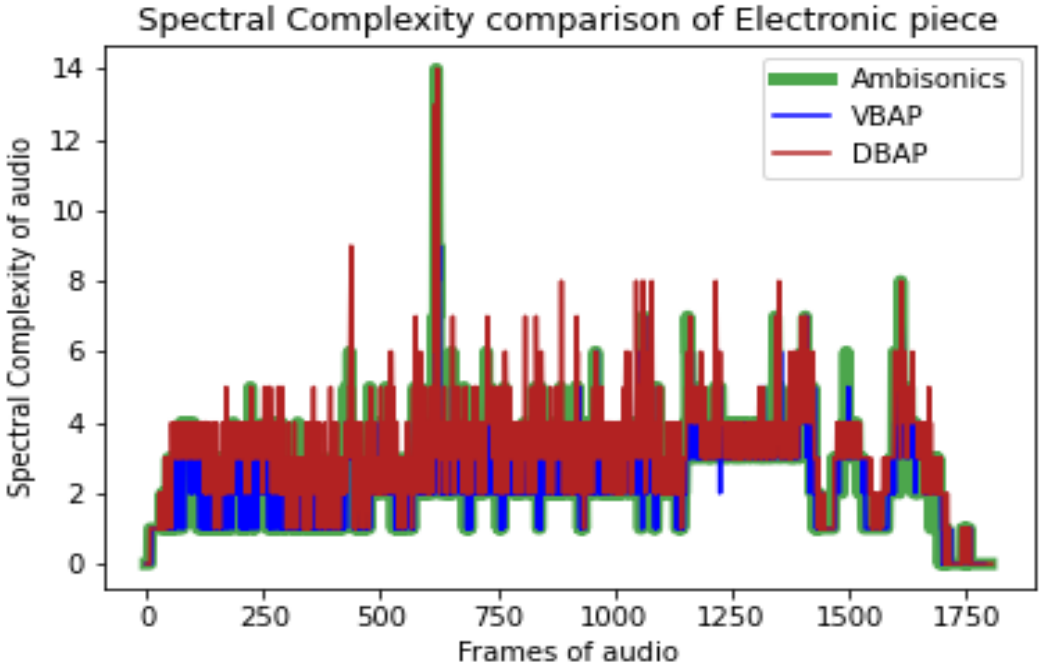

Figure 29: Framewise Spectral Complexity comparison of the Electronic piece across the three reproduction techniques

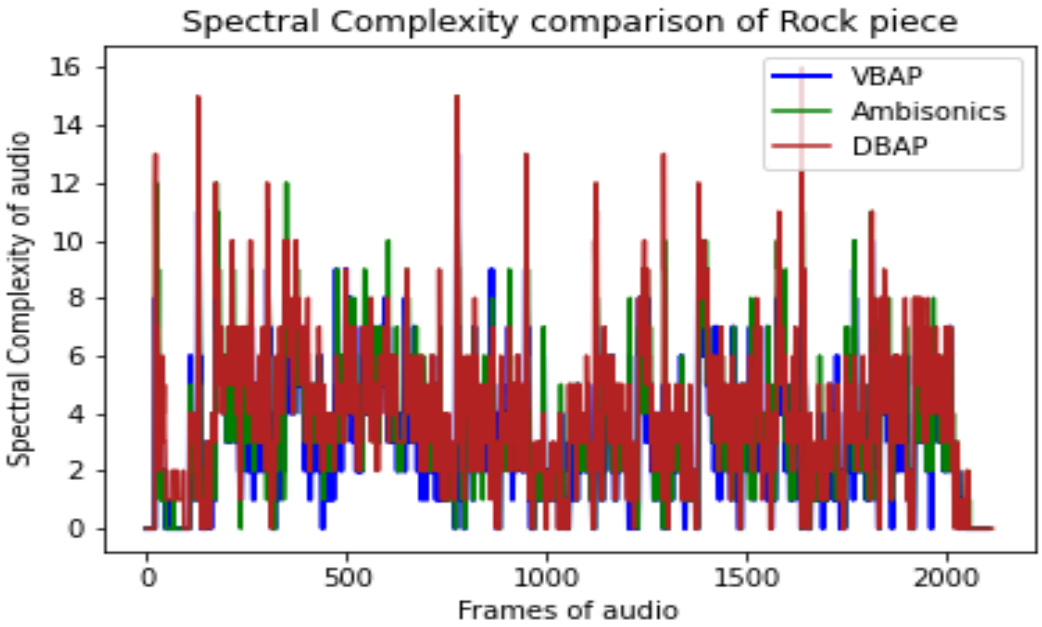

Figure 30: Framewise Spectral Complexity comparison of the Rock piece across the three reproduction techniques

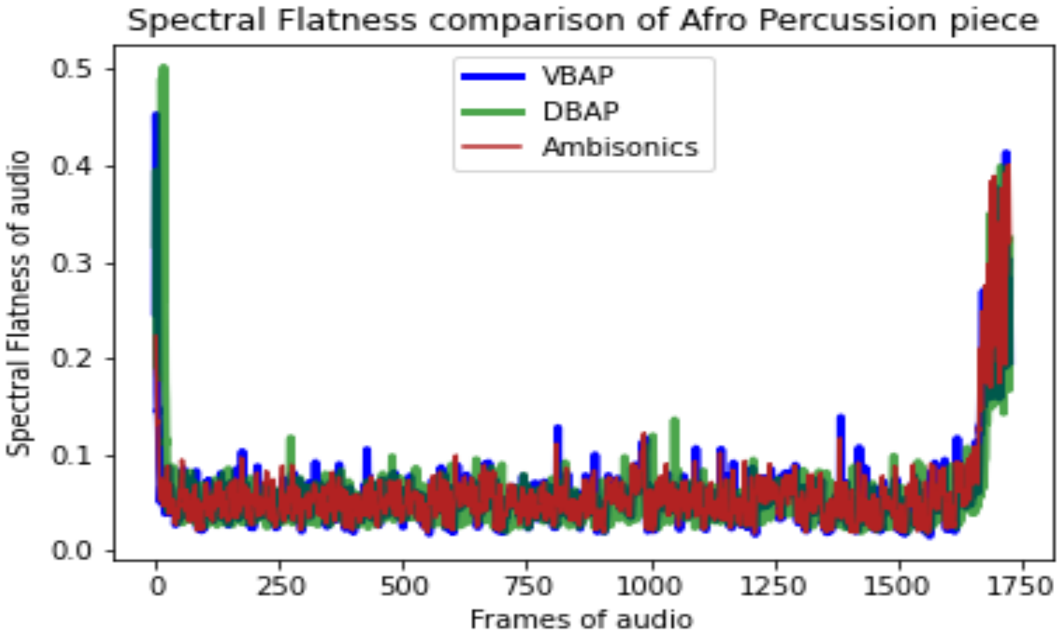

Figure 31: Framewise Spectral Flatness comparison of the Afro Percussion piece across the three reproduction techniques

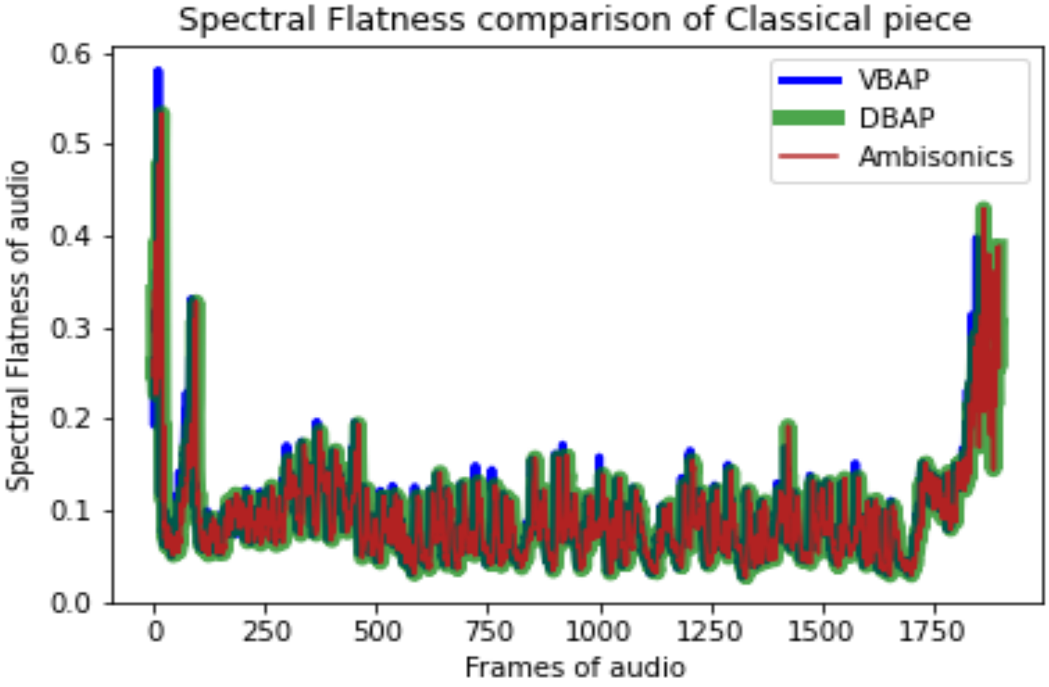

Figure 32: Framewise Spectral Flatness comparison of the Classical piece across the three reproduction techniques

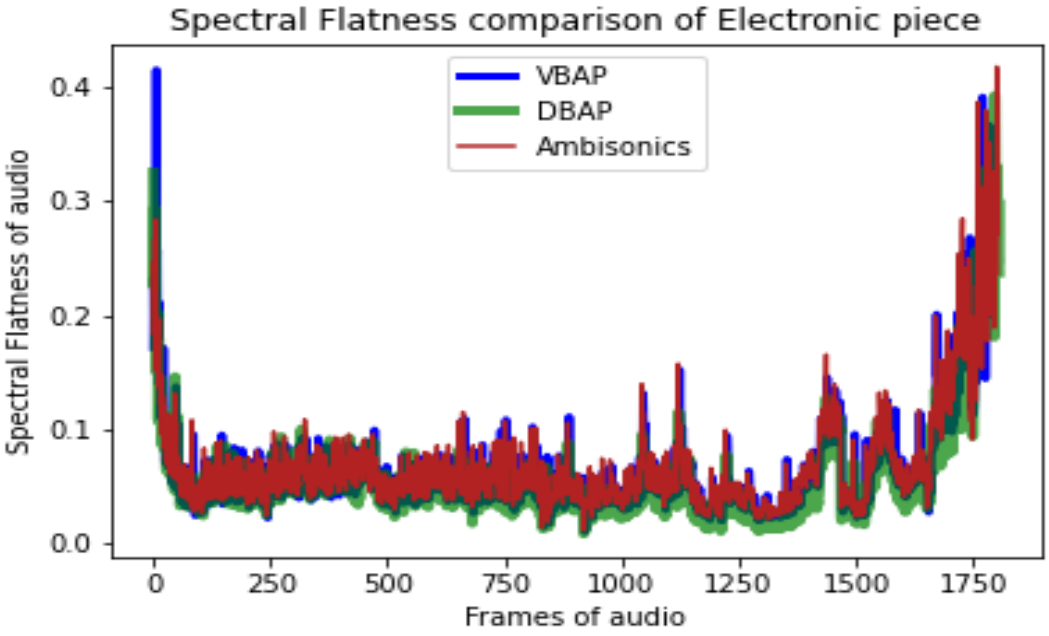

Figure 33: Framewise Spectral Flatness comparison of the Electronic piece across the three reproduction techniques

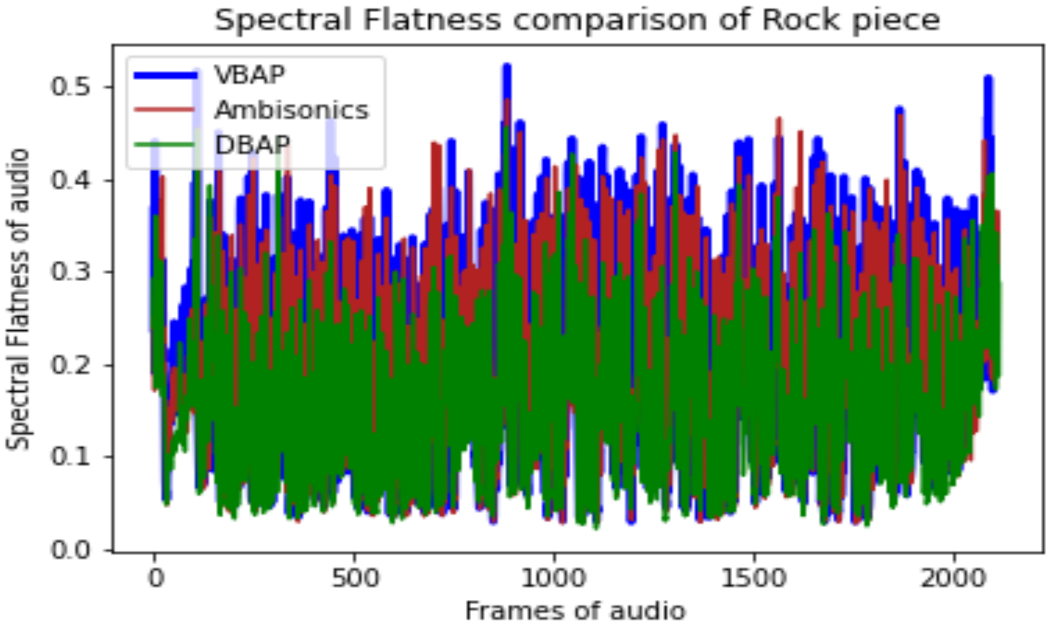

Figure 34: Framewise Spectral Flatness comparison of the Rock piece across the three reproduction techniques

5.4.2 Discussion

While the use of non-musical stimuli to conduct comparative studies has its merit, this study along with its counterpart (chapter 4) have shown that the comparisons using mainstream musical material can also be extremely useful and can provide interesting results.

The addition of the analysis of the Binaural results added an interesting layer to the results obtained from the loudspeaker study. The VBAP signal being the brightest with greatest spectral spread on more than one occasion could well be assumed as the primary factor for the selection of VBAP as the most preferred reproduction type. This could also be backed by the fact that on most occasions, the VBAP spectral complexity curve was the lowest. There was a clear correlation between brightness and spectral spread for other attributes as well. [105] tells us that a lower spectral complexity signal also contributes to a more relaxing piece. This is of course an intuitive result. Therefore, a more “relaxing” piece being rated as more preferred is also an intuitive conclusion. This was also seen in the rating of the attributes. Lastly, VBAP is the highest curve for most of the pieces, furthering the inferences drawn in chapter 4.

5.5 Summary

In conclusion, a comparative study of Ambisonics, VBAP and DBAP was performed using musical material of 4 different genres. It was observed how each of the techniques influence attributes such as Preference, Envelopment, Spatial Clarity, Sound Quality and Stability. In chapter 4, it was found how Preference of a technique could be described using the ratings of the other attributes. The possible reasons behind the results obtained in chapter 4 formed the key part of the analysis in this chapter. The analysis of the Binaural recordings done. It was seen how more bright, next spectrally complex and more spectrally flat signals could have resulted in the user ratings. This provides interesting insights into the reproduction techniques and how they interact with different musical genres.

6. The Virtual Concert Hall Prototype

This chapter describes the creation of the virtual concert hall prototype. This is a virtual environment created that uses the learnings from previous chapters to setup an environment that can be utilized by composers, electronic music performers, educators, and music enthusiasts.

6.1 Introduction

The creation of virtual environments has rapidly increased in the last decade due to the VR boom. At present there are several VR technologies, tools, applications that perform a multitude of tasks. As we have seen from the case studies section in Chapter 2, even music/audio related applications have a whole range of different creations. However, as I researched further into the topic in order, I realized that there isn't a tool that could give various options to the user to utilize for composition, educational or recreational purposes. That was the primary motivation behind the design and implementation of the virtual concert hall prototype.

6.2 Design of the Prototype

Having a solid motivation behind building a tool is only step one of the entire process. There were still many questions that were open. What software am I going to use? How would the VR experience be? Should I make it screen-based, or headset based? What headset to use if I choose to go that route? All these questions needed to be answered even before implementation began.

Ultimately, I chose to build a headset-based virtual environment. I felt that the kind of tool I was developing would find much more use in a headset-based VR experience rather than a screen-based model.

Most design decisions were taken drawing inspiration from the various case studies described in Chapter 2. For example, the Allosphere has multiple spatialization reproduction methods that creators can choose from. In the current prototype as well, I wanted to keep similar options. This particular decision was also driven by the comparative studies done detailed in Chapters 4 and 5.

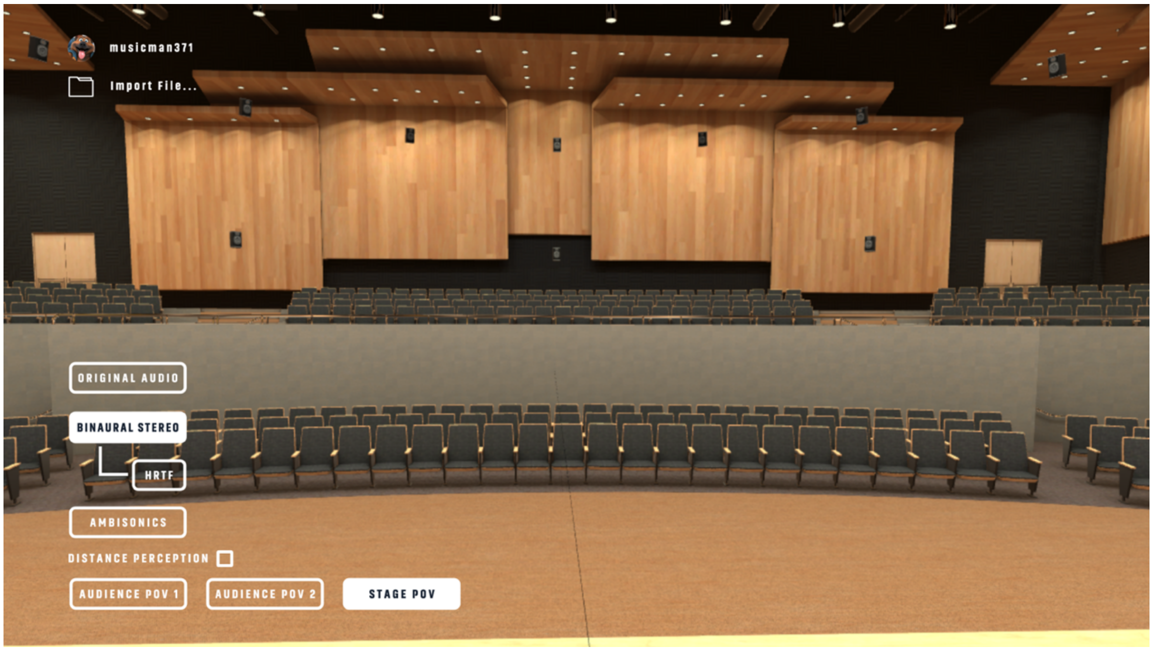

Figure 35: The view of the virtual concert hall from the stage

Above is a screenshot of the virtual concert hall built. It also shows all the various options a user has once they begin to use the tool. Before, I describe the various features and functionalities of the environment, let us first look at the decisions taken into building the hall itself.

For starters, I chose Unity as the development platform for this project. This project is done in collaboration with a researcher at Auckland University of Technology and together the two of us were most comfortable with this work environment and we felt it had everything we need to build the concert hall. It made the most sense to develop in this rather than spend time learning a new tool.

The next decision was designing the concert hall. Here, we could go two routes. Either build the whole thing from scratch or choose a pre-made asset available in the Unity asset store. We found an asset that suited our needs perfectly and instead of reinventing the wheel we chose to use this pre-made asset with slight modifications. The reasons this asset works really well for us are:

- It has two levels in the audience. This was ideal for us as one of the things we were thinking of was giving multiple perspectives a user can switch between.
- It was low polygon. So, rendering and immersion is a lot easier and the threshold for users to access this as a VR experience is much lower and computationally friendly.

The next important thing to get correctly was the lighting. This is one of the modifications done to the pre-made asset. The lighting was adjusted to give a realistic feel, one that was more representative of an actual concert hall with the house lights on. This ensures that when someone is working with the tool, they don't get the impression of a weirdly constructed environment. It needed to feel natural.

The next decision involved the VR elements within the concert hall. What it would look like when they're standing vs when they're sitting in their seat. We made the decision to not allow movement. The user could only select one of the three points of views (POV) for the experience. The reasoning behind this decision was that the hall was built to explore audio rather than explore the physical space, so we needn't give the user the ability to walk around the room. This meant that the cameras were locked, but we had to ensure we get the height correct.

Having made these initial decisions, it was time to decide on how the user interface (UI) would look. Here we selected a screen canvas and simple buttons that the user can point to and pull the trigger using the controller. The screen canvas ensures that the UI is locked to the user's head and upon rotation, the UI simply follows the person. We experimented with dropdown menus to reduce the size of the UI, but the interaction was much more difficult with dropdown menus. The experience of buttons was much cleaner and simpler.

Now, let us look at the features implemented. As alluded to earlier, there are three positions that the user can switch between. They are Stage POV, Audience POV 1 and Audience POV 2 (one for each level). This functionality gives a composer/performer multiple perspectives of the piece they are composing/performing. As a consumer also, the user gets multiple perspectives to enjoy.

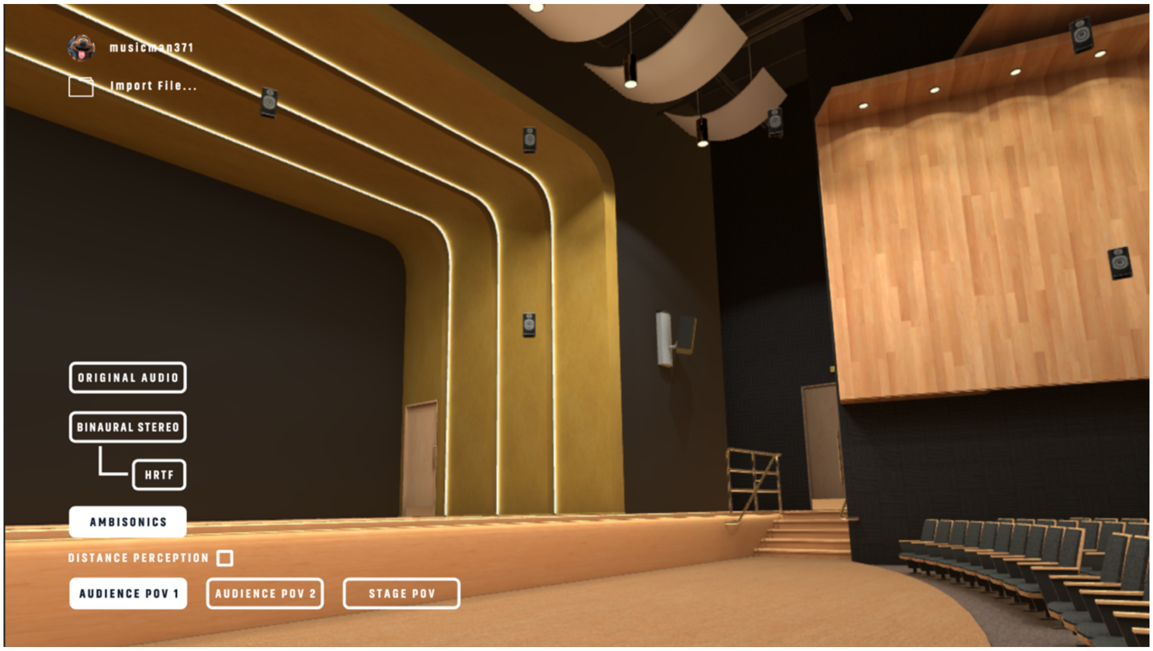

Figure 36: Lower-level audience point of view

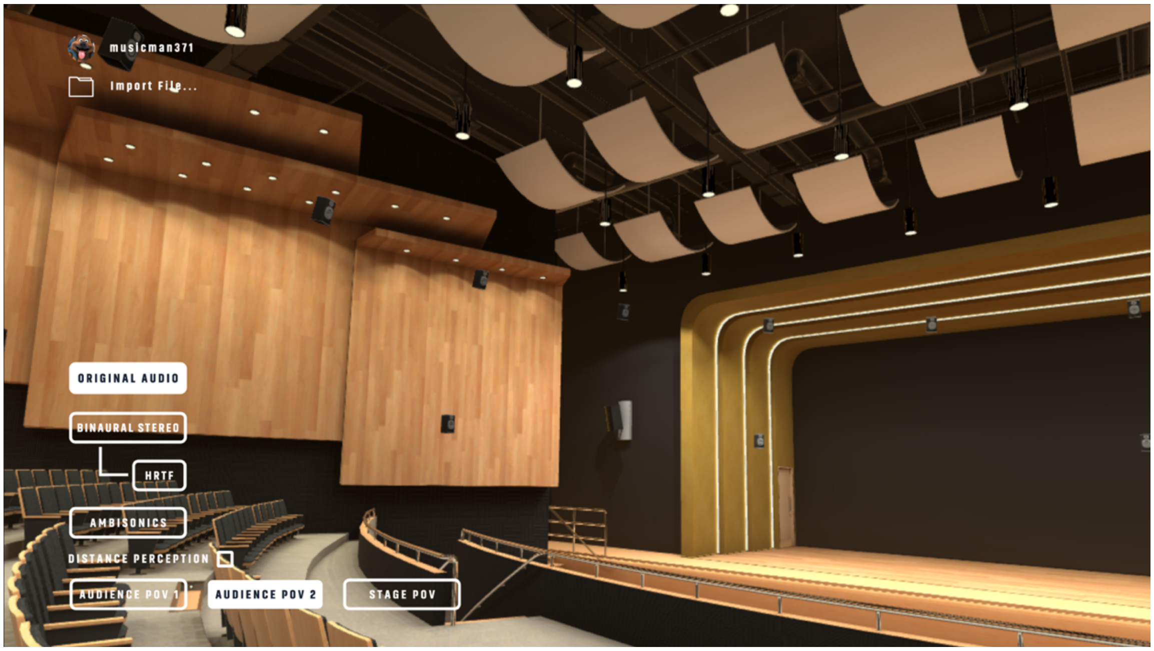

Figure 37: Upper-level audience point of view

There are three modes the audio can be listened in. One is the original audio that is imported. The second one is Binaural stereo. Within Binaural stereo, there is an HRTF option. This button is currently not functional, but it is already implemented with the future work in mind. It is to give the user the ability to import a personal HRTF or choose from a database of HRTFs to find the "best-fit" that would enhance their Binaural experience. The last option is Ambisonics. Here, expectation as of now is that the user has an Ambisonic encoded audio file. That is then decoded and listened to in the virtual environment.

The last button is the Distance perception. This is a toggle, based on the results shown in chapter 3. On the top left of the photo, you can see that upon startup, there is a profile that is loaded. For a new profile, the distance perception is off and when the tool is used more than once, the button is set to true. When it is set, the direct to reverberation ratio applied to the audio is adjusted to factor in the change of distance perception. As of now, this mimics the results from Chapter 3, however, as stated in the future work of that chapter, localization studies within the VR environment need to be done before accurate calibration can be done to factor this in perfectly. Currently, this is just a proof of concept. The reason this button is present in the UI is to give the composers an idea of how the music would sound before and after the user spends time listening to their piece in the environment.

6.3 Summary and Future Work

This chapter describes the design and implementation of a virtual concert hall prototype. This prototype currently is a proof of concept combining elements from previous chapters to build a tool that could be used by music composers, electronic music performers, educators and consumers. There are three points of view that the user can switch between to get different listening perspectives. This would help music composers and electronic music performers perfect their pieces while it's a nice feature for consumers to utilize and get a broader listening experience. The tool also gives multiple spatializing options and an option to listen to the original audio along without spatialization along with a distance perception toggle switch. Both these features are primarily designed with composers and educators in mind. For composers, this would give multiple perspectives on their piece and help them make informed decisions while creating their compositions. Educators on the other hand, can toggle between different spatializations and give a hands-on demo/experience for their students as they teach concepts of spatial audio.

This prototype has tremendous potential and a lot of room for expansion and future work. As a first step, I would be conducting two user studies. One with a combination of composers, performers, educators and consumers to get a sense of what each of them feels about the tool. What works, what doesn't work, what would be good to have etc. This will help refine this tool before it can be released for everyone to use. The second user study would be similar to the sound source distance study described in Chapter 3. So, the future work of that and this would be combined to help refine the distance perception button.

Following that, I also have plans to expand this tool to have more virtual concert halls or listening rooms to give many more perspectives for composers, educators, and consumers to use and play with.

7. Conclusions and Future Work

This chapter summarizes the main contributions as part of this dissertation. We shall revisit the research question from chapter 1 and see how the work done fits into it. We will also look at the broader impacts of this dissertation and the significance of this research. Lastly, future directions of this research in the field, that follows the work presented is discussed.

7.1 The Research Question Revisited

This dissertation attempts to answer the questions, **"What psychoacoustic factors affect listening to music in non-virtual listening contexts and how could they transform when listening in a virtual listening context? How do these psychoacoustic factors work?"**

As discussed in the introduction, these questions have many moving parts to it and a number of phrases with fluid definitions. Despite setting the scope of the problem, it can be difficult to find an approach that can give a robust solution to the problem. This is because music spatialization and the use of space in music in itself is a topic that has a myriad of viewpoints and music composers do not have a unified view on the matter. On the one hand, composer Dennis Smalley says, "Space is the whole thing. It is not usually something that people perceive as separate from the sounds themselves, although the composer might consider space separately – might blot out certain aspects of the sounds to consider purely spatial factors. For the listener, they're all molded into one. That's why we end up talking about the piece as a whole, because the whole is the space or spaces of the piece" [106] while on the other hand, Johannes Goebel argues that "In my opinion the spatial placement of sounds, whether instrumental or electronically, has about the same potential for aesthetic differentiation as loudness. Compared to pitch and timbre, localization yields far less potential for aesthetic differentiation, but on the other hand no one would deny that in quite a few pieces loudness is an important, highly sophisticated, composed part of music. And the same could be said for the distribution of sound in space." [107]. Music spatialization also includes various factors other than localization. There are factors such as distance, spatial clarity and stability that also need to be considered. With these large opinion differences in mind, the thesis attempted to answer the aforementioned research questions by studying some of the key aspects that were not yet explored in the field. The research centered around auditory perception as ultimately, in my opinion, the success of a music spatialization strategy is dependent on the perception of the listener.

In chapter 3, a study on the estimation of sound source distance was designed and carried out. The topic of sound source distance has been researched before, but this chapter introduces a novel approach to the problem. Building off prior studies on the estimation of sound source improving with familiar sound sources (e.g., a friend's voice), a hypothesis is made on the estimation of sound source improving with familiarity of the environment. A subjective experiment was designed to test this hypothesis and the results showed a 23% improvement in the estimation of sound source in a known environment.

Chapters 4 and 5 looked at the comparison of spatialization techniques with different music genres. Comparative studies of spatialization techniques primarily use non-music material. The reason being using musical material as stimuli increases the variables multiple fold and unwanted elements could affect the outcome of the comparisons. The studies in these chapters though, carefully controlled for format, mixing and mastering strategies etc. to design a unique two-part comparative study of spatialization techniques with music material. The loudspeaker study described in chapter 4 was a subjective analysis on different attributes to see which spatialization technique performed well for which attribute and which genre. Chapter 5 had a novel approach of analyzing the binaural recordings of the material presented as part of the subjective study in Chapter 4. This provided a new perspective on the various spectral, melodic, and rhythmic features present in the stimuli that could have led to the results seen in Chapter 4.

Finally, chapter 6 describes the creation of a virtual concert hall. This concert hall is an informed prototype that incorporates learnings from previous chapters. It is also a prototype that can aid electronic musicians during their composition and rehearsal processes while also being a tool that can be used by educators teaching topics such as 3D audio, virtual reality, audio spatialization and so on. Such a prototype is currently non-existent either in academia or industry.

7.2 Broader Impacts

The scientific and artistic exploration of sound has been present for over a century and yet I believe that today we stand at a time when we could spend another 100 years on this topic and still feel like we can spend hundreds of more years on the topic. New realities spring up each day and they open new paths to explore. As a researcher, musician, and educator at heart, with this dissertation, I tried to impact the field of 3D audio with each of those three in mind. Therefore, we can see that each chapter impacts the field in a slightly different way.

In chapter 3, we saw a novel study on the problem of sound source distance. With the boom in virtual reality, we increasingly see virtual spaces that are being created every single day. These may be fantasy spaces or actual spaces that are being realized in virtual reality to give a different perspective and experience for the consumer. Either way, creation of auditory displays, audio-visual installations and other audio content in these new environments would benefit from the understanding of how humans perceive sound source distance and how that perception changes over time as familiarity increases.

Comparative studies on music spatialization reproduction techniques in chapters 4 and 5 provide fresh perspectives that composers can make the most of while composing new music material. Lastly, the development of a unique virtual environment combining concepts learnt from previous chapters has uses for a musician, educator and a consumer thus rounding up my attempt to impact the field from a multitude of angles.

7.3 Research Timeline

This dissertation began in September 2016 and over the course of these six years has resulted in three publications (each of chapters 3, 4 and 5) and two more papers (chapters 2 and 6) are nearing completion. These are listed below.

Published articles:

- **Aswathanarayana, S**. (2021, May). Comparison of Spatialization Techniques with Different Music Genres II. In Audio Engineering Society Convention 150. Audio Engineering Society.
- **Aswathanarayana, S**. (2020, October). Comparison of Spatialization Techniques with Different Music Genres. In Audio Engineering Society Convention 149. Audio Engineering Society.
- **Aswathanarayana, S**. (2017, May). Effect of a Known Environment on the Estimation of Sound Source Distance. In Audio Engineering Society Convention 142. Audio Engineering Society.

Articles yet to be published:

- Development of music performance in virtual reality – a review paper synthesizing material from chapter 2. Planned to be submitted to the Journal of the Audio Engineering Society or Organized Sound or Journal of New Music Research or IEEE Signal Processing Planned publication: Summer 2023.
- Sangeeta Kalpanica Bhavan: A virtual music space for composers, performers, and educators. This paper will primarily detail work described in chapter 6 of the dissertation. The plan is to submit it to the Audio for Virtual and Augmented Reality (AVAR) conference in 2023.

7.4 Future Directions

The work presented in this dissertation proposed novel approaches in the research of sound source distance and comparison of spatialization techniques. These are by no means exhaustive and each of the topics has scope of expansions into more dissertations.

To begin with sound source distance, as stated in the broader impacts section, there is a current boom in virtual reality and increased creation of new virtual spaces. However, auditory perception studies on sound source distance have currently only taken place in non-virtual listening contexts and such studies are yet to take place in virtual listening contexts. Taking such source distance studies into virtual listening contexts will be very interesting and help us solidify our understanding of externalization of sound in headphones to create greater immersion for the listener.

Comparative studies described in chapters 4 and 5 used a limited number of genres and music material that can easily be extended to develop a much more comprehensive understanding of music spatialization with many more genres of music that are vastly different from the material used in this dissertation. This can definitely open up many creative ideas for composers.

The prototype developed in chapter 6 has many limitations and it’s only a proof of concept at this stage as described in the chapter. Building on this prototype itself is a big project that can be undertaken in the future to provide a tool that would have all the capabilities as described in the future work section of chapter 6.

On a personal level, I have recently become very interested in a project that is different from all that is described in this dissertation and yet the learnings of this dissertation can be applicable in my project. I am hoping to study the acoustics of South Indian Hindu temples. Hindu religious worship is vastly different from Christian church worship, so traditional techniques for characterization of worship spaces cannot be applied to Hindu temples. I plan to build on top of current geometric acoustic techniques to characterize Hindu temples and in the process devise new characterization methods that are more apt for Hindu worship spaces. Having devised new methods to characterize Hindu temples, I plan to build computer models of the same and bring the temples to life in the virtual world similar to the virtualization of the Notre-Dame Cathedral described in [60]. To accomplish this part of the project, and create novel listening experiences, learnings from this dissertation will be key.

In summary, the vastness of the field and the many possible future directions that the work presented in this dissertation can take leaves me very excited for a Soundtastic (made up word – a mash up of sound and fantastic) future ahead.

References

[1] Roads, C. (2015). Composing electronic music: a new aesthetic. Oxford University Press, USA.

[2] Chadabe J. (1997). Electric Sound: The past and promise of electronic music. Prentice Hall.

[3] https://www.dolby.com/technologies/dolby-atmos/

[4] Lee, K., Son, C., & Kim, D. (2010, May). Immersive virtual sound for beyond 5.1 channel audio. In Audio Engineering Society Convention 128. Audio Engineering Society.

[5] Hamilton, R. (2019, March). Collaborative and competitive futures for virtual reality music and sound. In 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR) (pp. 1510-1512). IEEE.

[6] Çamcı, A., & Hamilton, R. (2020). Audio-first VR: New perspectives on musical experiences in virtual environments. Journal of New Music Research, 49(1), 1-7.

[7] Lanier, J., Heilbrun, A., & Sacks, B. (1989). An interview with Jaron Lanier. *Whole Earth Review* (Fall Ed.), 109–119.

[8] Palumbo, M., Zonta, A., & Wakefield, G. (2020). Modular reality: Analogues of patching in immersive space. Journal of New Music Research, 49(1), 8-23.

[9] Heilig, M. L. (1960). *U.S. Patent No. 2,955,156*. Washington, DC: U.S. Patent and Trademark Office.

[10] Sutherland, I. E. (1965). The ultimate display. In *Proceedings of the IFIP Congress* (pp. 506–508).

[11] Milgram, P., Takemura, H., Utsumi, A., & Kishino, F. (1995, December). Augmented reality: A class of displays on the reality-virtuality continuum. In Telemanipulator and telepresence technologies (Vol. 2351, pp. 282-292). International Society for Optics and Photonics.

[12] Graham, R., & Bridges, B. (2016, June). Competing Attractions, Orbital Decay and the Music of the Spheres: Force–based relational dynamics for organizing space and timbre in performance using physical modelling. In Korean Electro-Acoustic Music Society Annual Conference. Korean Electro-Acoustic Music Society.

[13] Smalley, D. (1997). Spectromorphology: explaining sound-shapes. *Organised sound*, *2*(2), 107-126.

[14] Smalley, D. (2007). Space-form and the acousmatic image. *Organised sound*, *12*(1), 35-58.

[15] Serafin, S., Geronazzo, M., Erkut, C., Nilsson, N. C., & Nordahl, R. (2018). Sonic interactions in virtual reality: state of the art, current challenges, and future directions. *IEEE computer graphics and applications*, *38*(2), 31-43.

[16] Harley, M. A. (1997). An American in Space: Henry Brant's" Spatial Music". American Music, 70-92.

[17] Zelli, B. (2009, October). Spatialization as a musical concept. In 2009 IEEE International Symposium on Mixed and Augmented Reality-Arts, Media and Humanities (pp. 35-38). IEEE.

[18] Hamilton, R., Caceres, J. P., Nanou, C., & Platz, C. (2011). Multi-modal musical environments for mixed-reality performance. Journal on Multimodal User Interfaces, 4(3), 147-156.

[19] Begault, D. R. (2000). 3-D Sound for Virtual Reality and Multimedia. Moffett Field, CA: National Aeronautics and Space Administration.

[20] Blauert, J. (1997). *Spatial hearing: the psychophysics of human sound localization*. MIT press.

[21] Amet, E. H. (1911). Method of and Means for Localizing Sound Reproduction. US Patent 1,124,580.

[22] Blumlein, A. D. (1931). Improvements in and Relating to Sound-transmission, Sound-recording, and Sound Reproducing Systems. Great Britain Patent, 394,325.

[23] Torick, E. (1998). Highlights in the history of multichannel sound. Journal of the Audio Engineering Society, 372, 368-272.

[24] Malham, D. G., & Myatt, A. (1995). 3-D sound spatialization using ambisonic techniques. *Computer music journal*, *19*(4), 58-70.

[25] Teruggi, D. (2007). Technology and musique concrète: the technical developments of the Groupe de Recherches Musicales and their implication in musical composition. *Organised Sound*, *12*(3), 213-231.

[26] Stockhausen, K. (1992). Programme Notes for the 1956 World Premiere of Gesang Der Jiinglinge. *liner notes for the*, 1952-1960.

[27] https://www.wired.com/2009/05/dayintech-0512/

[28] Davis, M. F. (2003). History of spatial coding. *Journal of the Audio Engineering Society*, *51*(6), 554-569.

[29] Gerzon, M. A. (1973). Periphony: With-height sound reproduction. *Journal of the audio engineering society*, *21*(1), 2-10.

[30] Gerzon, M. A. (1975, March). The design of precisely coincident microphone arrays for stereo and surround sound. In *Audio Engineering Society Convention 50*. Audio Engineering Society.

[31] O'Donovan, A., Duraiswami, R., & Gumerov, N. A. (2007, October). Real time capture of audio images and their use with video. In *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (pp. 10-13). IEEE.

[32] Snow, W. (1955). Basic principles of stereophonic sound. *IRE Transactions on Audio*, (2), 42-53.

[33] Berkhout, A. J. (1988). A holographic approach to acoustic control. *Journal of the audio engineering society*, *36*(12), 977-995.

[34] Berkhout, A. J., de Vries, D., & Vogel, P. (1993). Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America*, *93*(5), 2764-2778.

[35] Spors, S., Rabenstein, R., & Ahrens, J. (2008). The theory of wave field synthesis revisited. *In 124th Convention of the AES*.

[36] Pulkki, V. (1997). Virtual sound source positioning using vector base amplitude panning. *Journal of the audio engineering society*, *45*(6), 456-466.

[37] Lossius, T., Baltazar, P., & de la Hogue, T. (2009, August). DBAP–distance-based amplitude panning. In *ICMC*.

[38] Kostadinov, D., & Reiss, J. D. (2009). Spatial Audio Matlab Toolbox. *Centre for Digital Music, Queen Mary University of London, London August*, *14*.

[39] Roginska, A., & Geluso, P. (2017). Immersive Sound. Focal Press.

[40] Wade, N. J., & Deutsch, D. (2008). Binaural hearing–Before and after the stethophone. Acoustics Today, 4(3), 16-27.

[41] Collins, P. (2008). Theatrophone: the 19th-century iPod. *New Scientist*, *197*(2638), 44-45.

[42] Paul, S. (2009). Binaural recording technology: A historical review and possible future developments. *Acta acustica united with Acustica*, *95*(5), 767-788.

[43] Hammer, K., & Snow, W. (1932). Binaural Transmission System at Academy of Music in Philadelphia. Memorandum MM-3950, Bell Laboratories.

[44] Seeber, B. U., & Fastl, H. (2003). Subjective selection of non-individual head-related transfer functions. Proceedings of the 2003 International Conference on Auditory Display. Boston, MA, USA.

[45] Roginska, A., Santoro, T. S., & Wakefield, G. H. (2010, November). Stimulus-dependent HRTF preference. In Audio Engineering Society Convention 129. Audio Engineering Society.

[46] Carr, K., & England, R. (Eds.). (1995). Simulated and virtual realities: Elements of perception. CRC Press.

[47] Barfield, W., & Furness, T. A. (1995). Virtual environments and advanced interface design. Oxford University Press.

[48] Wickens, C. D., & Baker, P. (1995). Cognitive issues in virtual reality.

[49] Glowinski, D., Baron, N., Shirole, K., Coll, S. Y., Chaabi, L., Ott, T., ... & Grandjean, D. M. (2015). Evaluating music performance and context-sensitivity with Immersive Virtual Environments. EAI Endorsed Transactions on Creative Technologies, 2(2), e3.

[50] Bargar, R., Choi, I., Das, S., & Goudeseune, C. (1994, September). Model based interactive sound for an immersive virtual environment. In Proceedings of the

International Computer Music Conference (pp. 471-471). INTERNATIONAL COMPUTER MUSIC ACCOCIATION.

[51] Serafin, S., Avanzini, F., De Goetzen, A., Erkut, C., Geronazzo, M., Grani, F., ... & Nordahl, R. (2020). Reflections from five years of Sonic Interactions in Virtual Environments workshops. Journal of New Music Research, 49(1), 24-34.

[52] Sanchez-Vives, M. V., & Slater, M. (2005). From presence to consciousness through virtual reality. Nature Reviews Neuroscience, 6(4), 332-339.

[53] Cruz-Neira, C., Sandin, D. J., & DeFanti, T. A. (1993, September). Surround-screen projection-based virtual reality: the design and implementation of the CAVE. In Proceedings of the 20th annual conference on Computer graphics and interactive techniques (pp. 135-142).

[54] Amatriain, X., Höllerer, T., Kuchera-Morin, J., & Pope, S. T. (2007). Immersive audio and music in the allosphere. In ICMC.

[55] Höllerer, T., Kuchera-Morin, J., & Amatriain, X. (2007, August). The allosphere: a large-scale immersive surround-view instrument. In Proceedings of the 2007 workshop on Emerging displays technologies: images and beyond: the future of displays and interacton (pp. 3-es).

[56] Cabrera, A., Kuchera-Morin, J., & Roads, C. (2016). The evolution of spatial audio in the allosphere. Computer Music Journal, 40(4), 47-61.

[57] Graham, R., & Cluett, S. (2016, September). The soundfield as sound object: Virtual reality environments as a three-dimensional canvas for music composition. In Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality. Audio Engineering Society.

[58] Atherton, J., & Wang, G. (2020). Curating perspectives: Incorporating virtual reality into laptop orchestra performance. In Proc. Int. Conf. New Interfaces Musical Expression (pp. 154-159).

[59] Janer, J., Gomez, E., Martorell, A., Miron, M., & de Wit, B. (2016, September). Immersive orchestras: audio processing for orchestral music VR content. In 2016 8th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES) (pp. 1-2). IEEE.

[60] Postma, B. N., Poirier-Quinot, D., Meyer, J., & Katz, B. F. (2016). Virtual reality performance auralization in a calibrated model of Notre-Dame Cathedral. EuroRegio2016, Porto, Portugal.

[61] Zahorik, P. (2002, July). Auditory display of sound source distance. In Proc. Int. Conf. on Auditory Display (pp. 326-332).

[62] Zahorik, P., Brungart, D. S., & Bronkhorst, A. W. (2005). Auditory distance perception in humans: A summary of past and present research. ACTA Acustica united with Acustica, 91(3), 409-420.

[63] Zahorik, P., Kistler, D. J., & Wightman, F. L. (1994). Sound localization in varying virtual acoustic environments. Georgia Institute of Technology.

[64] Cochran, P., Throop, J., & Simpson, W. E. (1968). Estimation of distance of a source of sound. The American journal of psychology, 81(2), 198-206.

[65] Bronkhorst, A. W., & Houtgast, T. (1999). Auditory distance perception in rooms. Nature, 397(6719), 517-520.

[66] Bronkhorst, A. W. (2001). Effect of stimulus properties on auditory distance perception in rooms. Physiological and Psychological Bases of Auditory Function, 184-191.

[67] Mershon, D. H., & Bowers, J. N. (1979). Absolute and relative cues for the auditory perception of egocentric distance. Perception, 8(3), 311-322.

[68] Nielsen, S. H. (1992, March). Auditory distance perception in different rooms. In Audio Engineering Society Convention 92. Audio Engineering Society.

[69] Zahorik, P. (2001). Estimating sound source distance with and without vision. Optometry and vision science, 78(5), 270-275.

[70] McGregor, P., Horn, A. G., & Todd, M. A. (1985). Are familiar sounds ranged more accurately?. Perceptual and motor skills, 61(3).

[71] Coleman, P. D. (1962). Failure to localize the source distance of an unfamiliar sound. The Journal of the Acoustical Society of America, 34(3), 345-346.

[72] Philbeck, J. W., & Mershon, D. H. (2002). Knowledge about typical source output influences perceived auditory distance. The Journal of the Acoustical Society of America, 111(5), 1980-1983.

[73] Hartmann, W. M. (1983). Localization of sound in rooms. The Journal of the Acoustical Society of America, 74(5), 1380-1391.

[74] Holt, R. E., & Thurlow, W. R. (1969). Subject orientation and judgment of distance of a sound source. The Journal of the Acoustical Society of America, 46(6B), 1584-1585.

[75] Mershon, D. H., & King, L. E. (1975). Intensity and reverberation as factors in the auditory perception of egocentric distance. Perception & Psychophysics, 18(6), 409-415.

[76] Peters, N., Lossius, T., Schacher, J., Baltazar, P., Bascou, C., & Place, T. (2009, July). A stratified approach for sound spatialization. In Proceedings of the Sound and Music Computing Conference, Porto, Portugal.

[77] Kostadinov, D., Reiss, J. D., & Mladenov, V. M. (2010, March). Evaluation of distance based amplitude panning for spatial audio. In ICASSP (pp. 285-288).

[78] Benjamin, E., Heller, A., & Lee, R. (2006, October). Localization in horizontal-only ambisonic systems. In Audio Engineering Society Convention 121. Audio Engineering Society.

[79] Stitt, P., Bertet, S., & Van Walstijn, M. (2013, September). Perceptual investigation of image placement with ambisonics for non-centred listeners. In Proc. of the 16th Int. Conference on Digital Audio Effects (DAFx-13), Maynooth, Ireland.

[80] Bates, E., Boland, F., Furlong, D., & Kearney, G. (2007, June). A comparative study of the performance of spatialization techniques for a distributed audience in a concert hall environment. In Audio Engineering Society Conference: 31st International Conference: New Directions in High Resolution Audio. Audio Engineering Society.

[81] Pulkki, V. (2001). Localization of amplitude-panned virtual sources II: Two-and three-dimensional panning. Journal of the Audio Engineering Society, 49(9), 753-767.

[82] Frank, M. (2013). Phantom sources using multiple loudspeakers in the horizontal plane. na.

[83] Ono, K., Pulkki, V., & Karjalainen, M. (2002, April). Binaural modeling of multiple sound source perception: Coloration of wideband sound. In Audio Engineering Society Convention 112. Audio Engineering Society.

[84] Ahrens, J., Geier, M., & Spors, S. (2008, May). The soundscape renderer: A unified spatial audio reproduction framework for arbitrary rendering methods. In Audio Engineering Society Convention 124. Audio Engineering Society.

[85] Ramakrishnan, C., Goßmann, J., & Brümmer, L. (2006, June). The ZKM klangdom. In Proceedings of the 2006 Conference on New Interfaces for Musical Expression (pp. 140-143).

[86] McLeran, A., Roads, C., Sturm, B. L., & Shynk, J. J. (2008). Granular sound spatialization using dictionary-based methods. In Proceedings of the 5th Sound and Music Computing Conference, Berlin, Germany (No. 1).

[87] Gerzon, M. A. (1974). Surround-sound psychoacoustics. Wireless World, 80(1468), 483-486.

[88] Gerzon, M. A. (1977). Itesign of Ambisonic Decoders for Multispeaker Surround Sound.

[89] Ircam Spat URL: http://forumnet.ircam.fr/ product/spat-en/, June, 2017.

[90] https://steinhardt.nyu.edu/people/paul-geluso

[91] Marentakis, G., Zotter, F., & Frank, M. (2014). Vector-base and ambisonic amplitude panning: A comparison using pop, classical, and contemporary spatial music. Acta Acustica united with Acustica, 100(5), 945-955.

[92] Guastavino, C., & Katz, B. F. (2004). Perceptual evaluation of multi-dimensional spatial audio reproduction. The Journal of the Acoustical Society of America, 116(2), 1105-1115.

[93] Thurstone, L. L. (2017). A law of comparative judgment. In Scaling (pp. 81-92). Routledge.

[94] Peteres, N., Marentakis, G., & McAdams, S. (2011). Current technologies and compositional practices for spatialization: A qualitative and quantitative analysis. Computer Music Journal, 35(1), 10-27.

[95] Aswathanarayana, S. (2020, October). Comparison of Spatialization Techniques with Different Music Genres. In Audio Engineering Society Convention 149. Audio Engineering Society.

[96] Rumsey, F. (2002). Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. Journal of the Audio Engineering Society, 50(9), 651-666.

[97] Choisel, S., & Wickelmaier, F. (2007). Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference. The Journal of the Acoustical Society of America, 121(1), 388-400.

[98] Hammershøi, D., & Møller, H. (2002). Methods for binaural recording and reproduction. Acta Acustica united with Acustica, 88(3), 303-311.

[99] Møller, H. (1992). Fundamentals of binaural technology. Applied acoustics, 36(3-4), 171-218.

[100] Hammershøi, D., & Møller, H. (2005). Binaural technique—Basic methods for recording, synthesis, and reproduction. Communication acoustics, 223-254.

[101] Hughes, S., & Kearney, G. (2016, September). Moving virtual source perception in 2d space. In Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality. Audio Engineering Society.

[102] Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Boyer, H., Mayor, O., ... & Serra, X. (2013). Essentia: An audio analysis library for music information retrieval. In Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4-8; Curitiba, Brazil.[place

unknown]: ISMIR; 2013. p. 493-8.. International Society for Music Information Retrieval (ISMIR).

[103] Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the CUIDADO project. CUIDADO Ist Project Report, 54(0), 1-25.

[104] Streich, S. (2006). Music complexity: a multi-faceted description of audio content. Barcelona, Spain: Universitat Pompeu Fabra.

[105] Laurier, C., Meyers, O., Serra, J., Blech, M., Herrera, P., & Serra, X. (2010). Indexing music by mood: design and integration of an automatic content-based annotator. Multimedia Tools and Applications, 48(1), 161-184.

[106] Austin, L., & Smalley, D. (2000). Sound diffusion in composition and performance: an interview with Denis Smalley. Computer Music Journal, 24(2), 10-21.

[107] Brümmer, L., Rabl, G., Boehmer, K., Risset, J. C., Harrison, J., Bayle, F., ... & Stockhausen, K. (2001). Is tape music obsolete? Is spatialization superficial?. Computer Music Journal, 25(4), 5-11.