UNIVERSITY OF CALIFORNIA SAN DIEGO

Space Time Exploration of Musical Instruments

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Music

by

Isaac Garcia Munoz

Committee in charge:

Professor Tamara Smyth, Chair
Professor William S. Hodgkiss
Professor Miller Puckette
Professor Chinary Ung
Professor Shahrokh Yadegari

2020

The dissertation of Isaac Garcia Munoz is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2020

DEDICATION

To my family, distanced by borders yet united by culture.

EPIGRAPH

Happiness is only real,

when shared.

—Christopher McCandless

TABLE OF CONTENTS

Signature Page iii

Dedication iv

Epigraph v

Table of Contents vi

List of Figures ix

List of Tables xi

Nomenclature xii

Preface xiii

Acknowledgements xiv

Vita xv

Abstract of the Dissertation xvi

Chapter 1 Introduction 1
1.1 Overview 1
1.2 Auditory Perception of Sound Sources 2
1.2.1 Binaural Hearing 2
1.2.2 Loudness 4
1.2.3 Reverberation 6
1.2.4 Localization 9
1.3 Spatial Audio Synthesis 10
1.3.1 Binaural Synthesis 10
1.3.2 Head Related Transfer Functions 11
1.3.3 Audio Objects 15
1.3.4 Soundfields 16
1.4 Spatial Audio Software 18
1.4.1 Spatial Audio Software Development Kits 18
1.4.2 Auralization Software 18

Chapter 2 Introduction to Extended Reality Musical Instruments 22
2.1 Augmented Reality Musical Instruments 22
2.1.1 Augmented Reality Audio 22
2.1.2 Survey of AR Musical Instruments 24

2.2 Challenges in Musical Instrument Design for AR 25
2.2.1 AR Latencies 25
2.2.2 AR Reverb 27
2.3 Challenges in Musical Instrument Design for VR 28
2.3.1 Fear UnSound 28
2.3.2 Interaction with the Virtual World 30
2.4 Augmented Reality System Example 31
2.4.1 Computer Vision 31
2.4.2 System Design 32
2.4.3 User Interaction 34
2.4.4 Spatial Audio Block 35
2.4.5 Technical Constraints 36

Chapter 3 Spatial Reverb for Extended Reality Musical Instruments 38
3.1 Spatial Reverb 39
3.2 Directional Room Impulse Responses 40
3.2.1 Measuring DRIRs 41
3.2.2 Objective Analysis of DRIRs 45
3.3 XR Musical Instruments for Evaluation of DRIR 46
3.3.1 AR Electric Drumset 46
3.3.2 AR Electric Guitar 46
3.3.3 Matching the Rooms' Coordinate Systems 48
3.4 Exploring the Room 49
3.4.1 DRR and Clarity of the AR Electric Drumset 50
3.4.2 Synthetic DRIRs 51
3.5 Perceptual Comparison of Spatial Reverb 53
3.5.1 Informal Preference Test 53
3.6 Tuning DRR 60
3.6.1 Interpolation of DRIRs 61
3.7 Considerations of using DRIRs for Spatial Reverb 63

Chapter 4 Spatial Audio Effects for Extended Reality Musical Instruments 64
4.1 Spatial Looping 65
4.2 Spatial Delay 65
4.3 Spatial Feedback 66
4.4 Spatial Compression 67

Chapter 5 Extended Reality Audio Concert 72
5.1 Composing for XR Musical Instruments 73
5.1.1 *VR Singing Kite Concerto* 73
5.1.2 *Between a Log and a Pluck Place* 74
5.1.3 *Push Pull* 75
5.2 Setups for XR Musical Performance 76

5.2.1 Rehearsals 76
5.2.2 Concert Setup 76
5.2.3 Ambisonics Decoder Design 77
5.2.4 XR System 78
5.3 Reflection 79

Chapter 6 Conclusion 83

Appendix A DRIR Plots 86

Appendix B Energy Plots of Measured DRIRs 93

Appendix C *Push Pull* Score 97

Bibliography 100

LIST OF FIGURES

Figure 1.1: Uncoiled cochlea with basilar membrane. 3
Figure 1.2: Fletcher-Munson equal loudness curves. 5
Figure 1.3: A, B and C weightings. 6
Figure 1.4: Normalized impulse response, exponential decay fit, and integral decay curve linear regression fit. 7
Figure 1.5: Sections of reverberation. 9
Figure 1.6: Cone of confusion. 9
Figure 1.7: CIPIC HRTF. 12
Figure 1.8: Six Degrees of Freedom. 13
Figure 1.9: Euler angles. 14

Figure 2.1: Reactable. 24
Figure 2.2: Fear UnSound VR app. 29
Figure 2.3: Diagram of an example AR performance. 32
Figure 2.4: Diagram of log drum and audience. 33
Figure 2.5: 3D Audio Player diagram. 36

Figure 3.1: Convolution spatial reverb. 40
Figure 3.2: Time stretched pulse waveform. 41
Figure 3.3: IR Measurement Tool. 42
Figure 3.4: Sennheiser AMBEO VR Mic. 43
Figure 3.5: Zylia ZM-1 Microphone. 44
Figure 3.6: Ambeo A-B Format Converter. 44
Figure 3.7: DRIR energy plots with source in front. 45
Figure 3.8: Diagram of the setup for the AR drumset. 47
Figure 3.9: Diagram of the setup for the AR guitar. 48
Figure 3.10: Cambodian Singing Kite. 50
Figure 3.11: EVERTims raytracing visualized in Blender. 52
Figure 3.12: EvertSE. 52
Figure 3.13: Preference test MaxMSP UI. 55
Figure 3.14: Naturalness preference histogram. 56
Figure 3.15: AR/VR naturalness preference histogram. 57
Figure 3.16: AR electric guitar/drumset naturalness preference histogram. 58
Figure 3.17: Floorplan of test and rehearsal room 59
Figure 3.18: DRR, clarity and naturalness preference. 61
Figure 3.19: DRIR interpolation. 62

Figure 4.1: Block diagram of the spatial looping effect for an XR audio system. 66
Figure 4.2: Block diagram of the spatial delay effect for an XR audio system. 67
Figure 4.3: Block diagram of the spatial feedback effect for an XR audio system. . . . 68
Figure 4.4: Block diagram of the spatial compression effect for an XR audio system. . 69

Figure 4.5: Screenshot of Directivity Shaper VST Plugin by IEM. 70
Figure 4.6: Screenshot of Directional Compressor VST Plugin by IEM. 71

Figure 5.1: Spherical Sound Search concert flyer. 72
Figure 5.2: Experimental Theater seating chart. 77
Figure 5.3: Screenshot of the IEM AllRADecoder plugin. 78
Figure 5.4: Screenshot of the IEM SimpleDecoder plugin. 78
Figure 5.5: Screenshot of the IEM DistanceCompensator plugin. 80

Figure A.1: Ambeo VR Mic DRIR with impulse at crash. 87
Figure A.2: ZM-1 DRIR with impulse at crash. 87
Figure A.3: Ambeo VR Mic DRIR with impulse at room center. 88
Figure A.4: ZM-1 DRIR with impulse at room center. 88
Figure A.5: Ambeo VR Mic DRIR with impulse at hihat. 89
Figure A.6: ZM-1 DRIR with impulse at hihat. 89
Figure A.7: Ambeo VR Mic DRIR with impulse at kick. 90
Figure A.8: ZM-1 DRIR with impulse at kick. 90
Figure A.9: Ambeo VR Mic DRIR with impulse at snare. 91
Figure A.10: ZM-1 DRIR with impulse at snare. 91
Figure A.11: Ambeo VR Mic DRIR with impulse at low tom. 92
Figure A.12: ZM-1 DRIR with impulse at low tom. 92

Figure B.1: Energy plots for a source at the crash's position. 93
Figure B.2: Energy plots for a source at the center of the test room. 94
Figure B.3: Energy plots for a source at the hihat's position. 94
Figure B.4: Energy plots for a source at the kick's position. 95
Figure B.5: Energy plots for a source at the snare's position. 95
Figure B.6: Energy plots for a source at the low tom's position. 96

Figure C.1: Page 1 of the score for *Push Pull*. 98
Figure C.2: Page 2 of the score for *Push Pull*. 99

LIST OF TABLES

Table 1.1: 3D Audio Spatializers 19
Table 1.2: IEM Plugin Suite Descriptions [17] 20

Table 3.1: DRR and Clarity for Measured DRIRs 51
Table 3.2: Spatial Reverb Conditions for Two-Choice Informal Study 54
Table 3.3: Pilot Test Scenarios 54

NOMENCLATURE

3DoF	Three degrees of freedom
6DoF	Six degrees of freedom
AR	Augmented Reality
BRIR	Binaural Room Impulse Response
CPU	Central Processing Unit
CV	Computer Vision
DAW	Digital Audio Workstation
dB	Decibels
DRIR	Directional Room Impulse Response
DRR	Direct to Reverberant Ratio
DSP	Digital Processing Unit
FDN	Feedback Delay Network
GPU	Graphics Processing Unit
HMD	Head Mounted Display
HRIR	Head Related Impulse Response
HRTF	Head Related Transfer Function
ILD	Interaural Level Difference
ITD	Interaural Time Difference
MR	Mixed Reality
NN	Neural Network
SDK	Software Development Kit
SIL	Sound Intensity Level
SPL	Sound Pressure Level
VR	Virtual Reality
XR	Extended Reality

PREFACE

Upon much reflection during the development of this dissertation, the start of this concept traces back to a fourth grade science fair project. In this project, I built a wooden box with two circular openings to explore the effects of sound in an enclosed space. By pressing my ear upon one opening and inserting a 440Hz tuning fork into the other opening, I was able to hear the acoustics of that sound and how it reverberated throughout the confines of the box. The reverberation and directivity of that tuning fork now applies to this creation of extended reality musical instruments.

Later on in my academic career, I decided to major in Mechanical Engineering with the initial intension of understanding principles of acoustics governing what I had heard in that small wooden box. However, in the course of my studies, I realized that my real motivation was not to just measure sound, but to also modify and manipulate sound through digital signal processing. Through the Electrical Engineering (EE) curriculum, I was fortunate to have gained the technical foundation needed for the instruments developed in this dissertation.

Sounds are planted; they grow from a seed sown at a specific point in space and time by excitation energy. Musical instruments, the seed sowers, create this excitation energy in a patchwork of crops that can be reaped for performances. After completing my degree in EE, I traveled for a year as a Thomas J. Watson Fellow throughout Latin America to learn different ways in which musical instruments are made. My lessons and experiences were made manifest in my UCSD Master's thesis where I built electroacoustic musical instruments to explore modifying their natural sound. This process required amplification either over loudspeakers or headphones separating the modified sound from the instrument that germinated it. My purpose in creating extended reality musical instruments is to enable the harvesting of expressively rich sounds while exploring the instrument-sound relationship in both time and space.

ACKNOWLEDGEMENTS

Many thanks to all of the members of my committee for their guidance and support. A special thank you to my Committee Chair Tamara Smyth for being so understanding of my school-work balance. Thank you fellow Computer Music graduate students for all the fascinating concerts held in the Experimental Theater over the last decade. Jessica Flores and the production staff, thank you for managing the controlled chaos of those concerts. Thank you Technical Event Services for the opportunity to gain live sound engineering experience behind the mixing boards at The Loft, Price Center and Porter's Pub.

I am indebted to Josh Meyer for providing the rhythm at my concert and being so patient while I tuned the drumset during rehearsals. Thank you Melissa Olson and Emilia Meyer for your time spent listening to spatial reverbs. A heartfelt thank you to my wife for your patience all the times that I retreated to make sounds in my space at home.

Mom, thank you for helping me build that wooden box in 4th grade even though you had no experience being a carpenter. You continue to amaze me with your strength and perseverance.

VITA

2007	B. S. in Electrical Engineering, California Institute of Technology
2007-2008	Thomas J. Watson Fellow
2008-2011	Test Engineer, Broadcomm
2011-2013	Graduate Teaching Assistant, University of California, San Diego
2013	M.A. in Music, University of California, San Diego
2013-2017	Audio Test Engineer, Qualcomm Technologies, Inc.
2017-Present	Senior Engineer, Multimedia R&D - Audio, Qualcomm Technologies, Inc.
2020	PhD. in Music, University of California, San Diego

PUBLICATIONS

Garcia-Munoz, Issac. Transforming the teponaztli. *UCSD Master's Thesis*. 2013.

Davis, Graham and Andre, Schevciw and Munoz, Isaac and Peters, Nils, "Perceptual Evaluation of Personalized BRIRs and Headphone Compensation", *Audio Engineering Society Conference: 2019 AES INTERNATIONAL CONFERENCE ON HEADPHONE TECHNOLOGY*, 2019.

ABSTRACT OF THE DISSERTATION

Space Time Exploration of Musical Instruments

by

Isaac Garcia Munoz

Doctor of Philosophy in Music

University of California San Diego, 2020

Professor Tamara Smyth, Chair

Musical instruments are tools used to generate sounds for musical expression. Virtual Reality (VR) and Augmented Reality (AR) musical instruments create sounds that may be spatially disjointed from the instrument controls. Spatial audio processing can be used to position the Extended Reality (XR) musical instruments and their corresponding sounds in the same space. This dissertation investigates novel ways of combining spatial reverb models to improve the naturalness of XR musical instruments. Seven spatial reverb systems, combinations of a shoebox spatial reverb model, a raytracing spatial reverb model, and measured directional room impulse response convolution reverb, were compared in a pilot study. A novel hybrid system of synthetic early reflections and directional room impulse responses was preferred for naturalness when

tested over headphones with three instruments created by the author: AR electric guitar, AR electric drumset, and VR Singing Kite. This research culminated in a concert, *Spherical Sound Search*, which showcased the preferred hybrid system, the three XR musical instruments, and four re-contextualized spatial audio effects (spatial looping, spatial delay, spatial feedback, and spatial compression). The three pieces in the concert explored different aspects of XR modalities and presented the novel system with spatial audio effects to a larger audience by rendering to an octophonic loudspeaker layout.

Chapter 1

Introduction

1.1 Overview

Computer Music performances have showcased complex sound syntheses as well as real-time algorithms where audio is sampled and processed in discrete moments in time [68]. In these performances, it is common to see a performer looking down at their computer screen. The ability for a performer to impart musically expressive gestures has increased through novel computer interfaces that use pressure sensitive buttons [2, 30] or that track the performer's hands [117, 134, 141]. As technology improves beyond the computer and on to newer platforms there is potential for expanding upon these new digital processing interfaces and enhancing the musical experience. With these new technologies, there is a possibility that virtual sounds can occupy physical space and interact with the 'real world' just like acoustic instruments.

The primary goal of this dissertation was to explore advancements in spatial audio technologies with computer enhanced musical instruments to promote expressive sound generation. A novel approach to spatial reverb (Chapter 3) and re-contextualized spatial audio effects (Chapter 4) are original contributions to the field of Computer Music in this research—highlighted in a concert (Chapter 5) featuring three musical instruments designed to evaluate spatial reverb.

This chapter introduces the concepts of psychoacoustics and spatial audio to describe how sounds have a position in space. The first part will review binaural hearing which allows humans to localize the direction and distance (to a certain degree) of sound. The second part is an overview of spatial audio techniques used both to preserve a recorded sound's position in space and to give a virtual/synthesized sound a position in space. Knowledge of these concepts is integral in understanding the computer enhanced musical instruments that were developed in this research. These instruments were designed to explore the expressive creation of sound in a new reality, challenging preconceived notions of space and time.

1.2 Auditory Perception of Sound Sources

Spatial audio is a complex interdisciplinary field combining psychoacoustics and engineering. This section summarizes concepts of psychoacoustics[1] that illustrate how humans localize a sound source.

1.2.1 Binaural Hearing

The full neurological pathways of hearing are not fully understood, but studies have described several effects of how sounds are perceived. These effects include the ventriloquist effect (relating visual references to sound sources), the cocktail party effect (picking out particular voices from a crowd), and the precedence effect (hearing sounds as if they are coming from the closest object instead of the actual source farther away) [64].

Binaural hearing is the act of perceiving sounds by using both ears to convert and process the fluctuations in pressure that are funneled into the ears. Figure 1.1 shows the pinna (external part of the ear) and an uncoiled choclea in the inner ear. Sound waves are filtered as they reflect off the folds of the pinna and shoulders as they are transfered to the inner ear which analyzes

[1] For a comprehensive introduction into psychoacoustics please refer to Perry Cook's *Music, Cognition and Computer Sound* [64].

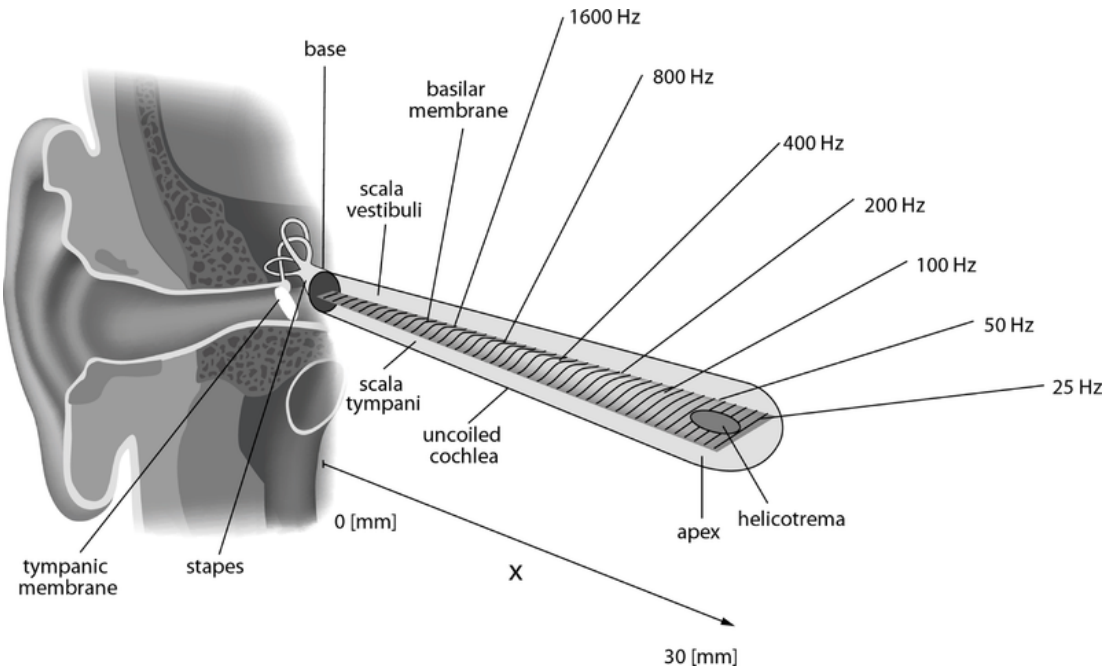


Figure 1.1: Kern A, Heid C, Steeb W-H, Stoop N, Stoop R. *The position x of the maximal amplitude of the travelling wave corresponds in a 1-to-1 way to a stimulus frequency.* http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1000161. 29 August 2008.

the energy of different frequencies of sound at different points along the choclea. Each person's ability to localize sounds is adapted to the unique shape of their pinnae. Binaural hearing weighs the differences in sound arriving at each ear in terms of the interaural time difference (ITD), which is the difference in time of arrival, and the interaural level difference (ILD), which is the difference in level. With this differential information, which is also frequency dependent, the brain develops an auditory image of the surrounding scene in time and space.

Visual references are used as cues for where sounds are perceived to be originating; if we see what we believe is the source of the sound, then binaural hearing aligns to what we see. Television and film are prime examples of the ventriloquist effect. We perceive the voices coming from the actors on the screen even if the speakers are not exactly co-located with the image of the actors. This effect has been studied to be an automatic response without the need for visual context, meaning that sounds will be aligned even if the listener has no previous reference of the sounds made by the visual objects. Therefore, if new computer enhanced musical instruments

make use of sound reproduction over headphones or loudspeakers, then the listener is likely to associate sounds as coming from the instrument even if they have never seen the instrument or heard how it sounds.

The field of audio spatialization (Section 1.3) was informed by binaural hearing perception and has focused on reproducing accurate localization of sound sources (Section 1.2.4), sometimes at the expense of sounding natural. Reverberation (Section 1.2.3) can be introduced or modified to improve the naturalness of audio spatialization—the subject of Chapter 3.

1.2.2 Loudness

Loudness relates to the perceived intensity of a sound at a given distance. The intensity is given by

$$I = \frac{p^2}{\rho c} \tag{1.1}$$

where p is the sound pressure, ρ is the density of air (kg/m^3), and c is the speed of sound (m/s). For sound sources that radiate energy equally in all directions, the pressure decreases by $\frac{1}{distance}$. Intensity is a measure of how much acoustic power is applied to an area. Humans hear changes in sound level on a logarithmic scale. Therefore, the Sound Intensity Level (SIL), which is reported in decibels (dB), is given by

$$SIL = 10\log_{10}\left(\frac{I}{I_o}\right) \tag{1.2}$$

where I_o is a reference intensity. Typically, pressure is easier to measure[2] than intensity making it more common to report sound levels using the Sound Pressure Level (SPL) given by

$$SPL = 20\log_{10}\left(\frac{p}{p_o}\right) \tag{1.3}$$

[2]Using calibrated microphones.

where p_o is a reference pressure. The reference intensity is usually set to the threshold of audibility, $I_o = 10^{-12}\frac{W}{m^2}$, and the reference pressure to the threshold of hearing, $P_o = 2\times 10^{-5}\frac{N}{m^2}$. When $\rho c \approx 400$, the SIL is equal to the SPL.

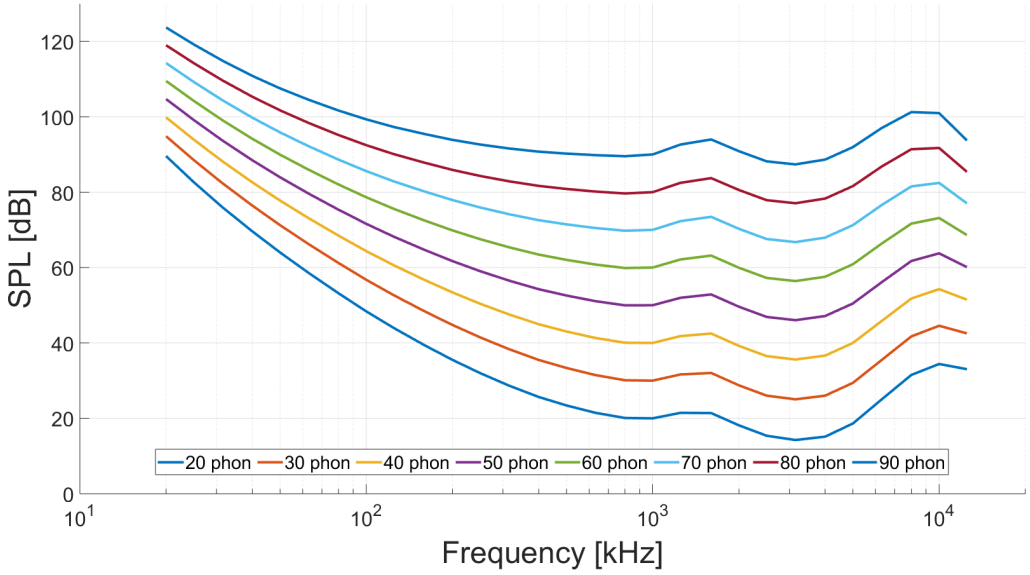


Figure 1.2: Fletcher-Munson equal loudness curves. The lines trace the SPL value required to perceive the same loudness across the frequencies in the human range of hearing. Each line represents a phon value ranging from 20 phon (lower line) to 90 phon (upper line) [145].

Perceived loudness depends on both the SPL and frequency content of the sound source. Fletcher and Munsen expressed this frequency dependence in phon[3] equal loudness curves [73]. Figure 1.2 shows how each phon represents the SPL required for sounds at different frequencies to be perceived equally loud.

SPL meters can apply different frequency dependent weighting scales (Figure 1.3) to their measurements. Of these scales, A-weighting is more aligned with the Fletcher-Munson curves in its attenuation of the low and high frequencies. To represent relative loudness of sounds there is yet another unit of measurement, the sone, which relates to phons by

$$phon = 40 + 10\log_2(sone) \tag{1.4}$$

[3]Phon are defined as the SPL in dB at 1kHz.

such that doubling the sone is equivalent to doubling the perceived loudness[4] [133]. Spatial audio

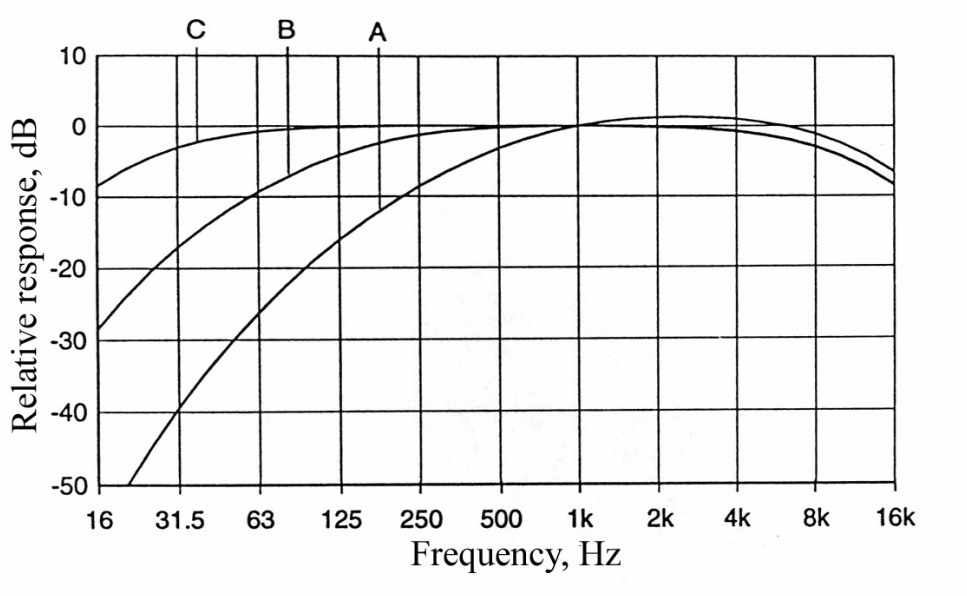


Figure 1.3: *A, B and C weightings.* https://sites.google.com/site/cliftonwindeducation/. February 2020.

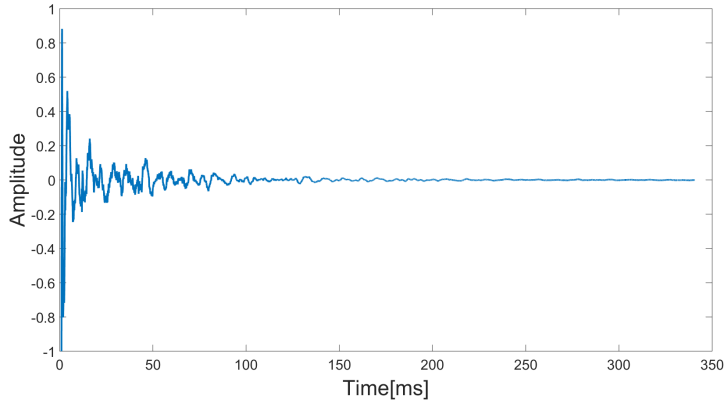
software presented later in Section 1.4 apply distance-based functions to sound sources in an attempt to reproduce the complex frequency dependent perception of loudness.

1.2.3 Reverberation

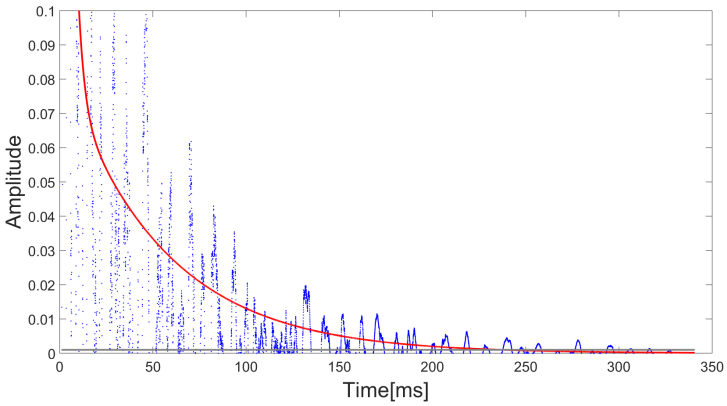
Reverberation (reverb) is the sum of all the surface reflections of a sound. Reverb times provide a partial description of reverberant spaces[5] and can be calculated from room impulse responses usings a variety of methods [138, 149, 151]. Impulse response measurements will be discussed in detail in Section 1.3.2 and Section 3.2.1. RT_{60} and RT_{30} are frequency dependent measurements of the time it takes the reverberation to decrease 60dB and 30dB respectively. Figure 1.4 shows two methods which give different RT_{60} times for the same impulse response. The first method depicted in Figure 1.4b fits an exponential decay curve (red line) to the amplitude

[4]Empirically derived for pure tones.

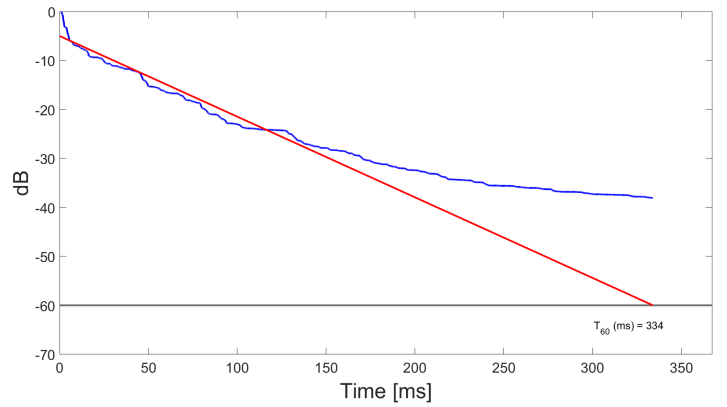
[5]Other measures of reverb, in addition to reverb times, are defined in ISO 3382-1/2 Standard [47].



(a)



(b)



(c)

Figure 1.4: A normalized impulse response is shown in (a). Two methods for calculating RT_{60} time are shown in (b), an exponential decay fit, and (c), an integral decay curve of the impulse response with a linear regression fit [58]. $RT_{60} \approx 240$ms for (b) and $RT_{60} \approx 334$ms for (c).

of the impulse response. The second method depicted in Figure 1.4c uses linear regression to fit a line (red line) to the integral decay curve approximated by

$$D(k) = \sum_{n=k}^{N} h^2(n) \tag{1.5}$$

where $h(n)$ is the impulse response, n are integer samples, k is discrete time and N is the length of the impulse response. The RT_{60} times are the time values where the fit lines intersect -60dB (grey horizontal line). In practice, the noise level of the recording equipment, or room, may be higher than the -60dB point of the impulse response. This is illustrated in the difficulty to visualize the RT_{60} point[6] in Figure 1.4b because it is located on the flat part of the decay curve near the noise floor. The same is manifest on Figure 1.4c as a gap between the integral decay curve and the linear regression fit. In such a scenario, reporting the RT_{30} time would be a more accurate description of the reverberant space.

Larger, more reverberant rooms like auditoriums generally have a larger RT_{60} when compared to smaller rooms. Materials can transmit, reflect and absorb sound energy which is characterized by propagation coefficients. Materials in the room, in combination with the dimensions of the room, affect the reverberation time. The Sabine Formula [71] given by

$$RT_{60} = 0.161 \frac{V}{S\alpha} \tag{1.6}$$

is an empirically derived relation between the volume, V, the total surface area, S, and average absorption, α, of a room that provides a measure of the room's RT_{60}. The more reflective the material (less absorption), the higher the RT_{60} time. The reverb of acoustic environments can be broken down into three parts: early reflections from the closest surfaces, late reflections arriving from surfaces farther away and from multiple reflections from nearby surfaces, and a diffuse field of decaying sound energy (Figure 1.5).

[6] 10^{-3} amplitude when the impulse response peak is normalized to 1.

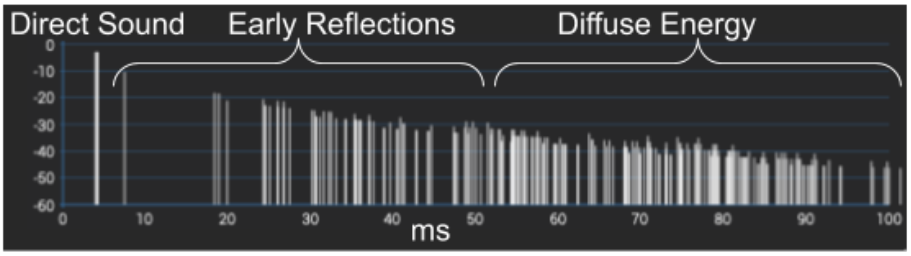


Figure 1.5: Sections of reverberation.

Some of the spatial audio tools of Section 1.4.1 provide parameters associated with reverb time and reflections to tune the reverb effect, but not necessarily to reproduce the reverb of a specific room. The auralization software in Section 1.4.2 models specific rooms more precisely which may result in more realistic RT_{60} times but make it more difficult to tune the reverb.

1.2.4 Localization

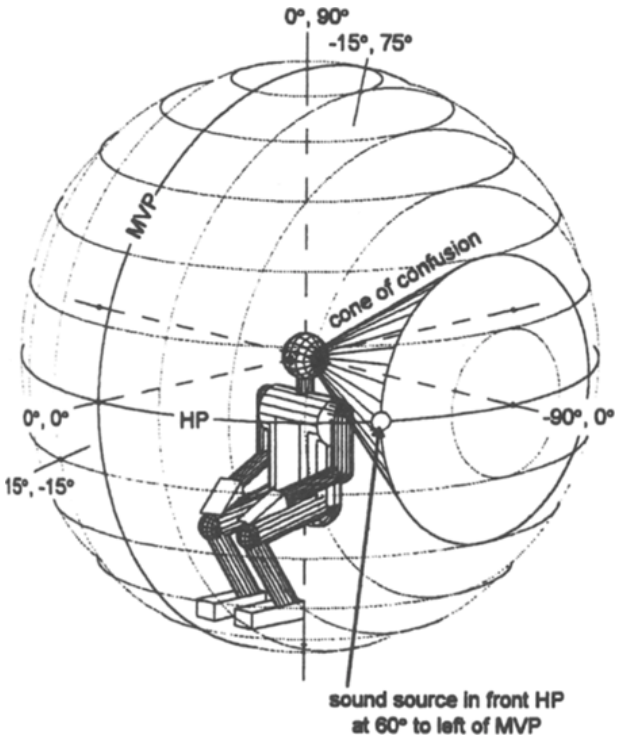


Figure 1.6: Perrett, Stephen and Noble, William. Schematic of spatial references relative to a listener,showing a cone of confusion at a leftward azimuth angle of 60^{o} [123].

Localization is the ability to discern the direction of a sound source, usually given in polar coordinates (azimuth^{o} and elevation^{o} as shown in Figure 1.6). Humans can distinguish sound sources within a few degrees on the horizontal plane, but have less definition with respect to elevation. There are also cones of confusion where the sound sources are perceived to be coming from the rear when they are presented at the front and vice versa [123].

Humans have a poor judgment of absolute distance. Without reverberation (in an anechoic chamber) it was shown that test subjects perceived sound to arrive from the closest position for static sources [64]. In other words, without reverb humans find it difficult to judge distance. Moving sources provide distance cues which improve distance localization [99]. The doppler shift effect (think of a passing ambulance), gives a sense of motion towards and away from a listener. Even for sounds with reflections, studies have proven that it is challenging to judge absolute distances [152]. It is a complex combination of sound level (Section 1.2.2), Direct to Reverberant Ratio (Section 3.2.2), spectral cues, binaural cues (Section 1.2.1), and dynamic cues that makes up the auditory perception of distance [99].

1.3 Spatial Audio Synthesis

1.3.1 Binaural Synthesis

Binaural synthesis enables the use of sound objects with directionality in the sound design of real-time applications such as Virtual Reality video games. One way this 3D audio effect can be realized is by processing mono audio clips with Head Related Transfer Functions (HRTFs) and listening to the output over headphones.

Techniques for creating an immersive audio experience range from simple stereo widening in which a delay is added to broaden the stereo image [78], to multi-channel surround sound processing [88]. Alternatively, binaural synthesis (also referred to as binauralization or binaural rendering), is a way to replicate natural hearing by reproducing sound cues at the ears of the

listener using two channels, normally through headphones.

1.3.2 Head Related Transfer Functions

A basic approach to create spatial audio is to apply HRTFs to a mono signal. This process is a type of audio filtering which imparts the aural cues of a precise point in 360^{o} space when listening with headphones. These HRTFs reproduce sound cues of binarual hearing (Section 1.2.1); ITD, ILD, and tonal information. HRTFs can be created from impulse responses using real or manequin head and torso simulators (HATs). The recordings are responses to a sound excitation at various positions around the head. The set of recordings is called Head Related Impulse Responses (HRIR)[7] and have been commonly acquired using HATs like the Knowles Electronics Manikin for Acoustic Research (KEMAR) or the KU100 (Neumann) [54]. Researchers at the University of California Davis have provided the CIPIC HRTF database [51] where they took 1250 impulse response measurements from human subjects as well as the KEMAR at locations shown in Figure 1.7. The MIT Media Lab played pseudo-random binary pulses from speakers 1.4 meters from the KEMAR to obtain impulses at 710 different positions [77]. However, HRIR does not account for situations where the sound objects move or the listener changes position. Currently, there are a number of spatialization software development kits (SDKs) that provide different ways of applying HRTFs, sampled discretely in space, to accomodate for both moving audio sources and moving listener position (see Section 1.4).

Measuring HRIR is another area of research which will only be mentioned briefly here. In purely digital systems, HRIR can be measured by capturing the output of a system when a unit impulse is given as the input. In acoustic measurements, however, loudspeakers are poor at creating a unit impulse requiring other means to obtain the response. A comparison of HRIR measurement techniques shows that excitation energy given by Maximum Length Sequence

[7] HRTFs and HRIRs may refer to the same dataset stored as audio files in the case of HRIRs and as HRTFs when the HRIRs are brought into the frequency space.

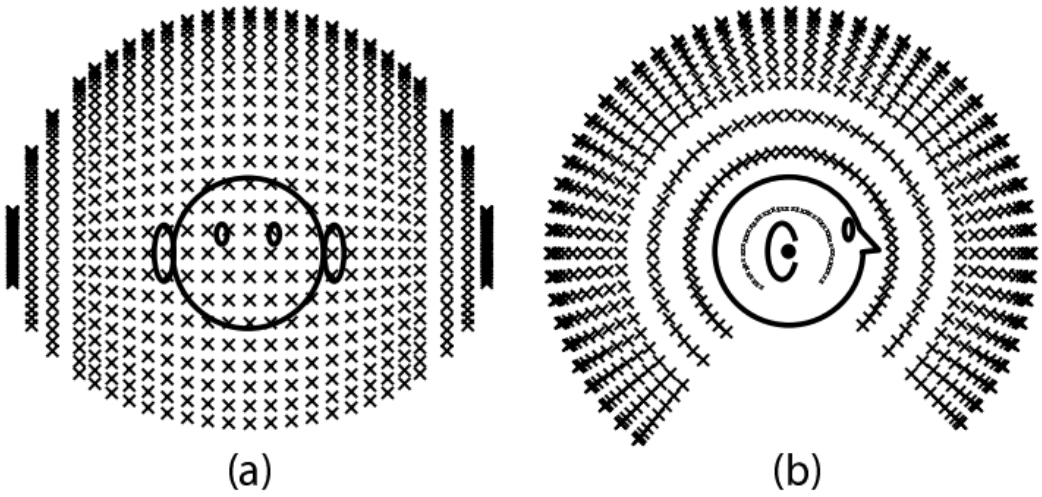


Figure 1.7: Locations of data points from the front (a) and the side (b) for the CIPIC HRTF database [51].

(MLS) and Inverse Repeated Sequence (IRS) are good at rejecting colored noise[8], while Time-Stretched Pulse and SineSweep have less distortion for quiet recording environments [143]. For the measurements performed in Chapter 3, the MaxMSP [23] project IR Measurement Toolbox [20] was used to implement the Time-Stretched Pulse method[9].

Listening over headphones removes the precedence effect experienced when listening to loudspeakers [150], as the audio signals played through headphones arrive simultaneously to each ear. This also means that there is no crosstalk interference. The recorded impulse responses are then used as digital audio filters. When those HRTFs are convolved with a mono input signal, the sound source is reconstructed at the elevation and azimuth of the HRTF set. It is important to note that the pinna (outer ear) is vital for human sound localization (Section 1.3.4).

HRTF binaural synthesis is also used in the decoding of sound fields formats. In theory, the binaural signals are derived by integrating plane waves with the HRTFs over the unit sphere [104].

Listening to binaural synthesis over headphones is an approximate spatialization effect. Perfect sound reconstruction would require a personalized set of HRTFs for every person because

[8]Colored noise is filtered white noise.
[9]HISSTools [84] is another option for MaxMSP users to obtain impulse responses.

of the unique shape of their pinnae as mentioned previously in Section 1.2.1. Measurements in hearing aid research show how the ITD and ILD compare with a spherical model of the head [97] where the differences are due to reflections from the pinna and body. Having musicians record their own HRIR with in-ear microphones in anechoic chambers would capture their unique pinna and body reflections but it is impractical and has led to a one-size-fits-all model.

Front-to-back confusion is a known issue in both binaural synthesis and binaural hearing. Although if the trajectory of the object is known and/or if the player rotates their head then this minimizes the font-to-back confusion [90]. Along the circle created by the *cone of confusion* seen previously in Figure 1.6 the ITD and ILD values are the same making it hard to specify the exact azimuth and elevation [62]. Cheng also points out that sounds along the median plane, at 0^{o} azimuth, may sound like they are inside the head.

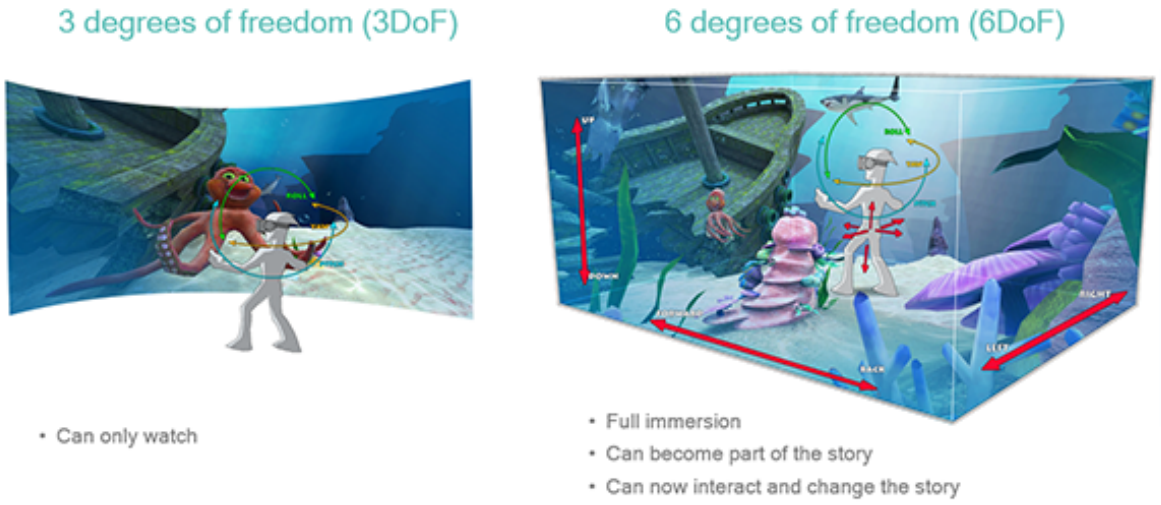


Figure 1.8: Hiren Bhinde. *Experience Six Degrees of Freedom in XR Development*. https://developer.qualcomm.com/blog/experience-six-degrees-freedom-xr-development. 23 October 2017.

When only the listener's head orientation (the direction where the listener is looking) is tracked, it is referred to as three degrees of freedom (3DoF) because it can be described in three Euler angles (roll, pitch, yaw or ϕ,θ,ψ as shown in Figure 1.9). If the listener's position in the real world is also tracked in space (for example; x-axis, y-axis, z-axis) then this is refered to as

six degrees of freedom (6DoF). In practice, changing the orientation of objects can be performed

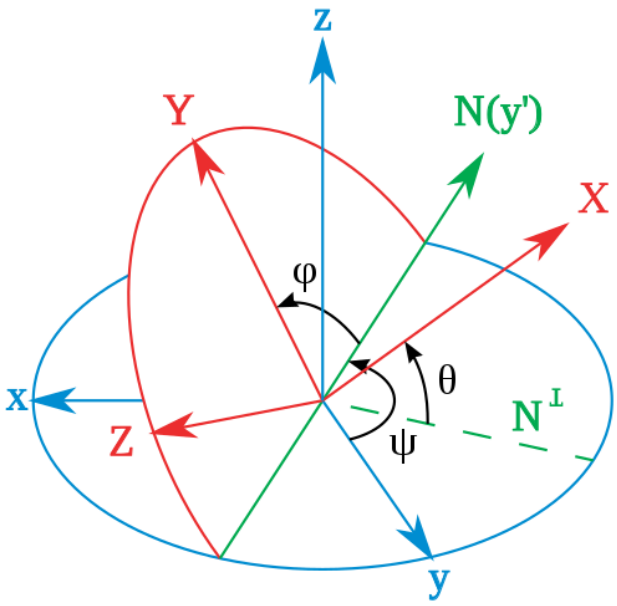


Figure 1.9: Juansempere. *Tait-Bryan angles, ZYX convention*. CC BY 3.0. 20 July 2009.

by 3*x*3 matrices for each angle. Alternatively, rotations can be performed using quaternions [87],

$$q = q_0 + q_1\,\mathbf{i} + q_2\,\mathbf{j} + q_3\,\mathbf{k} \tag{1.7}$$

where i, j, k are complex[10], and q_0, q_1, q_2, q_3 are real numbers constrained to have a unit norm:

$$\sqrt{q_0^2 + q_1^2 + q_2^2 + q_3^2} = 1. \tag{1.8}$$

Quaternions are used for describing motion of HMDs because they are not susceptible to gimbal

[10] $i^2 = j^2 = k^2 = ijk = -1$

locks[11] as are Euler angles. Conversion from quaternion to roll, pitch and yaw is given by

$$\begin{bmatrix} \phi \\ \theta \\ \psi \end{bmatrix} = \begin{bmatrix} \arctan \frac{2(q_0 q_1 + q_2 q_3)}{1-2(q_1^2+q_2^2)} \\ \arcsin 2(q_0 q_2 - q_3 q_1) \\ \arctan \frac{2(q_0 q_3 + q_2 q_2)}{1-2(q_2^2+q_2^3)} \end{bmatrix} \tag{1.9}$$

for scenarios where the HMD provides sensor data in quaternion and spatial audio tools require Euler angles.

1.3.3 Audio Objects

An audio object is defined to represent a singular sound source. It should contain at least one audio signal usually referred to as a mono sound source. What differentiates and audio object from an audio signal is that audio objects also contains metadata with properties about itself which may include its position, orientation and directivity.

Rotating one's head may help alleviate these confusions, but does not help when dealing with distance. If there is only one set of HRTFs then a distance model is used to place the object perceptually closer or farther away by scaling the volume of the sound object with a gain factor. In video game audio plugins there are usually two modes when dealing with distances. One where the gain is linear and another where it is logarithmic with respect to the distance. And if there are different radii of HRTFs available then there is the issue of interpolating along the radial direction. There will be an issue for near-field binaural synthesis if the HRTFs used are available at only one distance then only applying a gain. The addition of reverb helps determine the distance of a sound source [133]. Moore's model for spatial processing of sounds is a ray-tracing model which was initially intended for loudspeaker setups [107], but can also be used in binaural synthesis can now be achieved thanks to GPU optimizations [34]. Section 2.2.2 describes another solution for

[11] Gimbal locks occur when the listener is oriented parallel to the vertical axis loosing one degree of freedom.

dealing with reverb for binaural synthesis in the context of XR musical instruments.

If the number of sound objects to be rendered increases then binauralization can become computationally intensive due to the processing HRTFs for each object. If optimizations are made to estimate the HRTF with less computationally demanding filters then the result is degraded down to that of the spherical head model in [97]. To increase the number of audio objects that can be played, optimizations in the processors (CPU, DSP, GPU) have been made. Another area of spatial audio describes the audio scene as a whole instead of in individual audio objects.

1.3.4 Soundfields

Soundfields represent the sound coming from every direction and arriving at a single position. There are soundfield microphones and programs that can create soundfield content.

Spherical Harmonic Processing

The fundamental difference between soundfields and audio objects is that soundfields sample sound waves at positions on a sphere whose origin is the intersection of the sound waves' vectors. Spherical harmonic processing applies the wave equation solution on a unit sphere to encode spatially sampled audio by taking the spherical harmonic transform:

$$X_{lm} = \int_{\psi=0}^{2\pi} \int_{\theta=0}^{\pi} X(\theta,\psi) Y(\theta,\psi) \sin\theta d\theta d\psi \tag{1.10}$$

where $Y(\theta,\psi)$ is the spherical harmonic, $X(\theta,\psi)$ is an integrable function on a sphere where θ,ψ are spherical coordinates [128]. In Figure 1.9, θ denotes elevation angles, ψ denotes azimuth

angles and they are related to Cartesian coordinates by

$$
\begin{aligned}
x &= r\sin\theta\cos\psi, \\
y &= r\sin\theta\sin\psi, \\
z &= r\cos\theta,
\end{aligned}
$$

where r is the radius.

Ambisonics is the resultant audio format of spherical harmonic processing with order l and Ambisonic Channel Numbering (ACN) given by $l(l+1)+m$ [128]. Ambisonic microphones are made up of at least four individual microphones whose positions are the spatially sampled audio for the $X(\theta,\psi)$ function in Equation 1.10.

Rendering and Reproduction

Once the soundfields are in Ambisonics format, they can be rendered (played back) over headphones or loudspeakers. For headphones, the process is also called binauralization which too makes use of HRTFs mentioned previously for audio object rendering. For loudspeakers, rendering requires an Ambisonics decoder designed for the order of Ambisonics input and the position of the loudspeakers relative to the listener. Both of these methods enable the listener to experience the audio in 3DoF if the listener's head movements are tracked.

Soundfield Navigation

In 6DoF audio playback, where the listener is mobile, multiple Ambisonics signals sampled at different positions can be used. Princeton researchers have demonstrated this virtual navigation of Ambisonic soundfields [148]. Zylia provides a 6DoF SDK [44] that uses nine ZM-1 [46] microphones for 6DoF capture and playback of soundfields [121].

1.4 Spatial Audio Software

While spatial audio is not yet widely adopted into audio mixing workflows, there are an increasing amount of spatial audio frameworks available as both audio plugins and standalone software.

1.4.1 Spatial Audio Software Development Kits

This section lists spatial audio software targeting different platforms from low level embedded systems to high level game engines. OpenAL [89] provides support for 3D audio on desktop platforms and OpenSL ES [82] offers a low level audio API targeting embedded devices. Spatial Audio Designer is a software for mixing 3D audio [35]. Table 1.1 outlines three major 3D audio spatializers and their supported platforms.

Resonance Audio by Google Inc. provides support for easy integration into major video game engines and mobile devices [79]. Spatializer plugins like the Oculus VST/AAX are now supporting Head Mounted Displays (HMD) to enable head tracking in DAW workflows, where as before this was limited to VR game sound design. The Designing Sound article *Let's Test: 3D Audio Spatializer Plugins* provides a thorough review of Oculus Spatializer, Resonance Audio, and Steam Audio spatializers [80]. IEM Plugins [18] are a comprehensive suite of VST plugins for real-time audio spatialization using Ambisonics [53]. Table 1.2 lists their main features. These plugins became the building blocks of the XR musical instruments introduced in Chapter 3 because of their range of audio effects and their ability to interface with tracking data from HMDs.

1.4.2 Auralization Software

There are currently several software suites for reconstructing acoustics. Spat software, developed by Ircam, offers tools to artificially create reverb and spatialize audio [61, 92]. RWTH

Table 1.1: 3D Audio Spatializers

Spatializer	**Game Engine Audio Plugins**	**DAW Plugins**	**Platforms**
Steam Audio [36]	• Unity [39] • Unreal [40] • FMOD [13]		• Win • Mac • Linux • Android
Oculus Spatializer [28]	• Unity • FMOD • Wwise [43]	• VST [37] • AAX [1]	• Win • Mac
Resonance Audio [41]	• Unity • Unreal • FMOD • Wwise	• VST	• Android • iOS • Web

Aachen University has developed a Virtual Acoustics auralization framework for research experiments [91]. SMIR-Generator, developed by AudioLabs Erlangen, simulates sound pressure signals at arriving at spherical microphone arrays [93]. EVERTims [10] is an open source real-time auralization framework with an interface in Blender [8] 3D modeling software [127]. SoundParticles [33] offers a computer graphics approach to 3D audio rendering. Project Acoustics [29], developed by Microsoft, is an acoustics engine running in the cloud for use with Unity and Unreal game engines. As virtual reconstruction of the real world is done faster and

Table 1.2: IEM Plugin Suite Descriptions [17]

AllRADecoder	Used to design an Ambisonic decoder for an loudspeaker layouts using the AllRAD approach [155].
BinauralDecoder	Renders the Ambisonic input signal to a binaural headphone signal using the MagLS approach proposed in [137] using Neumann KU 100 HRTFs.
CoordinateConverter	Converts parameters from a spherical representation to a cartesian, and vice versa.
DirectionalCompressor	Ambisonic compressor/limiter to control dynamics for different spatial regions.
DirectivityShaper	Directional filter with four frequency bands which can be applied to directivity.
DistanceCompensator	Gain compensation applied to loudspeakers based on listening position.
DualDelay	Dual rotational delay.
EnergyVisualizer	Used to visualize the energy distribution on the sphere of an Ambisonics signal using an energy preserving projection.
FdnReverb	Feedback delay network reverberation.
MatrixMultiplier	Applies matrix multiplication to the input signal.
MultiBandCompressor	Four frequency band compressor.
MultiEncoder	Encode up to 64 audio signals into Ambisonics.
MultiEQ	Multichannel equalizer for up to 64 audio signals.
OmniCompressor	Omnidirectional compressor.
ProbeDecoder	Sampling the Ambisonics input at a given direction.
RoomEncoder	Virtual shoebox room model for listener and source positions.
SceneRotator	Ambisonics rotation.
SimpleDecoder	Decodes to loudspeakers using JSON config file from AllRADDecoder.
StereoEncoder	Mono and stereo encoder to Ambisonics.
ToolBox	Operations on an Ambisonics input signal.

with greater detail, these auralization programs begin to sound as natural as real world recordings, and in combination with spatial audio SDKs, may play a greater part in synthesizing the sounds of XR musical instruments.

The spatial audio tools presented in the previous sections provide methods for binaural synthesis to create localizable sounds informed by human perception of sound sources. The next chapter begins with an overview of musical instruments that incorporate interactions with virtual sound sources. The virtual audio game, *Fear UnSound*, and an electroacoustic drum system developed by the author, are used as examples to highlight the challenges of integrating spatial audio into musical instrument design. Chapter 3 describes a pilot preference study created for this research aimed at determining listeners' preference in spatial reverb processed musical instruments based on naturalness. This study also includes a new hybrid approach to spatial reverb. Sections 3.3.1 and 3.3.2 illustrate the designs of the computer enhanced musical instruments based on an electric drumset and electric guitar respectively. A new virtual musical instrument was also developed for the pilot study (Section 3.4). The resultant recontextualization of traditional audio effects[12] used with the musical instruments is described in detail in Chapter 4. This research culminated in a spatial audio concert, *Spherical Sound Search*, in which the author performed three new compositions featuring the spatial audio effects using the preferred spatial audio reverb. The setup, rehearsal, and performance of the concert are discussed in Chapter 5.

[12]Spatial looping (Section 4.1), spatial delay (Section 4.2), spatial feedback (Section 4.3), and spatial compression (Section 4.4).

Chapter 2

Introduction to Extended Reality Musical Instruments

Extended Reality (XR) is an umbrella term encompassing Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR). VR creates an entirely artificial visual and auditory world, AR superimposes images or sounds on top of the real world, and MR describes systems that are a combination of VR plus AR. This chapter provides a history of and current state of XR musical instruments. The challenges in developing immersive audio applications for each environment are discussed by means of the VR application *Fear UnSound* and an electroacoustic log drum AR musical instrument system developed for this research.

2.1 Augmented Reality Musical Instruments

2.1.1 Augmented Reality Audio

Augmented Reality audio was first defined as creating an artificial reverberant room and most of the AR audio applications were navigation based [90]. Tony Huang, an Associate Professor at the University of Technology Sydney, makes a distinction that AR audio requires live

sounds to be captured and augmented. In navigation based AR audio applications, the listener triggers sounds based on the listener's position in the room. Porting VR musical instruments into AR platforms [96, 154] does not necessarily make them AR audio musical instruments. Examples of single user AR musical instruments are those created to teach users how to play notes on a traditional musical instrument [60]. *Augmented Groove* was the first collaborative AR musical experience using mobile devices inside head mounted displays (HMD) to track physical cards used as the interface for music playback control [129, 130]. More recently, the AR art installation *Listening Mirrors* makes use of mobile devices playing audio through bone conduction headphones to show how immersive the experience can be [98] when mixing live and processed sounds. Another advantage of having an audience listen to spatial audio over headphones is that they are not constrained to a 'sweet spot' area in which the 3D audio effects are realized for the listener.

This chapter focuses on AR audio because it facilitates the use of physical musical instruments as the interface for new XR modalities. There are two methods for creating AR audio with XR musical instruments. In the first method, a live sound is recorded in real-time, processed by the computer, and then played back. This would be similar to using effects pedals with an electric guitar. In the second method, the sound or impulse from one instrument can be used as a trigger for the synthesis or playback of other sounds. An example of this would be an electric drum pad where the drum hits are inaudible impulses recorded using piezoelectric contact microphones to trigger the playback of preloaded drum samples. This method is used in the AR drum set created for this research and will be presented in Section 3.3. The ideal AR musical instrument would be sharable, similar to *Augmented Groove*, so that more than one person could experience the performance simultaneously. This is possible if every listener is tracked in 6DoF and receives a personalized binaural synthesis audio stream. However, this is not easily scalable, Chapter 5 summarizes a demonstration of how XR musical instruments can be performed over loudspeakers to accommodate larger audiences.

2.1.2 Survey of AR Musical Instruments

Arguably, the most popular AR instrument has been the *Reactable* (Figure 2.1) because it was popularized in performances by Bjork [94]. The *Reactable* is a collaborative synthesizer that uses cameras to recognize objects on a table whose shapes and relative distances are the synthesizer controls [96]. Another AR instrument, *Augmented Groove* was a predecessor developed in 2000, five years before the *Reactable*. *Augmented Groove* used Computer Vision to track physical cards and mapped their movement to musical parameters: pitch, distortion, amplitude, filter cut-off, filter resonance, delay [130]. The limitations of that system were mobility due to the headset camera being tethered to a computer and audio quality because the synthesis was MIDI based. Most applications of AR musical instruments have been for musical education or rehabilitation through music [60, 65, 108]. An AR Dombyra (Kazakh two string musical instrument) from 2014 used Vuforia on an Android mobile device to display a 3D model of the instrument on top of a predefined image [154]. The Dombyra's sounds were a set of pre-recorded clips triggered only by virtual buttons limiting its playability.

Figure 2.1: Fäscht. *Reactable*. CC by 2.0. 18 October 2008.

2.2 Challenges in Musical Instrument Design for AR

2.2.1 AR Latencies

Audio Processing Latency

Audio latency is not a specific issue to AR, but necessary to consider for musical applications because the threshold of perception for a delayed signal is 10ms before the delay becomes noticeable [133] and can interfere with musical performance. Latency was a challenge in a telematics seminar with Mark Dresser where musicians at UCSD improvised with a group at UCI using bidirectional audio and video transmission. Both parties in those performances had to adjust and recalibrate their playing to account for the delay. It was similar to playing an organ because there is a delay between pressing the keys and sound being generated through the pipes. In AR audio, a delay could be heard as comb filtering in the final mix due to the direct sound and augmented sounds being out of phase. Furthermore, if the delay time is not static (systems with high jitter) then it becomes more difficult for the player to compensate during a performance.

The sources of latency in AR systems include the time it takes to record, digitize, process and play back audio over headphones/speakers. Audio interfaces used in recording and playback add to audio latency in AR systems. Adding resampling on top of the previously mentioned latency sources introduces another source of latency. HRTF databases usually have one set of sampling rate that an XR application has to resample. Audio game engines may bypass spatialization when sample rates do not match. If the audio framework does resample properly it would still add extra latency to the system. Overall, the data collected in this research shows that the required buffer size sets the minimum latency of an AR audio system. Spatial audio processing methods process audio in block sizes and need a minimum amount of time to process the block.

If the AR system is extended to use spatial audio techniques previously described in Section 1.3, the player would need to use headphones for binaural synthesis which would muffle

the direct sound. This is mitigated by either having headphones that do not cover the ear canal (such as open back headphones and bone conduction headphones) or having microphones pass through the direct sound [83] which is similar to a binaural recording method [85]. Other options would require sound source separation techniques [136] in which audio is recorded by microphone(s) and the acoustic sound of the physical musical instrument is extracted for use in AR audio processing.

Motion Latency

Tracking a player's head is required for spatialization in XR. The following latency considerations apply to all head tracked systems. Motion to Sound Latency is the latency of the system from the time the headset is moved to the time the sound object plays with the new position. Huang defines this as the total system latency to head turns. This may be more important in AR systems because there will be lower visual delay since they headset is not rendering virtual objects to show the player the new visual position in the scene. For example, if a user were to look straight at a sound object and then quickly turn their head a quarter rotation (in a system with high motion to sound latency) then the sound will still be perceived as in front even though the object is already physically at the user's side. In VR, the delay in updating the graphics has been known to induce nausea. One method of actually measuring this delay requires a white box approach. Access to the audio framework is required in order to play an initial tone/impulse before the head movement and then play another tone/impulse from the sound object only after the system has received the new head position coordinates. The time between tones is the motion to sound latency.

Response Time

The response time is the time it takes to find a sound object which increases proportionally with the distance of the object and decreases with the size of the sound object [90]. The motion

to sound latency above 73ms is known to increase the response time for short sounds, but is acceptable up to 243ms for continuous sounds [59]. Response time was the concept for the VR game *Fear UnSound* discussed in Section 2.3.1.

2.2.2 AR Reverb

The time of decay of a space's reverb is generally proportional to the ratio of a room's volume over the area as was previously discussed in Section 1.2.3. Matching an audio object's processed sound with the room acoustics becomes an issue in an AR setup. Apart from HRTFs, information of an object's position is obtained from the room reflections. If no reverb is applied then the processed sound may be perceived as unnaturally disjointed from the source. If reverb is applied incorrectly, the it may interfere with the object's perceived distance. The method of creating a "reverberation fingerprint" to generate a binaural room impulse response model (BRIR) has been used to solve this issue [95]. This model includes the calculation of early reflections as a combined set of delayed echo instead of ray-tracing all the reflections for the sake of efficiency. The inputs to the model are the room's volume and the decay rate for all frequencies. The Reference BRIR is measured in a real room beforehand, then adapted to the room in which the AR headset is being used. The adaptation of the live room power spectrum is given by

$$P_{live}(f) = P_{ref}\frac{V_{ref}}{V_{live}} \tag{2.1}$$

where the power spectrum of the reference room, P_{ref}, is scaled by the ratio of the reference room volume, V_{ref}, to the live room volume, V_{live}.

The reference power function was obtained by taking the short-time Fourier transform of the impulse response of the reference room and dividing it by the frequency dependent decay time. By working backwards it was possible to get the estimate of the live room decay. If there are obstacles between the object and the listener then an occlusion model would be used to block

the direct sound but maintain the reverb. In the simplest models this is implemented by a lowpass filter of the direct sound. Chapter 3 further investigates the challenge with spatial reverb for AR.

2.3 Challenges in Musical Instrument Design for VR

The state of the art in VR instruments was surveyed by Serafin et. al. in a review which includes some guiding design principles [140]. The design principles include: Design for Feedback and Mapping, Reduce Latency, Prevent Cybersickness, Make Use of Existing Skills, Consider both Natural and "Magical" Interactions, Consider Display Ergonomics, Create a Sense of Presence, Represent the Player's Body [140]. Many VR instrument designers in the past have used an Oculus Rift [27] VR headset with the Leap Motion [21] sensor. There have been more VR musical instruments released since then with the player being tethered by controllers or sensors to a PC. Only recently have VR platforms, like the Oculus Quest [26], gone completely wireless and include hand tracking [116]. This new wireless technology will push the boundaries of VR musical instruments like the VR Singing Kite discussed in Section 3.4. Challenges facing musical instrument design for VR are framed by analyzing the VR game *Fear UnSound* [74].

2.3.1 Fear UnSound

Fear UnSound is an Android VR game developed for the purpose of using spatial audio to invoke fear and test the limits of a player's ability to withstand listening to sounds. In this game, the player localizes sound objects placed anywhere in 360^{o} space. To silence the sounds and move on to the next level the player must rotate their head to localize the direction of the virtual sound source and then hold their gaze until the sound is eliminated. The basis for the sound design is these seven characteristics of annoyance [101]: surprising sounds, interfering sounds, inappropriately timed sounds, intermittent sounds, high reverb sounds, loud sounds, high frequency energy sounds (chainsaw, heart beat, smoke alarm, snarling, screaming, howling

wolves, phone off-the-hook tone, zombies, and explosions). Beyond annoyance, the game, *Silent Hill*, demonstrated how destabilizing the player's sense of control invoked fear in the listener [63]. Zombie's sounds are usually produced to reflect a human voice which has been altered to be unintelligible animal growling [126]. In *Fear UnSound*, the destabilization comes from objects appearing when they are localized against a pitch black black screen.

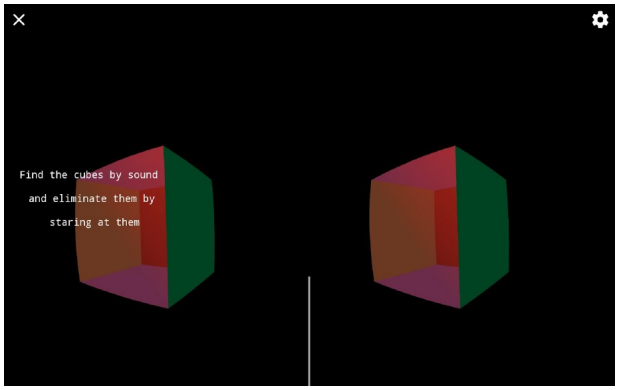


Figure 2.2: Screenshot of *Fear UnSound* VR Audio Game played on Android Daydream.

Fear UnSound was demonstrated at the Che Cafe on April 14, 2017 as part of performances and installations exploring sonic violence [32]. During the event, players had trouble with localizing the phone-off-the-hook sound because it was a pure tone that provided less information for binaural hearing than more broadband sounds. The audio object's (Figure 2.2) starting positions were selected randomly each time changing the player's response times. Similarly, there was an MR game developed by Moustakas et al., which employs head tracking and multiplayer functionality with a similar concept to hunt and destroy objects based solely on their sounds [109]. By including these additional functionalities in the hunting game, players were able to observe an increased sense of immersion from the AR audio system.

2.3.2 Interaction with the Virtual World

Head Tracking

In an AR environment, the objects are tied to the physical room through CV. In contrast, a VR world needs to have the position of the player and all the objects constantly updated. This constraint begins to set limits on how many objects can be processed simultaneously within the latency tolerance of the processor. Implementing head rotation in *Fear UnSound* was simple because Android's VR framework takes the accelerometer data from the mobile device's sensors and provides the head rotation information. The downside is that there is no head position tracking available yet for Android mobile devices like there is for systems like HTC Vibe. Therefore, being constrained to a setup is a requirement for head position tracking. The use of GPS for head position tracking would be an option but the position accuracy is more suitable for massive multiplayer games and less so with small movements. To create a 3D visual effect, a stereoscopic display must be implemented on the headset screen. This means that there is a need to draw twice the number of frames per second because the right and left sides of the device are unique for each eye. This could add to the total system latency compared with AR.

The minimum audible movement angle (MAMA) is defined as the smallest angle a sound object can move and still be perceived as stationary [90]. Ideally, the head rotation resolution is as fine as the MAMA.

VR Controllers

In AR, the human body is expected to be used as the interaction with the room, but for a completely simulated world the controls also have to be created. The HTC Vibe, for example, has two wireless hand controllers as the means to interact with the virtual world. For increased mobility, headsets can make use of near field detection sensors like Leap Motion to detect the hands without the use of controllers. The Leap Motion controller was used in a 2013 performance

by the author involving a set of five Cumbia percussion loops each controlled by a digit of the performer's left hand and five voice samples controlled by the digits of the right hand [110]. The vertical distance from the fingers to the sensor was mapped to volume control of the drum loops and pitch shifting of the voice samples. The resolution of the sensor was good enough to track each finger individually which allowed for precision control of the playback, but the tactile feedback of a physical button or touchscreen was missing. Haptics is the use of physical feedback through motions or sensation and commonly implemented by vibrating a controller. This works fairly well for musical VR games like Beat Saber [7] and has been studied for musical instruments [55, 56], but may be restrictive compared to AR musical instruments.

2.4 Augmented Reality System Example

This section provides an example of a 6DoF AR audio system designed by the author for use with a percussion musical instrument to diagram the technical requirements of spatializing the audio effects that are applied to a live sound object in real time.

2.4.1 Computer Vision

The use of CV in current AR technology affects the design of AR musical instruments[1]. Visual markers have been used to detect and track the neck of a guitar [108]. AR applications for mobile devices, such as facial recognition, rely heavily upon cloud-based computing to move away from markers. Mobile applications like Snapchat [31] and Facebook [11] transmit a reduced parameter set from the images captured by the camera to the cloud for processing using artificial neural networks (NN). There are open source tools [9, 38] available for running NN recognition. For mobile devices there is a neural processing engine SDK which optimizes the execution of Deep Neural Network (DNN) programs on mobile devices taking advantage of their CPU, GPU

[1] For more information on visual perceptual issues in AR refer to [100].

and DSP [132]. There is an expectation for dedicated hardware for DNN which will move the computation from the cloud to the mobile device and therefore reduce the latency for audio/visual object recognition.

2.4.2 System Design

The system architecture was designed for one electroacoustic drum performer and up to N audience members, as shown in Figure 2.3. The signal flow, described in Figure 2.4, shows the audio from the instrument in the center going into the mobile device of the performer to be processed and then transmitted wireless to the mobile devices of the audience. As the performer strikes the drum, the signal and position of the strike is passed through the Audio Effects Processing block and then to the 3D Audio Player (Section 2.4.4) which uses spatial audio processing so that the output is perceived to be originating from the same location as the mallet strike.

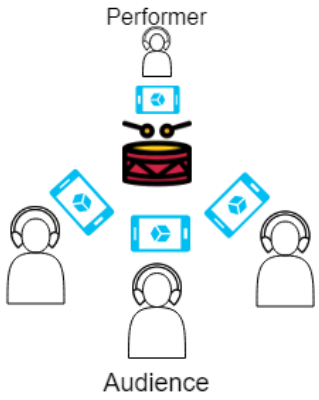


Figure 2.3: Representation of an AR performance in which the performer and audience are wearing headphones and have mobile devices tracking the instrument.

Physical Instrument

The musical instrument at the core of this system, shown at the center of Figure 2.4, is an electroacoustic, two-tongued log drum of Aztec origin called the teponaztli. While any

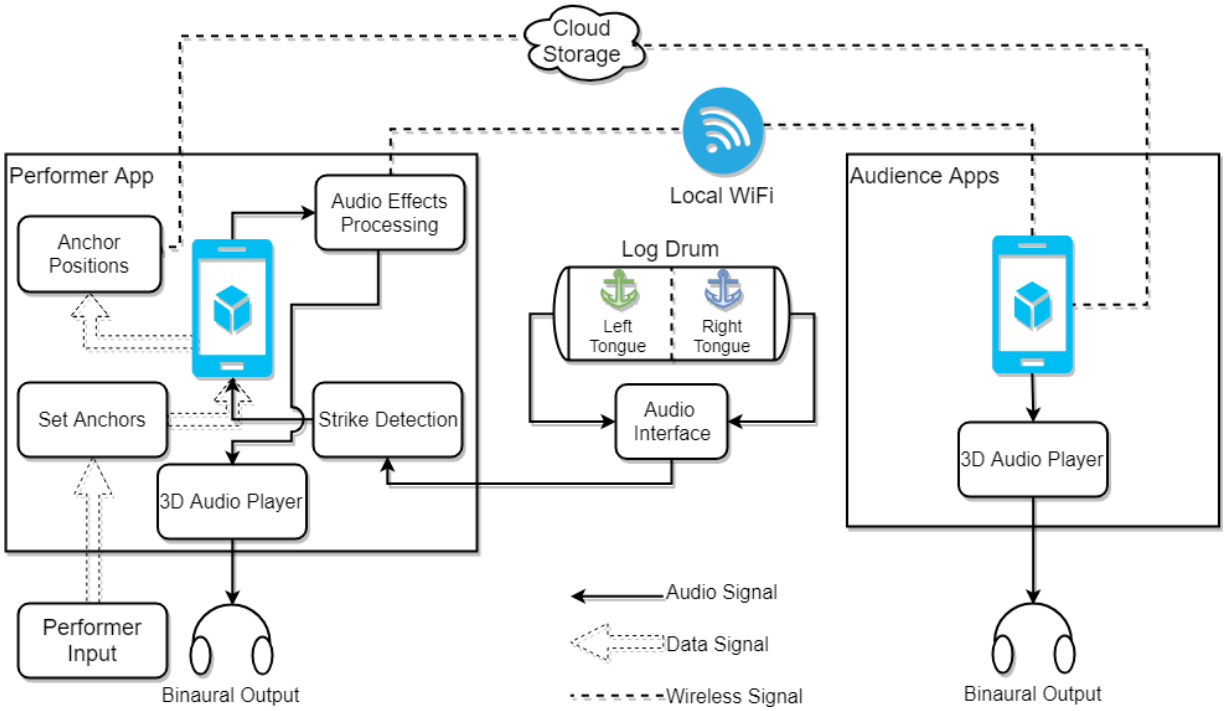


Figure 2.4: System overview diagram showing the connections between the teponaztli log drum (center), the performer's (left), and the audience's (right) mobile devices.

electroacoustic musical instrument could have been used in this type of system, the teponaztli was chosen because signal processing methods were available to automatically determine which of the two tongues was struck [75]. This simplified the work of matching the event of the performer striking the drum with the sound produced by the drum without needing to add CV tracking or sound source separation. A mobile device was used to augment the sound of the drum and add spatial audio.

Mobile AR Platform

Fear UnSound was transformed into an AR game [12] with Unity3D [39] and Vuforia [42]. The AR solution chosen for this system, however, was Google's ARCore [5] due to the ability to create a virtual representation of the room in the form of a mesh using a single camera from the mobile device. In this manner, the system enables a new position-based design for AR musical

instruments where the performer can choose points in the real world for sound augmentation.

2.4.3 User Interaction

The goal of this system was to abstract as much of the human-computer interaction as possible in order to make playing the AR instrument feel natural. Therefore after a short setup, described in the following sections, the performer no longer needs to interface with the mobile device in order to play the instrument.

Setting Anchors

The first step is to create a virtual mesh representation of the room (including the instrument) by scanning the room with the camera on the mobile device. Then, the performer will place two anchors, one over each tongue of the log drum, by tapping the screen at the corresponding points. The position of these anchors is transmitted to the cloud using the ARCore framework so that those positions can be shared with the audience members' mobile devices. This way of sharing locations is the key to everyone receiving personalized position-based (6DoF) binaural audio.

Setting AR Audio Parameters

Next, the performer has a choice of what audio effects are used for each anchor point. This choice can be static (no change after setup), dynamic (during the performance), or programmatic (script to change values based on audio-visual feedback during the performance). For demonstration purposes, the Audio Effects Processing block of Figure 2.4 consists of a delay effect which gives the log drum a sustained sound effect.

2.4.4 Spatial Audio Block

Spatialization for an AR musical instrument required adding both direction and distance cues to the binaural headphone playback in order to accurately place synthetic sounds in a real room. This section describes the 3D Audio Player block of the system.

3D Audio Player

The 3D Audio Player block spatializes the post-effects processed audio of the instrument. Unlike the AR examples described in Section 2.1.2, this system places the computer generated sounds at the same perceived location in space as the instrument. The performer and audience need to perceive the new synthesized sounds as coming from the same location in the room in order to be fully immersed in the realism of the AR musical instrument [90]. Spatialization in this system takes the mono audio signal from each anchor point and applies two rendering paths; one direct sound path (object rendering), and one room acoustics path (Ambisonics) as shown in Figure 2.5. Both paths utilize Unity3D platform's built in object spatializer framework.

Convolution reverb takes a room, impulse responses and combines them with sounds to make it seem like the sounds originate in those rooms. In this system, the soundfield audio rendering path was used to give the room reverb directivity. The reverb model was implemented using Ambisonics to encode the reflections in 360^{o} so that when the listener moves their head the sense of direction is maintained.

AR Musical Instrument Latency

Unless the percussion strike is known a priori, there will always be a delay between the strike and the spatialized audio heard on the headphones. The amount of delay depends on the latency of the 3D Audio Player. The use of bone conduction headphones mitigates this latency in the system and has been shown to reproduce spatialized audio [102]. Since the ears of the listeners are not covered, the initial percussive strike of the instrument will also be heard naturally.

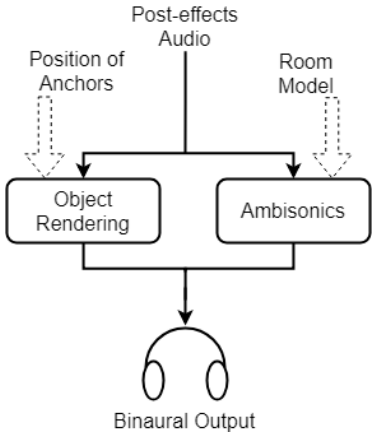


Figure 2.5: The 3D Audio Player takes in the post-processed audio from the musical instrument and applies two paths for spatialization; a direct object rendering path and a room reverb path.

As previously mentioned, open backed headphones or passthrough audio headphones would also work. The augmented audio that follows is a feature of the instrument (e.g. a gong strike has an immediate attack envelope and then a resonance). The attack is the real world acoustic sound of the instrument and the sustain is the computer generated sound from the 3D Audio Player.

2.4.5 Technical Constraints

The mobile devices in this system should have the same orientation as the listeners' heads, achieved by pointing the camera in the direction where the listeners' heads are facing. The direction of sound is determined by the anchored instrument. The 3D Audio Player directs the sound using the mobile device's coordinates as a reference point. For example, if the anchors are shown on the right side of the camera view, then the audio will be perceived as coming from the right side of the camera. Therefore, if the performer or audience chooses to look at the instrument directly instead of through the camera view, then their orientation should remain aligned with the camera as much as possible. The system presented here makes use of ARCore supported devices [6] because these devices have already saturated a majority of the market. One option is to place the mobile devices in an AR HMD. Another option is to use standalone AR platforms such as Google Glass [14] or Microsoft HoloLens [24].

The reverb should match the acoustic properties of the room, but the system has only been tested using a standard room model. This system can be improved by using ARCore's mesh representation of the room to adjust the reverb model based on the room size. The following Chapters present three XR musical instruments developed to investigate spatial audio reverb and to create an interactive performance with the use of position-based audio effects.

Chapter 3

Spatial Reverb for Extended Reality Musical Instruments

This chapter reviews the creation of a hybrid spatial reverb system, which consists of synthesized early reflections combined with measured late reflections for real time XR musical instruments. The hybrid system was created after comparing Directional Room Impulse Response (DRIR) convolution reverb with synthetic spatial reverb (IEM RoomEncoder). Sounds processed by DRIR convolution reverb were perceived by the author to be more natural but less localizable than sounds processed by synthetic spatial reverb. Therefore, combining the early reflections from the synthetic reverb with the later part of the DRIR resulted in a hybrid system which was potentially both natural sounding and localizable. This possibility was explored via a pilot study designed to rank listeners' preference in perceived naturalness of seven spatial reverb (including variations of the hybrid spatial reverb) for use with XR musical instruments. AR electric drumset, AR electric guitar, and VR Singing Kite systems were developed for the subjective listening tests. The measured DRIRs were derived from swept sine Ambisonic microphone recordings, and the synthesized DRIRs were generated from two spatial reverb models [10, 18]. The measured DRIRs had a lower Direct to Reverberant Ratio (DRR) which was suspected to be the main factor in

reduced localization when they were used for convolution reverb.

3.1 Spatial Reverb

The use of spatial reverb for VR environments [28,36] is well established for game engines to simulate multi-directional room reflections of sound sources which makes the experience more immersive. This process is done for audio reproduction through headphones and by tracking the user's movements with a Head Mounted Display (HMD). The tracking data can have either three degrees of freedom (3DoF) where the user can rotate their head but is stationary or six degrees of freedom (6DoF) where the user's position is tracked. In VR, the audio designer tunes the reverb to match the virtual environment. Specifically for AR audio (mixing computer processed sound with a live acoustic source [98]), the listener has a reference of how the reverb actually sounds making it a challenge to maintain a realistic effect. AR audio instruments differ from electroacoustic instruments in that the sound from AR musical instruments is processed in order to make the sounds feel like they are coming from a physical location in the same room. In contrast, a listener would expect the sound of an electric guitar to come from the guitar and the amplifier. Spatial reverb for XR musical instruments is different because it dynamically takes into account the listener's and sound source's positions.

In this research, spatial reverbs (Table 3.2) were designed for use with XR musical instruments (Sections 3.3.1 AR electric drumset, Section 3.3.2 AR electric guitar, and Section 3.4 VR Singing Kite) to investigate how well DRIRs can tailor a convolutional reverb to the listener's performance space. The AR electric drumset had fixed sound sources and a fixed listener position (the drummer) in order to evaluate DRIRs specifically in a 3DoF scenario. The AR electric guitar was utilized to evaluate a fixed listening position (the drummer) and a moving source (the guitar). The VR Singing Kite was used to explore the reverb when both the sound source and the listener were moving in a 6DoF scenario.

3.2 Directional Room Impulse Responses

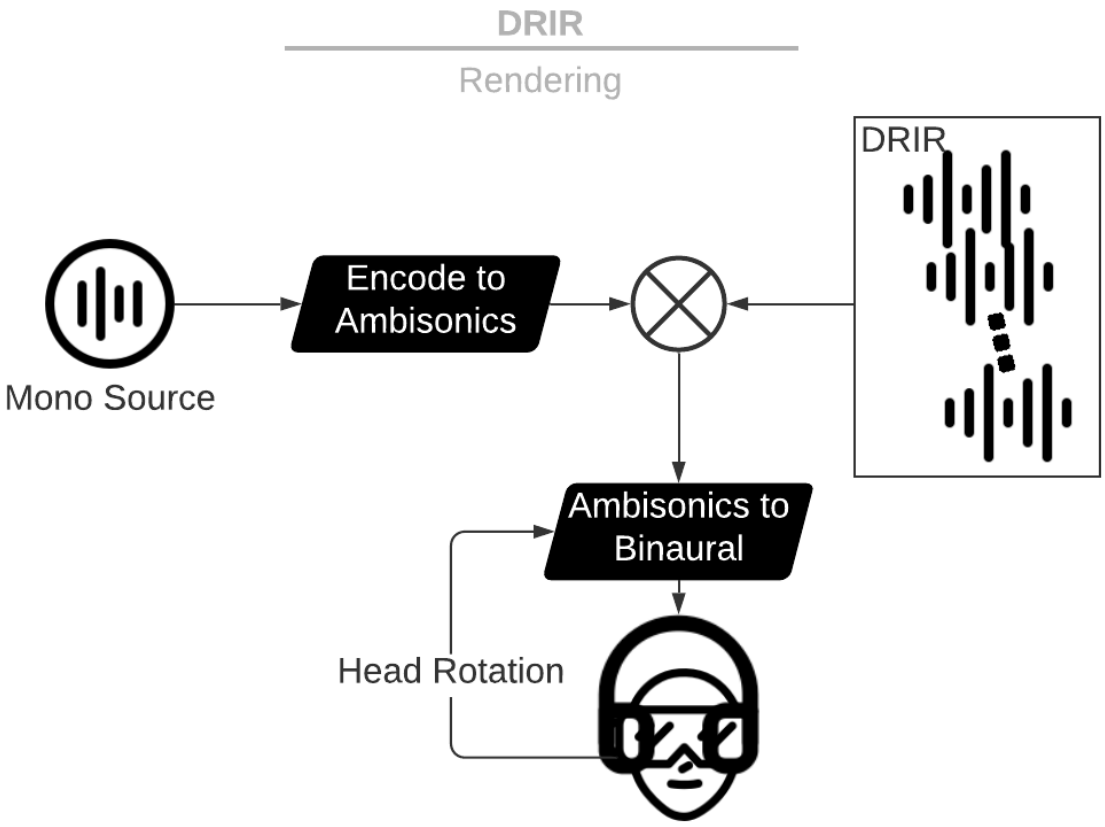


Figure 3.1: A high level diagram of convolution spatial reverb performed by convolving the Ambisonics encoded mono source with the Ambisonics encoded DRIR before being sent for 3DoF binauralization.

Reverb models from balloon pops [48] have been created for spatialization over loudspeakers to simulate vocalists singing in the acoustics of the measured room response. Rendering of spatial impulse responses has also been studied for loudspeaker reproduction by Merimaa et al [106, 131]. Ambisonics has been proposed as the convention for DRIRs in Spatially Oriented Format for Acoustics. Ambisonics DRIRs can be synthesized [115, 122] as well as measured [52]. The DRIRs can be convolved with a set of HRTFs from the sound source direction to create BRIRs for that direction. If these BRIRs are then convolved with the mono sound source, they impart the

reverb of the room onto the source, but only for that direction. Figure 3.1 shows a simple method to create the natural reverb effect for a mono sound source that can preserve the localization for all directions. The mono source is first encoded into an Ambisonics signal and then convolved with the DRIRs. This is done by first selecting spherical harmonic coefficients (Equation 1.10) for the desired azimuth plus elevation of the sound source and then mixing with the mono sound source. The result is an Ambisonics signal of the sound source which when convolved with the DRIRs now includes the room reverb and maintains the perceived location of the sound source. Convolving all the DRIRs directly with a mono source is like having the source arrive from every direction whereas convolving the DRIRs with the source converted to the spatial domain allows for the source to keep its direction with respect to the listener.

3.2.1 Measuring DRIRs

Measuring DRIRs begins by placing the source of the impulse at one location and a microphone or microphone array at another location. In 6DoF AR audio the listener can move anywhere in the room which makes the reverb position dependent. The challenge is to create a set of measurements that can adapt to any location in the room. The synthetic spatial reverb used in this study are position dependent as they take source and listener positions relative to a virtual room. This section steps through the measurement of the DRIRs for use in spatial audio reverb. The measurements were performed using Farina's a swept sine technique [72]. A sine sweep with

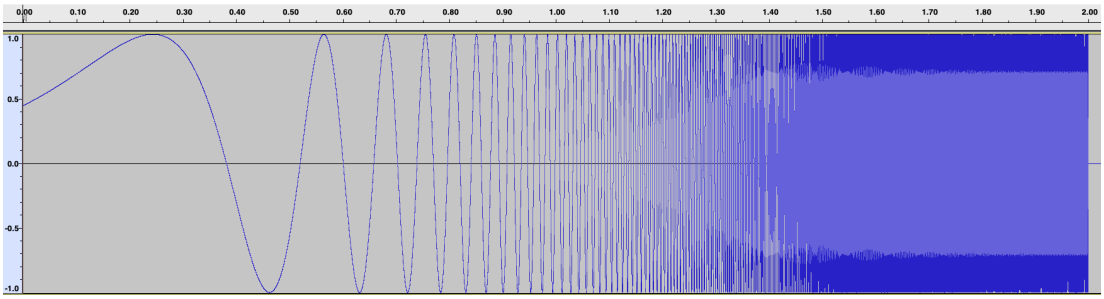


Figure 3.2: Waveform of the time stretched pulse used as the excitation energy for impulse response measurements. The x-axis is time from 0 to 2 seconds and the y-axis is amplitude.

frequency increasing exponentially, Time Stretched Pulse (TPS), shown in Figure 3.2, is played from a loudspeaker[1] and recorded by Ambisonics microphones. The IR Measurement MaxMSP patch (Figure 3.3) was used for convenience to synchronize the playback and record [20]. It was modified to simultaneously record 25 channels to support up to 4^{th} order microphones. The DRIRs were obtained by deconvolving the TSP from each microphone channel in Matlab because the IR Measurement MaxMSP patch only deconvolves one channel at a time. They were further processed by normalizing and trimming the starting silence[2].

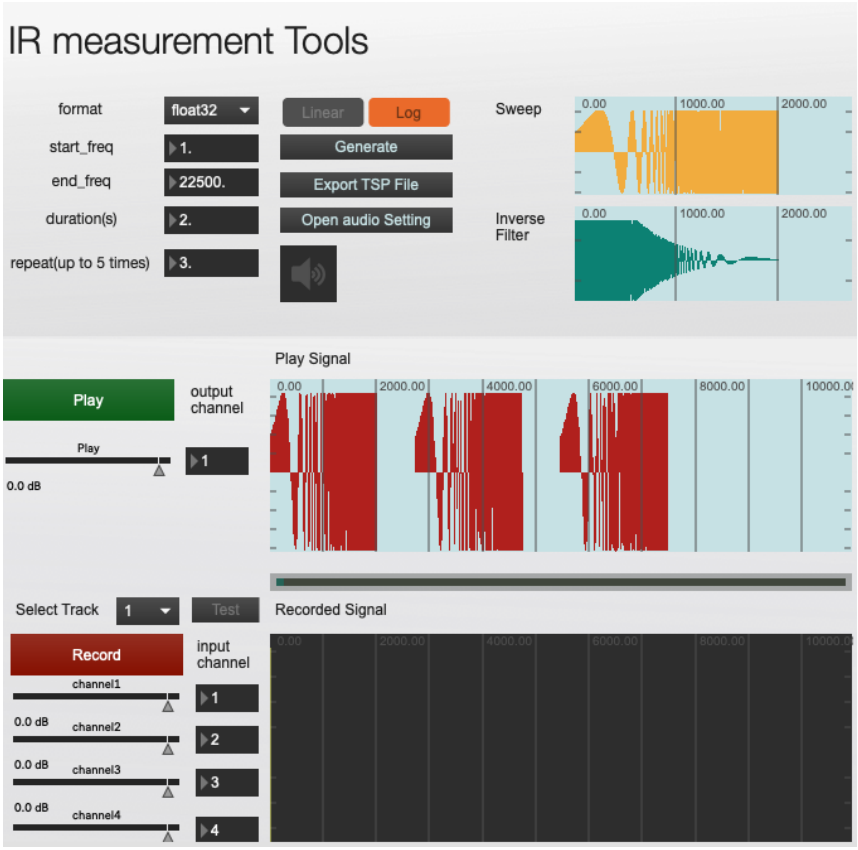


Figure 3.3: Screenshot of IR Measurement Tool MaxMSP Patch. The excitation energy comes from a time stretched pulse is shown in yellow at the top right. The number of times the signal is played can be varied to average out effects of background noise. Four channel recording is shown at the bottom. Only one track at a time can be processed into an impulse response.

[1]Loudspeaker equalization was not performed.
[2]The DRIRs are available at [113].

The two microphones[3] used were the Sennheiser AMBEO VR Mic [4] and the Zylia ZM-1 [46]. AMBEO VR Mic (Figure 3.4) is a first order Ambisonics microphone with an accompanying A-B Format converter audio plugin [3] shown in Figure 3.6 which was used to convert the four channel microphone impulse responses to B format first order Ambisonics. The ZM-1 is a third order Ambisonics (Figure 3.5) with accompanying software for recording and converting to Ambisonics.

Figure 3.4: Sennheiser AMBEO VR Mic. *Microphone 3D AUDIO*. [4].

When the mono sound source is encoded into Ambisonic format it can be represented as an SxN matrix of PCM values where S are the audio frame samples and N are the Ambisonic channels. Convolving directly with the DRIR, an MxN signal where M is the DRIR filter length, applies the spatial reverb from the measured DRIR to the source sample. The benefit of processing the source in this manner is that it can be easily rotated by a matrix multiplication. This affords a fast implementation of 3DoF rendering of the source (head rotation fed back into the binauralization block of Figure 3.1).

[3]The choice in Ambisonic microphones for the informal study was driven by their availability.

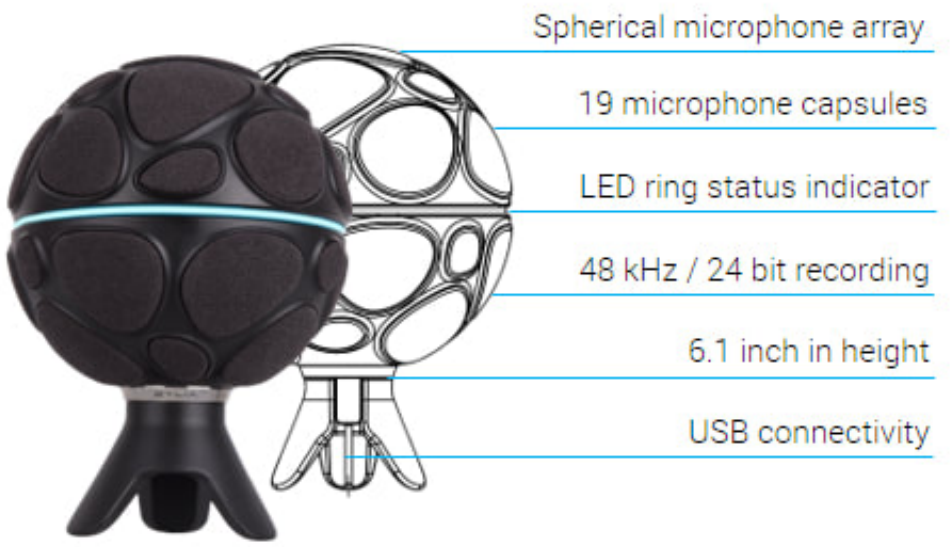


Figure 3.5: Zylia ZM-1 Microphone [146]. The Vocalist Studio.

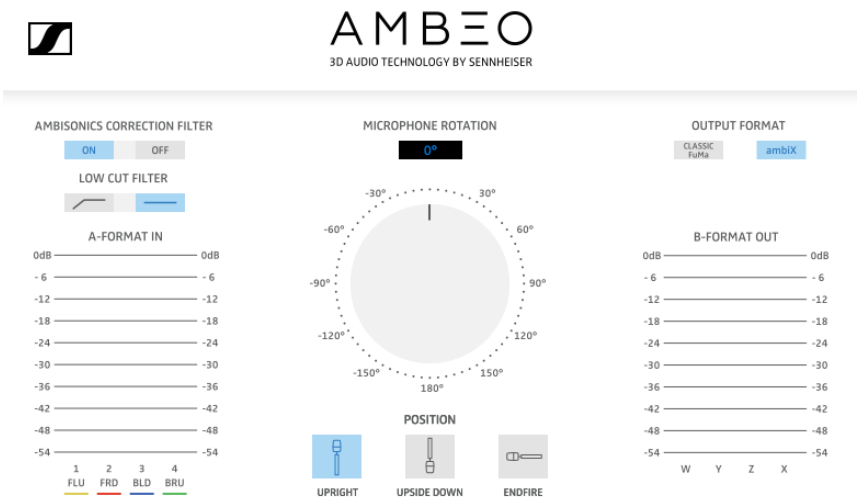


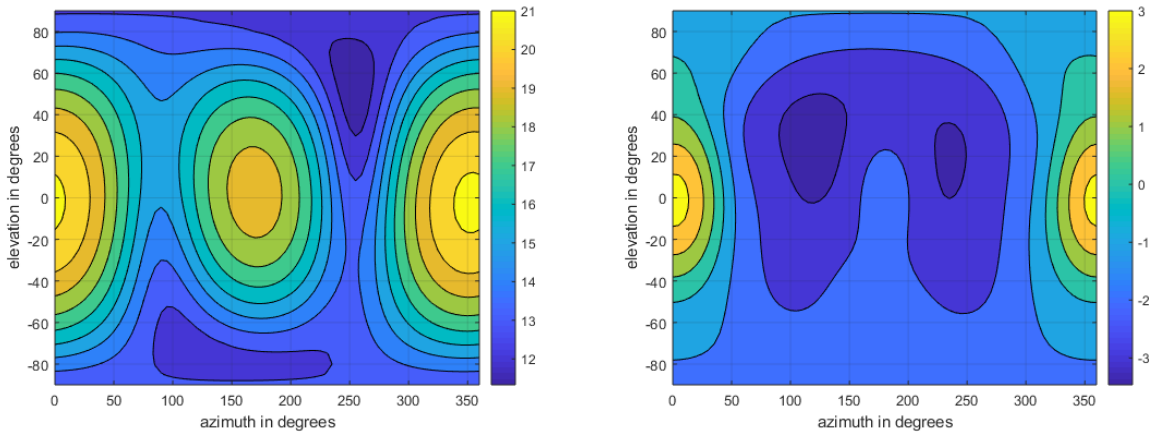
Figure 3.6: Screenshot of the Ambeo A-B Format Converter.

The third order Ambisonic DRIRs were recorded using Zylia's ZM-1 microphone. The raw recordings of the sine sweep contains 19 PCM tracks, one track for each of the digital MEMS sensors on the microphone. Deconvolution with the sine sweep was performed in Matlab to obtain

the impulse responses which were then converted into Ambisonics using Zylia's Ambisonic Converter (Figure 3.6) [45].

3.2.2 Objective Analysis of DRIRs

Directivity plots can be generated from the Ambisonic signal to display energy intensity over the unit sphere. These plots relate to the perceived localization. For example, the directivity plot of mono source at 0^o azimuth and 0^o elevation is shown in Figure 3.7 for 1^{st} and 3^{rd} order Ambisonics microphones. Note that the directivity plots do not contain the effect of binauralization which may affect localization on a per listener basis.



(a) (b)

Figure 3.7: DRIR energy plots for a source at 0 degrees azimuth and 0 degrees elevation recorded with (a) Ambeo VR Mic (1st order Ambisonics) and (b) Zylia ZM-1 (3rd order Ambisonics).

Direct to Reverberant Ratio (DRR) is given for sampled (discrete) impulse response $h(k)$ of length K by

$$DRR = 10\log_{10}\left(\frac{\sum_{k=0}^{f_s t_d} h^2(k)}{\sum_{k=f_s t_d}^{K} h^2(k)}\right) \tag{3.1}$$

where f_s is the sampling rate and t_d is the first 5ms of the response. DRR is the ratio of the sound's direct path to its reflections. Equation 3.1 describes an approach to measure the DRR of the DRIRs is to measure the energy of the direct sound by taking the ratio of the first 5ms of the response and the energy of all reverberation [70]. This approach can be taken either on the 0^{th} Ambisonic channel to get a sense of the omnidirectional DRR or on the final binaural channels to include effects of the HRTFs in the DRR.

3.3 XR Musical Instruments for Evaluation of DRIR

For subjective comparison of measured DRIR convolution reverb with synthetic spatial reverb three XR musical instruments were designed for use over headphones. Research has shown the importance that HRTFs play in the perceived spatialization because of the uniqueness of the human pinna [67]. The binauralization of the reverbs was performed by the IEM Binaural Decoder plugin [15] which uses HRTFs from the Neumann KU 100 [25].

3.3.1 AR Electric Drumset

An AR electric drumset is perfect a 3DoF scenario where the listener is not moving around the room. This simplifies the implementation because the DRIRs can be compared without interpolation since there is only one measurement position where the drummer is located. The sound sources are also known and are static. For the pilot test in Section 3.5 each drum sample was processed by all seven spatial reverb conditions with only one condition (selected by the user interface) at a time passing to the binaural output stage.

3.3.2 AR Electric Guitar

The AR electric guitar was designed in a way that the amplified sound is perceived to be coming from the same location as the electric guitar. An AR electric guitar allows the evaluation

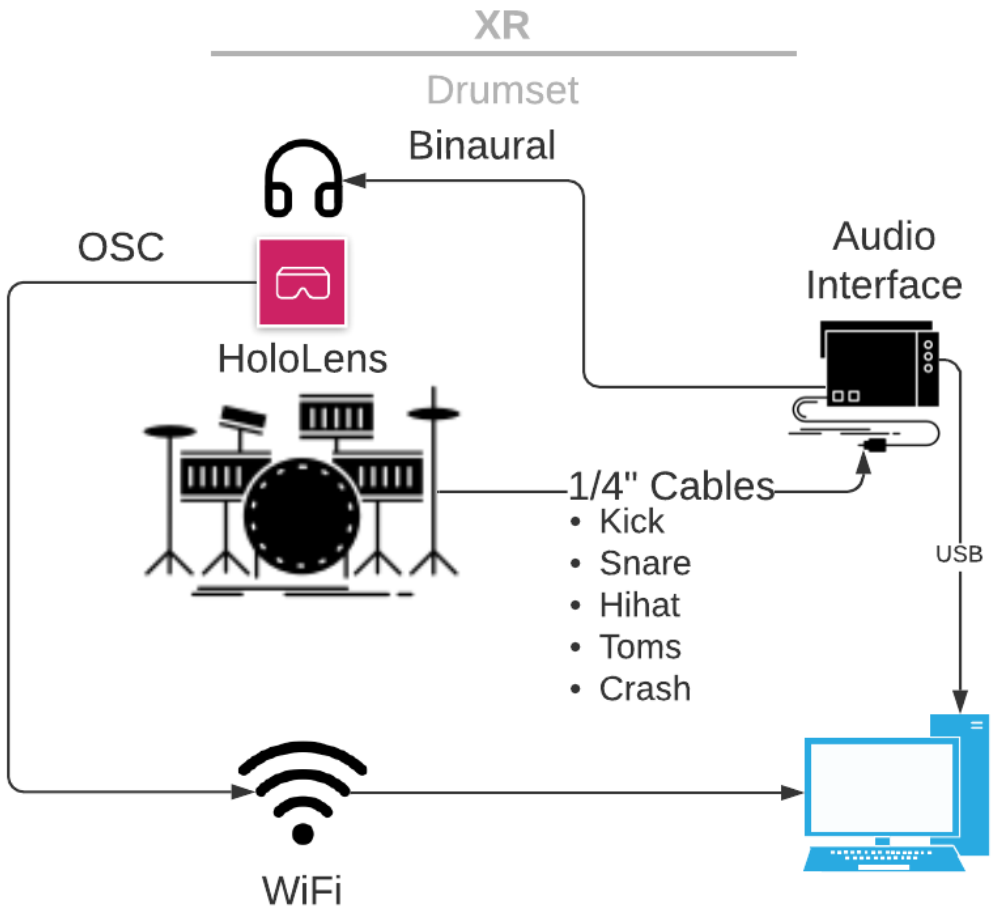


Figure 3.8: Diagram of the setup for the AR drumset.

in 6DoF because the guitar player is tracked using an HMD and can move around the room. A challenge compared with the AR electric drumset is that the samples are not pre-convolved with the DRIRs. Since this step has to be done in real-time because the guitar signal is not known a priori, a low-latency convolution method [76] is necessary to prevent a negative impact on the immersiveness in the augmented sound.

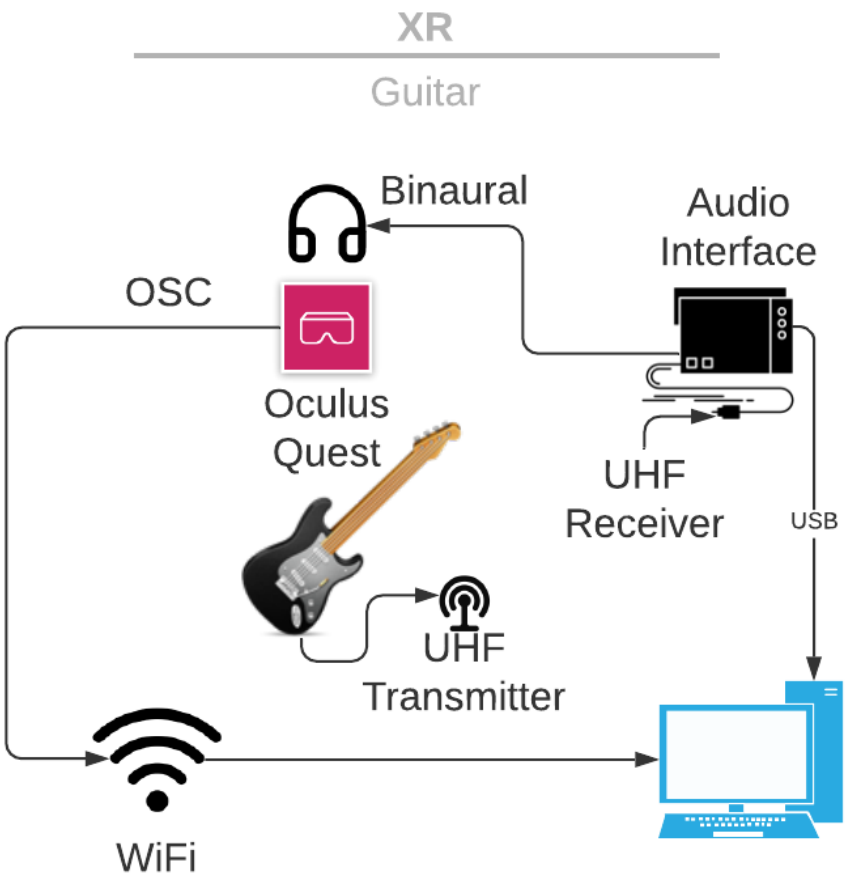


Figure 3.9: Diagram of the setup for the AR guitar.

3.3.3 Matching the Rooms' Coordinate Systems

To achieve the 6DoF AR effect the coordinates of the virtual world must be correctly overlaid onto the real world. This process should be the same regardless of what HMD (VR or AR) the listener is using. The locations of the measured DRIRs must be correctly placed in the virtual world.

The first set is to match or convert the coordinate system of the virtual world and the data of the real world which is obtained by the 6DoF HMD. Then the front directions must be aligned which can be tricky because 6DoF HMDs such as the Oculus Quest [26] may choose

front to be whichever direction the HMD is facing when the program is started. The Oculus Quest also has the concept of a "guardian" or boundary that is tied to the real world coordinate system. Therefore, if the program aligns itself to that boundary then the front directions of the virtual world and the real world can be matched.

Finally the system must be scaled. When creating virtual world scenes, there is a scaling parameter so the room size must be scaled correctly such that distances in the virtual world match those of the real world. This matching process was done for the test room/rehearsal space (Figure 3.17) and also for the concert venue (Figure 5.2).

3.4 Exploring the Room

Although the AR electric drumset and guitar are good means to evaluate the naturalness of the AR audio effects, these systems required multiple testers to investigate the sound at different locations. To explore the AR acoustics of the room a virtual reality application for the Oculus Quest was developed where the listener can control the position of a virtual sound source in a virtual room and also walk around independently.

Exploration of the different spatial reverb began by panning a white-noise audio object around a virtual room and listening over headphones. To turn this exercise into something musical, inspiration came from a composition course in which Chinary Ung's presented recordings from a Cambodian Singing Kite that produces a broadband sound from the reed blade at the front of the kite (Figure 3.10a) [144].

The VR Singing Kite (Figure 3.10b) was developed in Unity for the Oculus Quest. The position of the kite was sent from the Oculus Quest via OSC to a computer running MaxMSP for the audio processing. To operate the kite, the user controls the direction of the kite by grabbing a cylindrical handle with the Quest's right controller. The distance between the handle the kite is increased by pressing the trigger of the left controller mimicking the action of giving more slack

(a) (b)

Figure 3.10: (a) Cambodian Singing Kite. Prof. Chinary Ung on the left and YOS Chandara on the right. (b) VR Singing Kite app and the author controlling the kite using an Oculus Quest.

on the kite string. A complete description of the VR Singing Kite as a musical instrument is given in Chapter 5.

3.4.1 DRR and Clarity of the AR Electric Drumset

The importance of DRR for sound localization and distance has been studied with different sound sources [120]. Directivity plots in the previous sections show variations between the spatial reverbs, but the way spatial reverb affects DRR may play a larger role in both naturalness and localization. Reverberation clarity is defined in [70] as the Direct-to-Reverberant Ratio given in 3.1, but with $t_d = 50$ms for music and $t_d = 80$ms for speech. Clarity is important in distinguishing musical notes and can be reduced with an increase in reverb [81]. Table 3.1 lists DRR and clarity values of DRIRs for each microphone and for each source position to be discussed later in Section 3.5.1.

Table 3.1: DRR and Clarity for Measured DRIRs

DRIR Microphone	Position of Impulse	DRR [dB]	Clarity [dB]
ZM-1	snare	0.76	11.67
AMBEO VR Mic	snare	2.35	14.40
ZM-1	hihat	-0.16	11.41
AMBEO VR Mic	hihat	1.76	14.78
ZM-1	crash	-0.31	12.68
AMBEO VR Mic	crash	0.95	15.81
ZM-1	floor tom	-1.08	11.84
AMBEO VR Mic	floor tom	2.63	16.21
ZM-1	kick	-4.40	12.75
AMBEO VR Mic	kick	-3.57	11.02

3.4.2 Synthetic DRIRs

EVERTims was used to create synthetic DRIRs of a virtual room. A basic model of a spare room in the author's house was created using Blender. Figure 3.11 is a Blender visualization of the room with EVERTims raytracing model which was used to create second order DRIRs (Figure 3.12) at different source positions. The average[4] DRR of these synthetic DRIRs is higher than that of the measured DRIRs (Figure 3.18).

[4] Averaging the DRR of the binaural output of the three XR musical instruments.

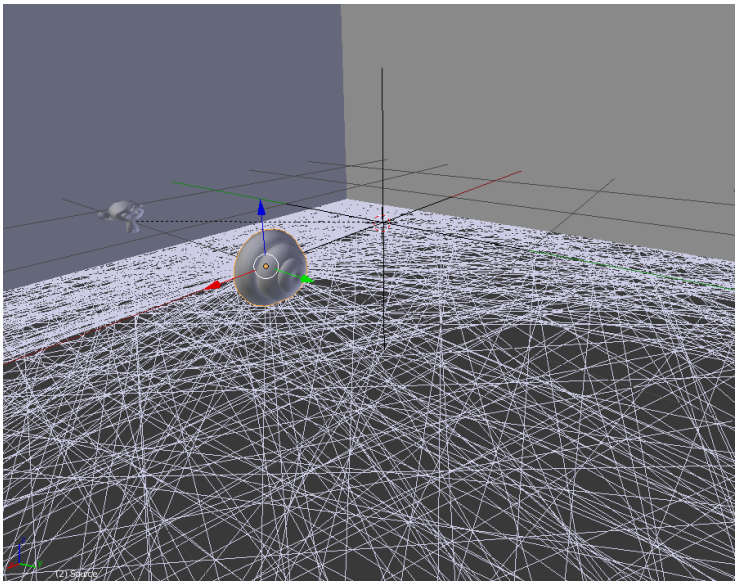


Figure 3.11: Raytracing visualization of the listener and source positions using Blender and EVERTims. This was the position used for creating the synthetic DRIRs (shown in Figure 3.12) for the guitar as the source.

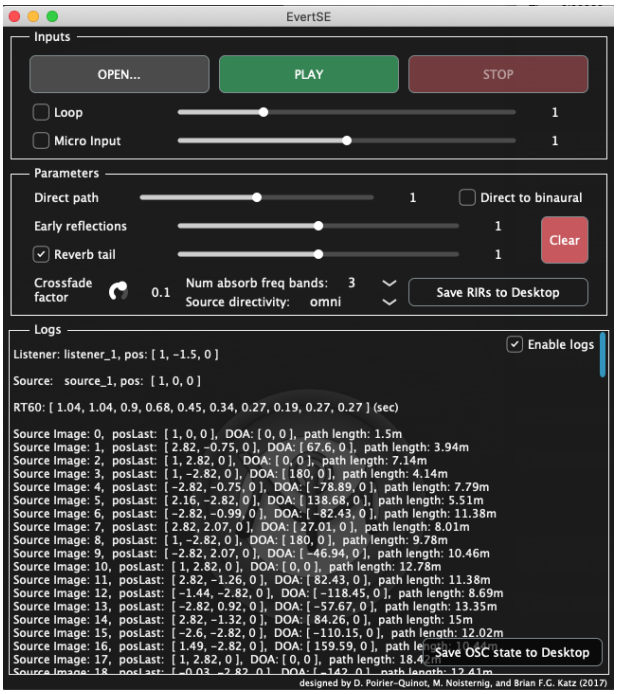


Figure 3.12: EvertSE screenshot showing settings for the generation of the synthetic 2nd order Ambisonic DRIRs used for the guitar position visualized in Figure 3.11.

3.5 Perceptual Comparison of Spatial Reverb

This section compares seven approaches to creating spatial reverb (Table 3.2). A 6*x*6m room (Figure 3.17) was used for all microphone measurements, room simulations, and listening tests in this evaluation. The shoebox reverb was evaluated in its own condition, SHO, and was also mixed with DRIRs conditions ZM, AMB, and EVR to create three hybrid spatial reverb conditions: H-ZM, H-AMB, and H-EVR. When SHO was mixed with the other conditions, the IEM RoomEncoder plugin's "Number of Reflections" parameter was reduced to permit only early reflections[5]. Two of the hybrid conditions used measured DRIRs and one used a synthetic DRIR created by the EVERTims framework. All the conditions were evaluated using the IEM BinauralDecoder plugin and the same pair of headphones[6].

3.5.1 Informal Preference Test

An informal study was performed by three subjects[7] for three XR musical instrument scenarios (Table 3.3) to evaluate spatial reverbs (Table 3.2). The subjects were asked to evaluate the naturalness[8] of the instruments beginning with the following prompt:

> *You will be presented with two version of spatial reverb for each XR musical instrument.*
>
> *Please select the one which makes the instrument sound more realistic for the room in which you are playing the instrument.*

The UI, shown in Figure 3.13, consisted of four buttons, "A", "B", "Next" and "Start". The subjects controlled the UI through a wireless mouse. The seven conditions in Table 3.2 were randomly paired in twenty one unique combinations each time the "Start" button was pressed. The test assigned "A" and "B" buttons to a unique pair of spatial audio reverb from the twenty

[5] Approximately 35ms ($\frac{12m}{343m/s}$) to capture one reflection when the listener and sound source are on opposite walls.
[6] Sennheiser HD 280 PRO.
[7] The tests were repeated three times by each subject.
[8] Using perceived naturalness as is commonly presented in 3D audio listening experiments.

Table 3.2: Showing the seven conditions for creating spatial reverb that were used in the two-choice informal study. The first condition, SHO, is the fully synthetic shoebox model spatial reverb. The DRIR measurements obtained with two Ambisonic microphones, Zylia ZM-1 (ZM-1) and Sennheiser Ambeo VR Mic (AmbeoVR), are used in convolution reverb for conditions ZM and AMB. EVR is a simulated DRIR using EVERTims software. The three hybrid reverbs (H-ZM, H-AMB and H-EVR) are a combination of the DRIRs and SHO. Note that where there is no microphone specified, the condition is completely synthetic and that where there is no software specified, the condition uses convolution reverb shown in Figure 3.1.

Label	**Condition**	**Microphone**	**Spatial Audio Software**
SHO	Shoebox		IEM RoomEncoder [19]
ZM	3rdOrder Ambisonics Measured DRIR	ZM-1	
H-ZM	ZM+SHO	ZM-1	IEM RoomEncoder
AMB	1stOrder Ambisonics Measured DRIR	AmbeoVR	
H-AMB	AMB+SHO	AmbeoVR	IEM RoomEncoder
EVR	2ndOrder Synthetic DRIR		EVERTims [10]
H-EVR	EVR+SHO		EVERTims and IEM RoomEncoder

one combinations each round and clicking "Next" recorded the preferred spatial until all twenty one combinations were tested.

Table 3.3: Pilot Test Scenarios

1	AR Electric Drumset
2	AR Electric Guitar
3	VR Singing Kite

The RMS amplitude output of the binauralization was measured for all conditions when feeding white-noise into each condition. Then the gains of each spatial reverb were adjusted so that all conditions had the same binaural output level. This was done to remove the effects of perceived

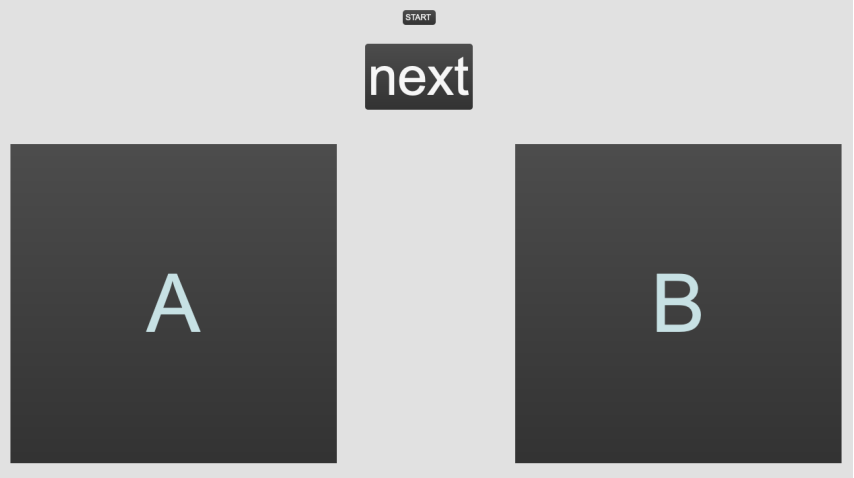


Figure 3.13: Preference test MaxMSP UI.

loudness on preference. The results from the pilot test presented in the following sections were analyzed to inform the use of spatial reverb with three specific XR musical instruments and a formal study would be required to confirm the observations.

Discussion of Results

Figure 3.14 displays a histogram of the informal test results that showed a preference for the hybrid spatial reverb Condition H-AMB[9]. The DRIR convolution reverbs (Conditions ZM, AMB, and EVR) rank lower in preference. It was expected that they not perform well in a 6DoF experience because the binaural output of these did not change in relation to the distance between the source and listener. Of the two DRIRs from microphone measurements (Conditions ZM and AMB) the test subjects chose the Ambeo VR Mic DRIR more than the ZM-1 DRIR for naturalness of the musical instruments. Comparing the energy plots of the two microphones shows that the energy of the sound source is localized more precisely in the ZM-1 DRIR energy plot than in the Ambeo VR Mic's (Figure 3.7 and Annex B) because the ZM-1 is a higher order Ambisonics microphone with greater spatial resolution. Further study is required to investigate interplay between localization and reverb on the sense of naturalness. The type

[9]Significance evaluated by a chi-quare test with $p < 0.001$

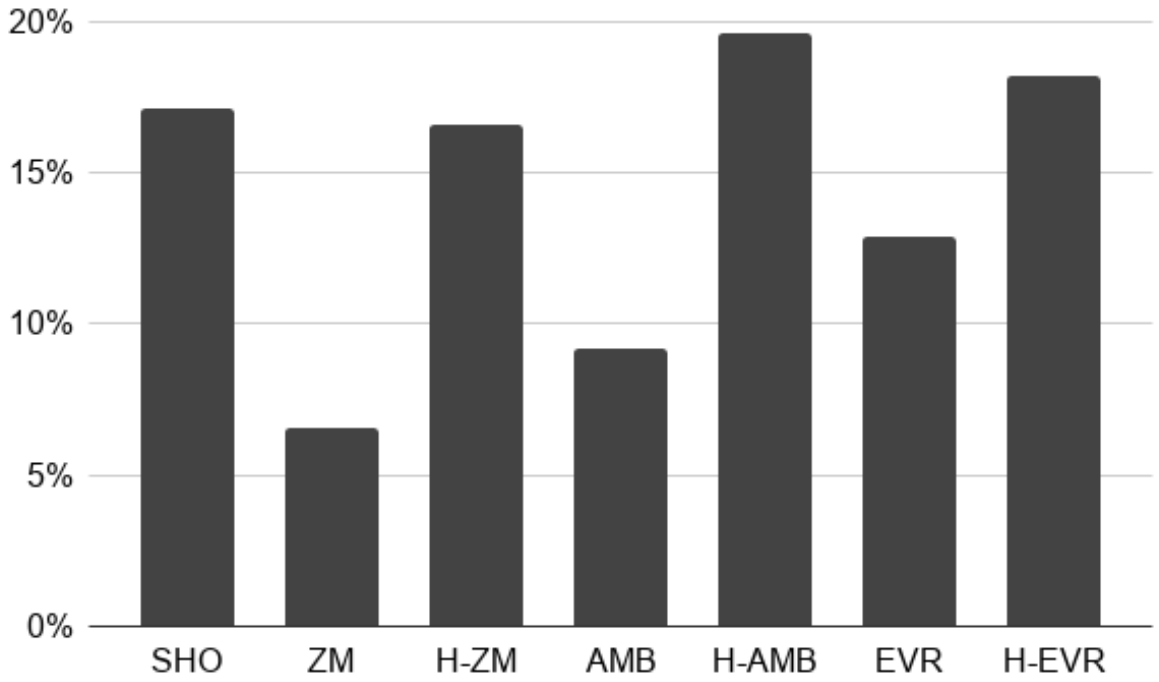


Figure 3.14: Histogram of informal test studying preference in naturalness between seven spatial reverb conditions (Table 3.2).

of acoustic sensor used in the two microphones may help to explain the results since Condition ZM uses MEMS microphones and Condition AMB uses condenser capsules. A 2017 study at New York University [153] compared the Ambeo VR Mic (Condition AMB) to a first order Ambisonics MEMS microphone and reported that subjects found the Ambeo VR Mic had better Naturalness[10]. The hybrid system with the ZM-1, Condition H-ZM, was preferred almost as much as the shoebox model Condition SHO but not as much as the other two hybrid spatial reverbs (Conditions H-AMB and H-EVR).

In a post-hoc comparison of the AR electric guitar/drumset and the VR Singing Kite results there is less of a difference between the hybrid spatial reverbs and shoebox model as seen in Figure 3.15. This may have been due to the way the VR Singing Kite was implemented because there was a lot of jitter on the position of the kite which may have had a negative impact

[10]Naturalness queried by ”Does the performance appear to take place in an appropriate spatial environment?”

on the naturalness of all spatial reverb.

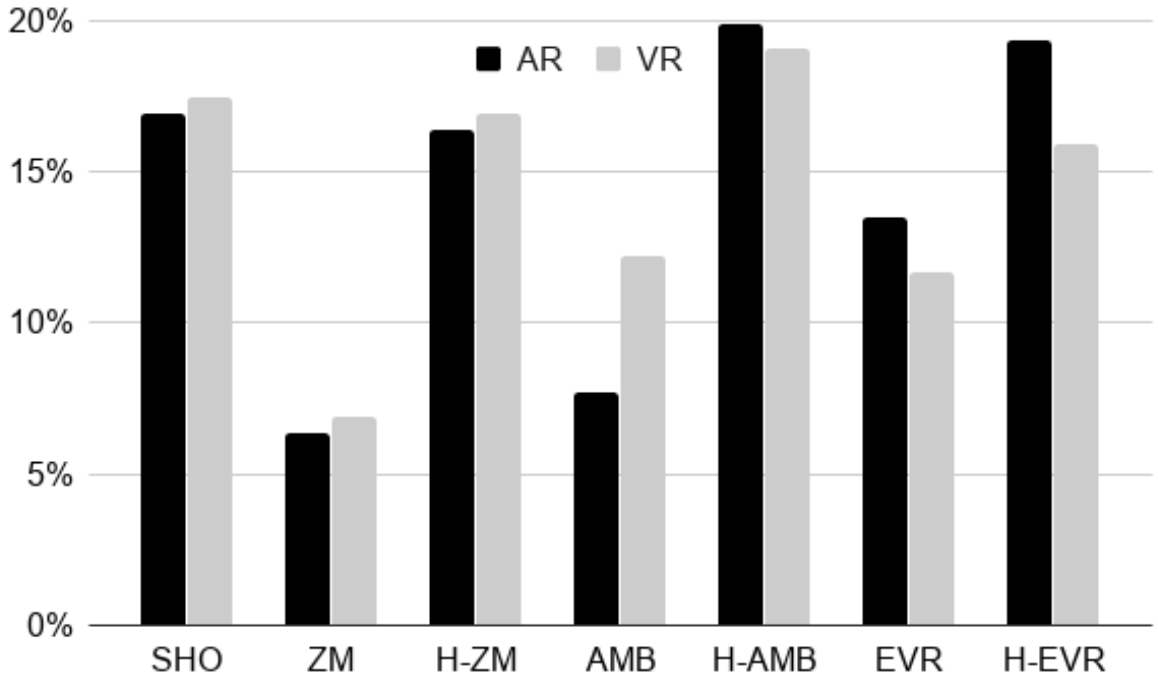


Figure 3.15: Histogram of the seven spatial reverb conditions from Table 3.2 separated into preferences with AR and VR instruments.

Figure 3.16 compares results from the AR electric guitar and the AR electric drumset. Condition H-AMB, hybrid with measured DRIRs, performs well for the drumset which is 3DoF and Condition H-EVR, hybrid with synthetic DRIRs, performs well for the guitar which is a 6DoF case. Further study would be required to investigate the relationship of spatial reverb in 3DoF versus 6DoF scenarios and AR versus VR.

Measured DRIRs vs Synthetic Reverb

The breakdown of the results into the AR electric guitar and AR electric drumset results is given by Figure 3.16. A flip in preference between the two instruments is observed with hybrid spatial reverb of Condition H-AMB and fully synthetic reverb Condition H-EVR. The AR electric drumset results show a preference for the Ambeo VR Mic shoebox hybrid whereas

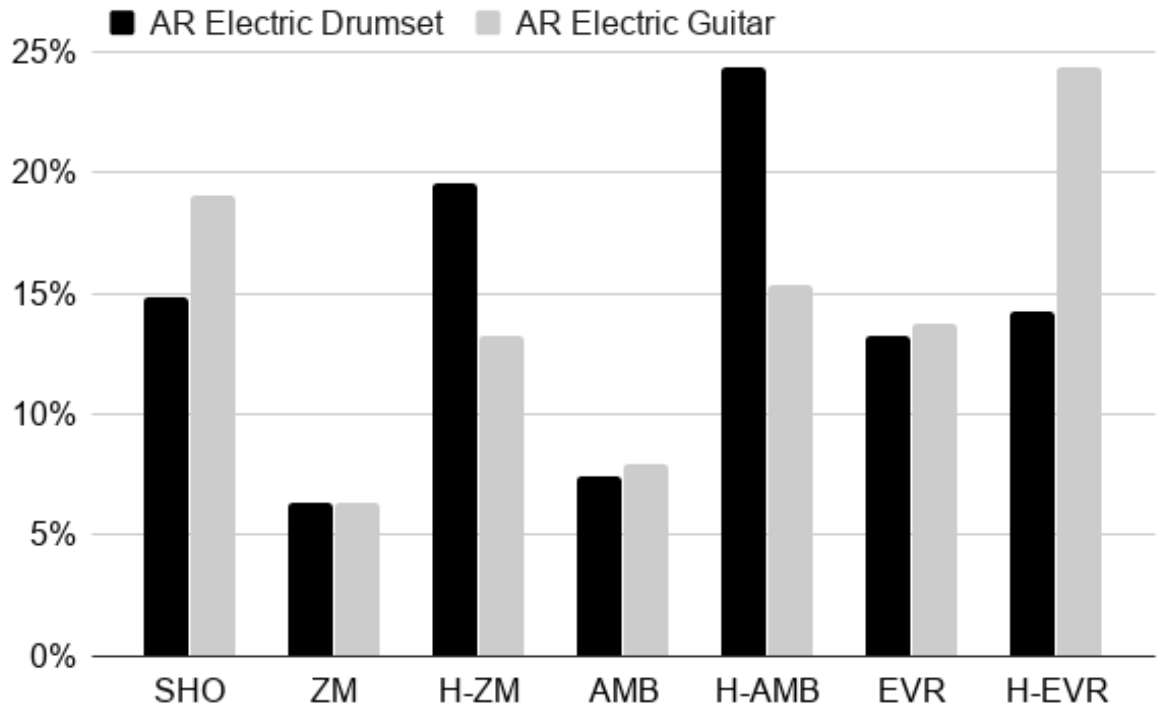


Figure 3.16: Histogram of preference by scenario (AR electric guitar and AR electric drumset) for the seven spatial reverb conditions (Table 3.2).

the AR electric guitar shows a preference for the shoebox raytracing hybrid. The drummer was in a fixed position (3DoF) with the sources (drums) at the same distance as when the DRIRs were measured. In the case where the subject is listening to the AR electric guitar (from the drummer's position) the synthetic spatial reverb was preferred due to the decrease in localization of the measured DRIR when the guitar player moved away from the measured DRIR position. The guitar (Figure 3.17) sound source was at a distance, S, from the listener farther than the distance to the closest reflection. Further study is required to evaluate if interpolation of many DRIR positions (Section 3.6.1) or extension of the measured hybrid spatial reverbs can improve the naturalness for measured DRIRs when S is dynamically changing.

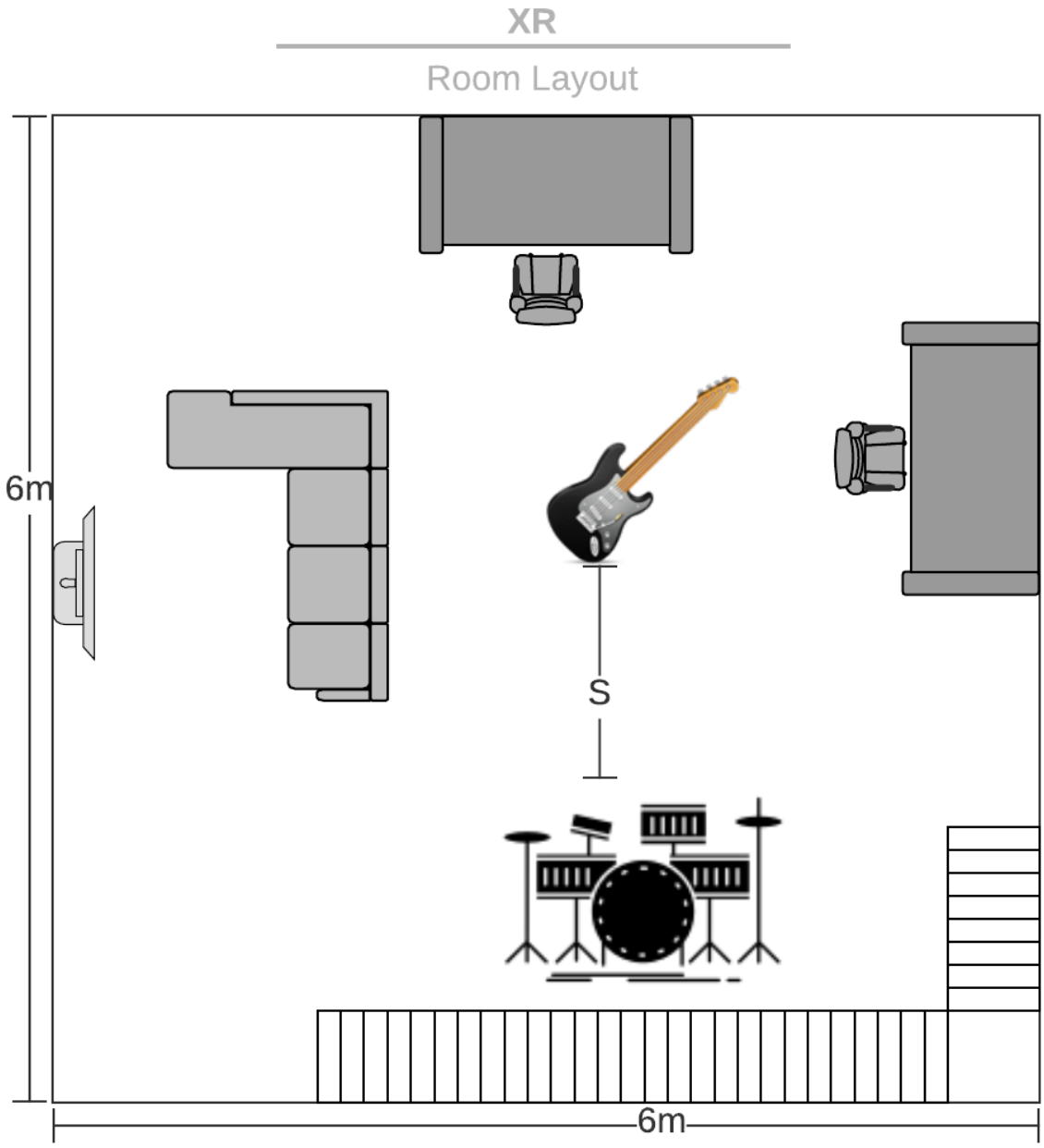


Figure 3.17: Floorplan of the 6*x*6*m* room where the XR musical instruments were evaluated. This was also the rehearsal space for the XR concert (Chapter 5).

3.6 Tuning DRR

The DRR and clarity results in Figure 3.18 were calculated from recordings of the binaural output[11] to include both the spatial reverb and binaural filters to align with the binaural synthesis that the test subjects experienced (DRR and Clarity in Table 3.1 were calculated from the microphone recordings). Although no conclusions may be drawn from the informal test, there did not seem to be a clear relationship between DRR and preference for naturalness because Condition H-AMB was more natural than other spatial reverbs with higher DRR values. Regarding clarity, the synthetic shoebox reverb (Condition SHO) had the greatest clarity value, but was less natural than Condition H-AMB which had 8dB lower clarity. This result might be attributed to the low frequency drums (kick and low tom) which sounded hollow with Condition SHO. Future studies may add a position based Feedback Delay Network (FDN) reverb to the shoebox spatial reverb model to improve the low frequency instruments.

The RT_{60} time was calculated for all the DRIRs of Figure 3.18 including individual drums. These RT_{60} measurements of binaural output were performed to gather information on the effect of spatial reverb on decay times. The mean was 450ms and the median 430ms. All spatial reverb conditions were within one standard deviation, 258ms, of the median except for Condition ZM whose mean and median was 620ms. Before the study was performed an exponential decay to the measured DRIRs was made to increase the DRR which may have also adjusted the RT60 time. The adjustment did improve the localization (perceived by the author), but was removed for the study because it had a detrimental impact on the rotation effect (3DoF). Modifications to the measured DRIRs could have been performed in a systematic way (in addition to modifying the decay), but were not pursued because it was thought to lead to a similar approach that creates the reverb fingerprint described by [95].

[11] An impulse (click~) was sent into the spatial reverbs A to G in the MaxMSP patch and recorded with the sfrecord~ object.

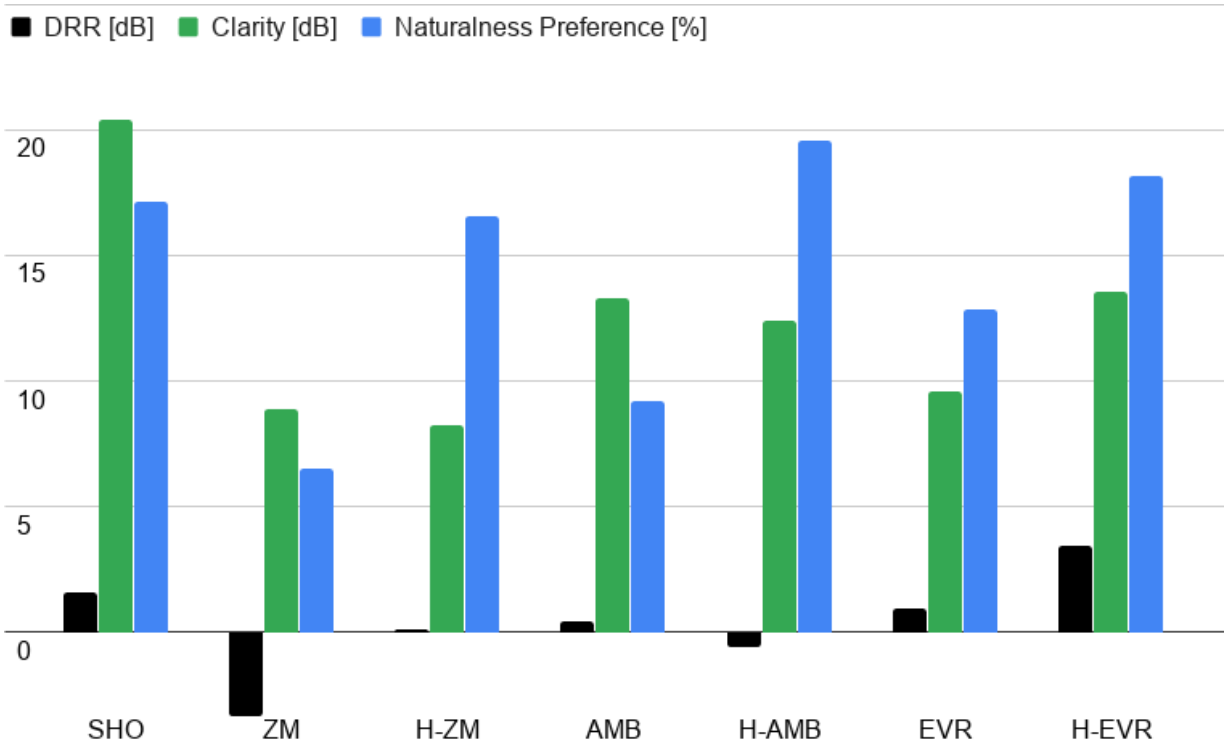


Figure 3.18: DRR, clarity and naturalness preference for the seven spatial reverb conditions of Table 3.2.

3.6.1 Interpolation of DRIRs

Another approach to compensate the DRIRs for XR spatial reverb is to create a database of DRIRs that map out listener and source position for a given space and interpolate between them as the listener and sources move in that space. Two interpolation methods were studied for use with DRIRs, linear and biharmonic spline, to evaluate their impact on spatialization and audio quality. Starting with a grid of triangularly spaced positions 1.5m apart, the interpolation methods were used to double and quadruple the density. The biharmonic spline solution [69] given by

$$s(x) = \sum_{j=1}^{n} w_j g(x, x_j), \tag{3.2}$$

$$g(x_i, x_j) = |x_i - x_j|^2 (\ln(|x_i - x_j|) - 1)$$

was used to interpolate a two dimensional grid of impulse responses (Figure 3.19).

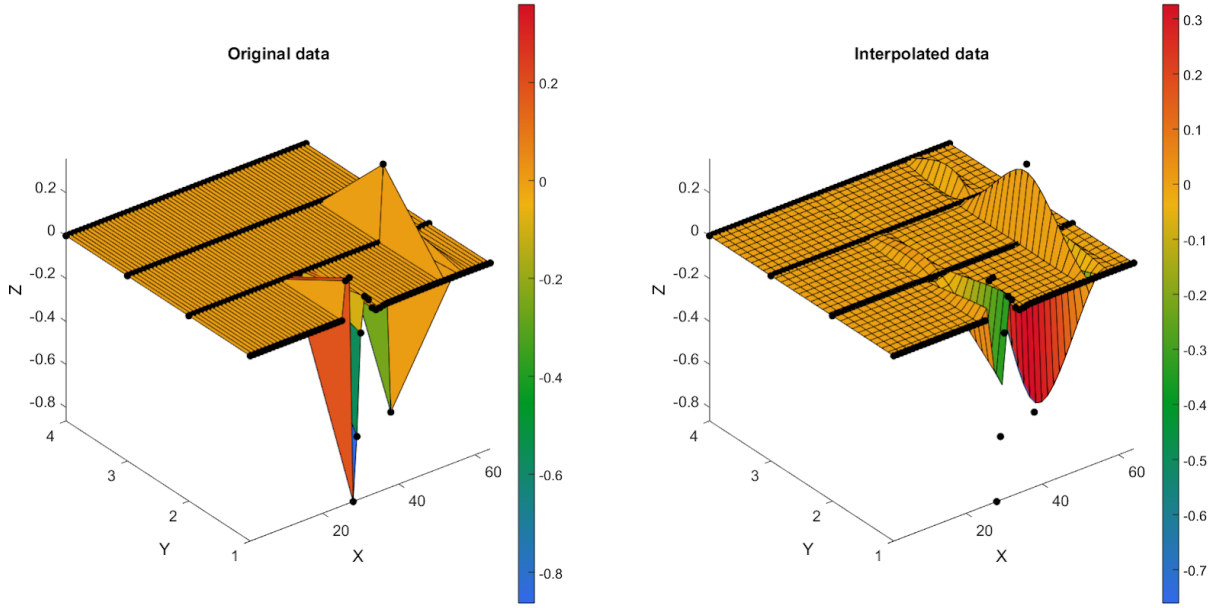


Figure 3.19: Interplolation plots of DRIRs. The x-axis is samples, z-axis is amplitude and y-axis is the distance between the impulse responses.

When the listener's position moved to the next closest DRIR position, the processing in the efficient convolution method with the old position's DRIR continued while starting to fill the new frames with the new DRIR position.

More investigation is required to find out how processing two neighboring DRIRs and crossfading between them compares with the changeover method described above. An open source 3D Tune-In Toolkit presents means of interpolating HRIRs which may also be used for DRIRs [66]. The University of Huddersfield has created a database of BRIRs using a custom made head-rotation system to measure directional BRIRs as well as FOA and omni impulse responses which would be useful in the comparison [52].

Using machine learning to model impulse responses by estimating acoustic parameters has been shown in [119] but evaluating this technique in a realtime audio is beyond the scope of this dissertation.

3.7 Considerations of using DRIRs for Spatial Reverb

Performers may not be able to measure their own DRIRs for every venue. A solution could be to make databases [49] [142] of impulse responses readily available. If the models used with EVERTims raytracing are given more detail either through graphic designer or computer vision mesh creation there may not be a need to obtain microphone measurements. For performers and composers who do not wish to model their spaces, taking a DRIR measurement and applying hybrid spatial reverb may be the best approach. This was the case for the performance reviewed in Chapter 5.

A hybrid system of synthetic first reflections and measured spatial reverb was developed to enhance the immersiveness of AR musical instruments. The DRR was tuned based on a distance model for 6DoF performances. It was shown that spatial reverb can be used to make AR musical instruments sound more natural in the space.

Chapter 4

Spatial Audio Effects for Extended Reality Musical Instruments

The set of XR musical instruments described in Section 3.3 were initially designed to explore spatial reverb audio effects of instruments with different positions. The AR electric guitar's position was tied to the performer's position, which was tracked by the 6DoF HMD. The AR electric drumset's position was static and predefined in the spatial audio system. The VR Singing Kite's virtual position was mapped to real world coordinates. This chapter will demonstrate how four traditional audio effects (looping, feedback, delay, and compression) were re-contextualized for use with those XR musical instruments. Spatial looping, spatial feedback, spatial delay, and spatial compression described in the following sections will be categorized as XR audio effects since virtual and real world coordinates are input parameters affecting the audio processing.

4.1 Spatial Looping

The VR Singing Kite was programmed to emulate a kite flying in the wind in order to have a constantly moving sound source. Like a real kite, the user could set the string handle down and allow the kite to fly on its own which enabled the listener to perceive distance effects without the cognitive load of controlling the kite. The instrument was extended to create and control multiple copies of the kite. In traditional audio mixing, looping is an audio effect that enables replaying a certain section of audio by triggering the start of the loop and specifying the loop duration. When spatial looping is triggered, the position of the kite is held constant and a section of the kite's recording is looped in time. Creating a copy of the kite enables the exploration of binaural masking where the player loops the copied kite ipsilateral to one ear and then moves the original kite to the opposite ear.

Figure 4.1 is a flow diagram of the spatial looping effect. The 6DoF tracking information and sound source's position are sent via OSC into the MaxMSP patch which waits for the user input trigger (also sent via OSC) to initiate the looping. When the trigger is received, a copy of the spatial audio reverb is enabled with the positions held static until the loop is disabled. The creation of many spatial loops generates audio layers that can form an orchestral backing for the VR Singing Kite Concerto discussed in Chapter 5.

4.2 Spatial Delay

The inspiration for spatial delay came from stereo delays as applied to guitar tracks which dynamically pan the guitar between the left and right channels. For the AR electric guitar, the delay is tied to the position where the guitar was binarualized. As the guitarist moves about the space they leave a trail of sound. This is implemented by adding a delay to the binaural output as seen in Figure 4.2.

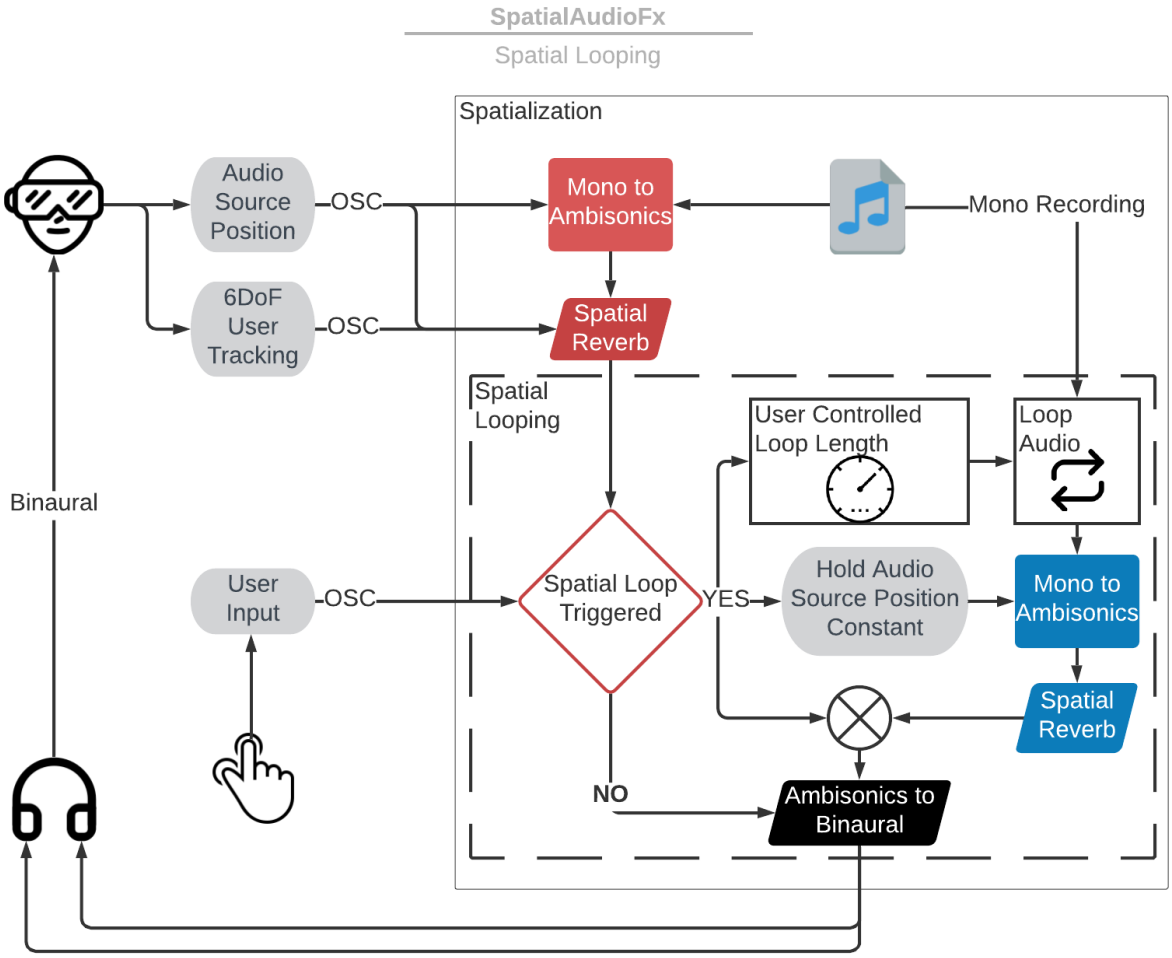


Figure 4.1: Block diagram of the spatial looping effect for an XR audio system.

4.3 Spatial Feedback

The spatial feedback effect arose from an idea to turn a room into a musical instrument. Alvin Lucier's piece, *I Am Sitting in a Room*, [103] presented a similar concept where the number of times he re-records himself reading in a room increase the effect of the room's reverb. Spatial feedback is similar in that the output of the spatial reverb is fed back and mixed with the mono input signal, as seen in Figure 4.3.

There is a one second delay introduced so that the feedback does not become immediately unstable. If the position of the sound source is stationary, then the feedback can get out of control

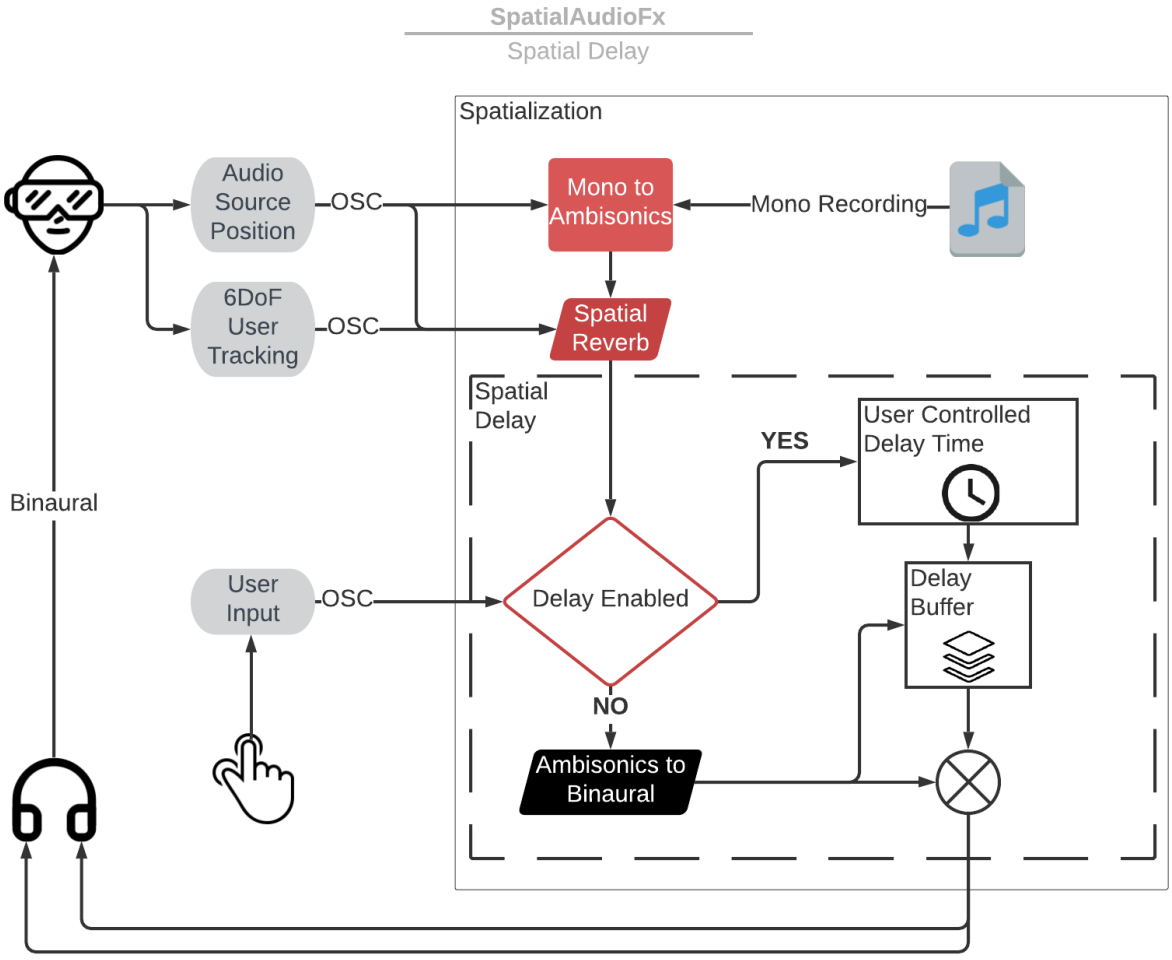


Figure 4.2: Block diagram of the spatial delay effect for an XR audio system.

(even with the delay) because the spatial reverb amplifies the same early reflections. If the sound source is moving, then the feedback amplifies different directions as the source moves creating more interesting textures.

4.4 Spatial Compression

The spatial compression presented in Figure 4.4 is different from the audio coding spatial compression used to reduce the bitrate required to transmit 3D audio formats [118, 124, 125, 139]. The re-contextualization of compression presented in this research is more analogous to the

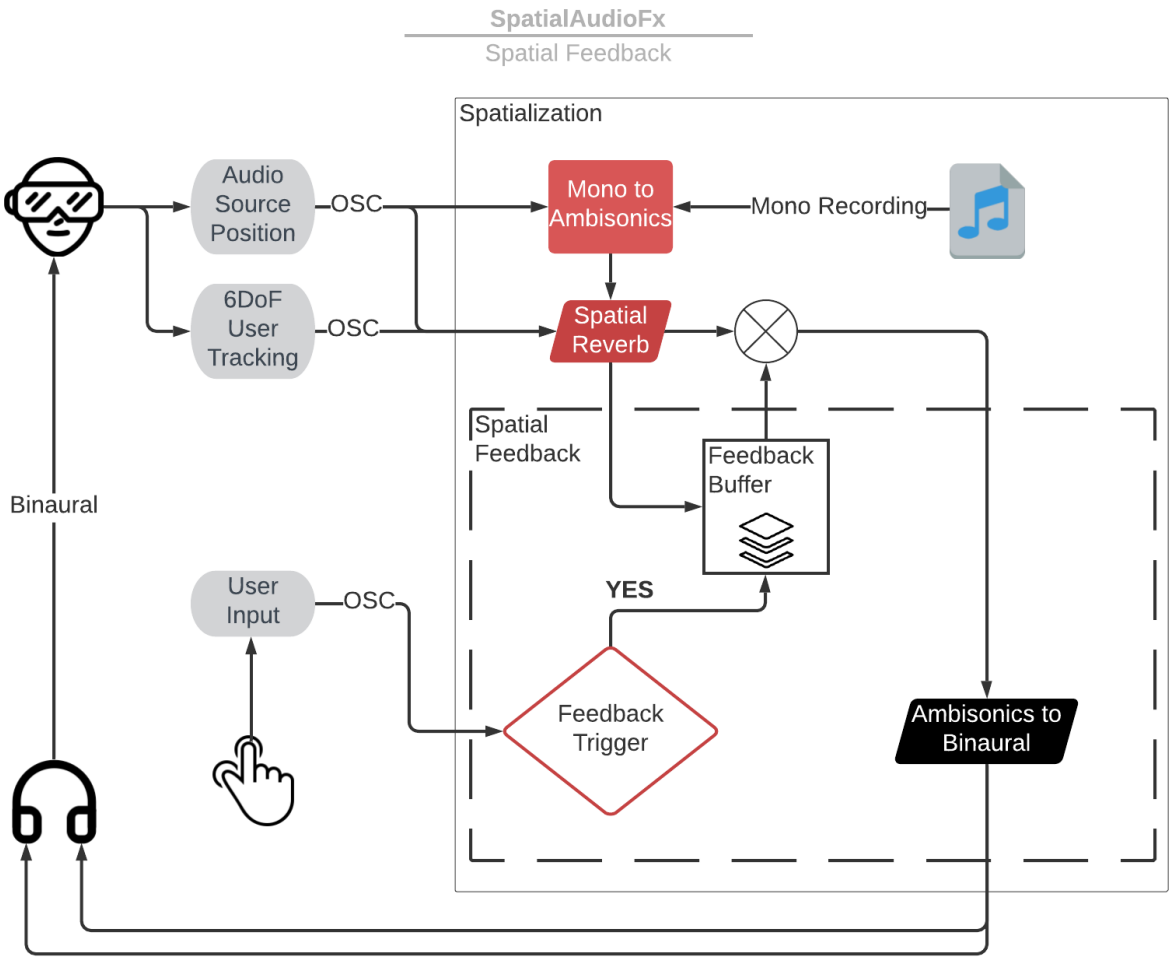


Figure 4.3: Block diagram of the spatial feedback effect for an XR audio system.

amplitude compression of DAWs. To understand how spatial compression differs, the directivity of sound sources must first be discussed.

Sound-emitting sources have frequency dependent patterns of sound radiation. For example, a person speaking will sound differently in both loudness and timbre if they are facing towards the listener or away from the listener. Directivity of microphones and loudspeakers is typically measured by recording the SPL at different angles and reported in the form of frequency response polar plots [22]. Directivity can be imparted on XR musical instruments to give them more complex characteristics, but currently most game engine audio spatializers only allow

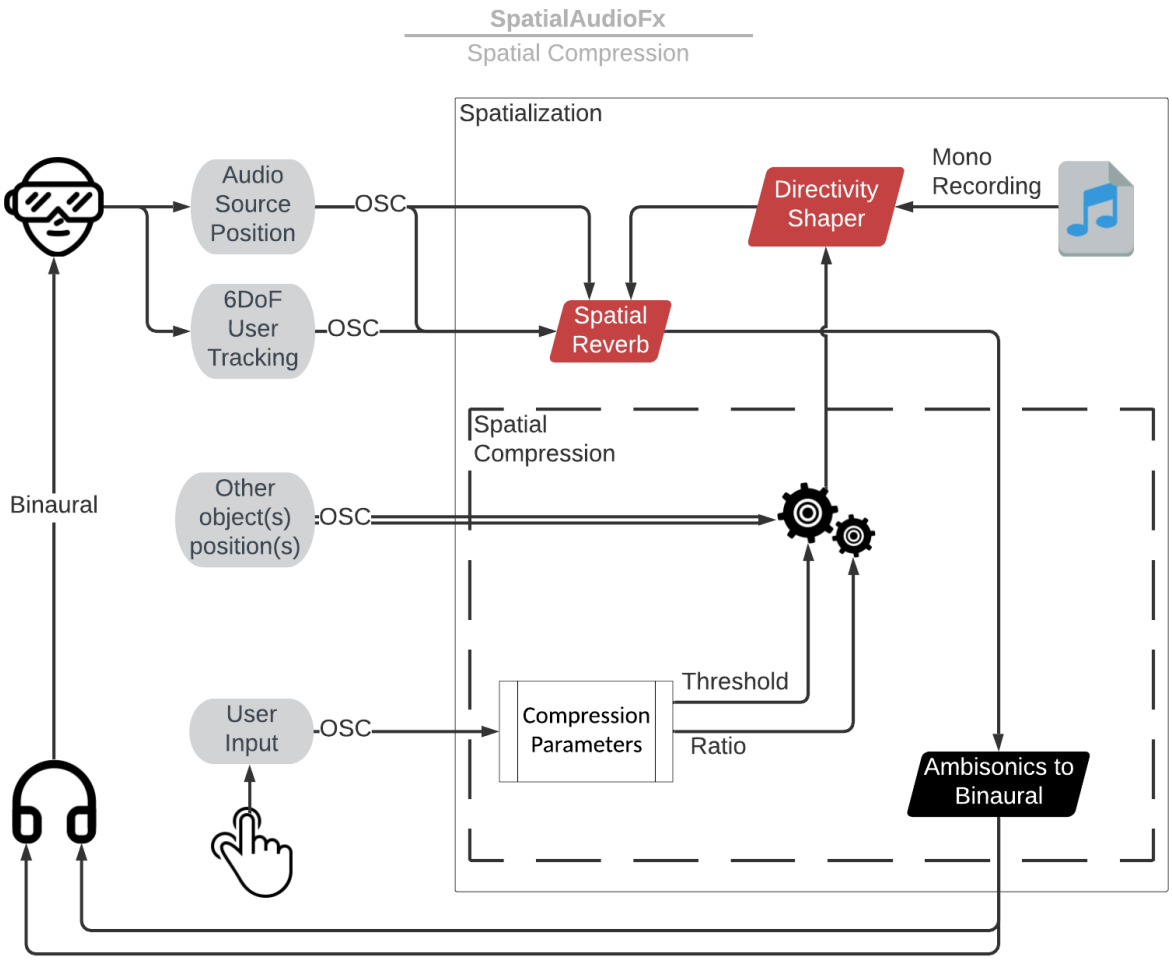


Figure 4.4: Block diagram of the spatial compression effect for an XR audio system.

content generators to modify the sound source's frequency and not the directivity. IEM plugins working in the Ambisonics domain provide a Directivity Shaper [16] (Figure 4.5 shows the UI of the Directivity Shaper plugin) that can apply directivity onto a mono source. SOFiA [57] is sound field analysis toolbox that can be used for further research into directivity of sound fields.

Spatial compression is the effect of modifying the directivity of the sound source which can be both volume dependent and position dependent. User input and object positions are inputs to the Directivity Shaper. This works in combination with the Directional Compressor plugin (Figure 4.6) to compress the amplitude of the modified directivity.

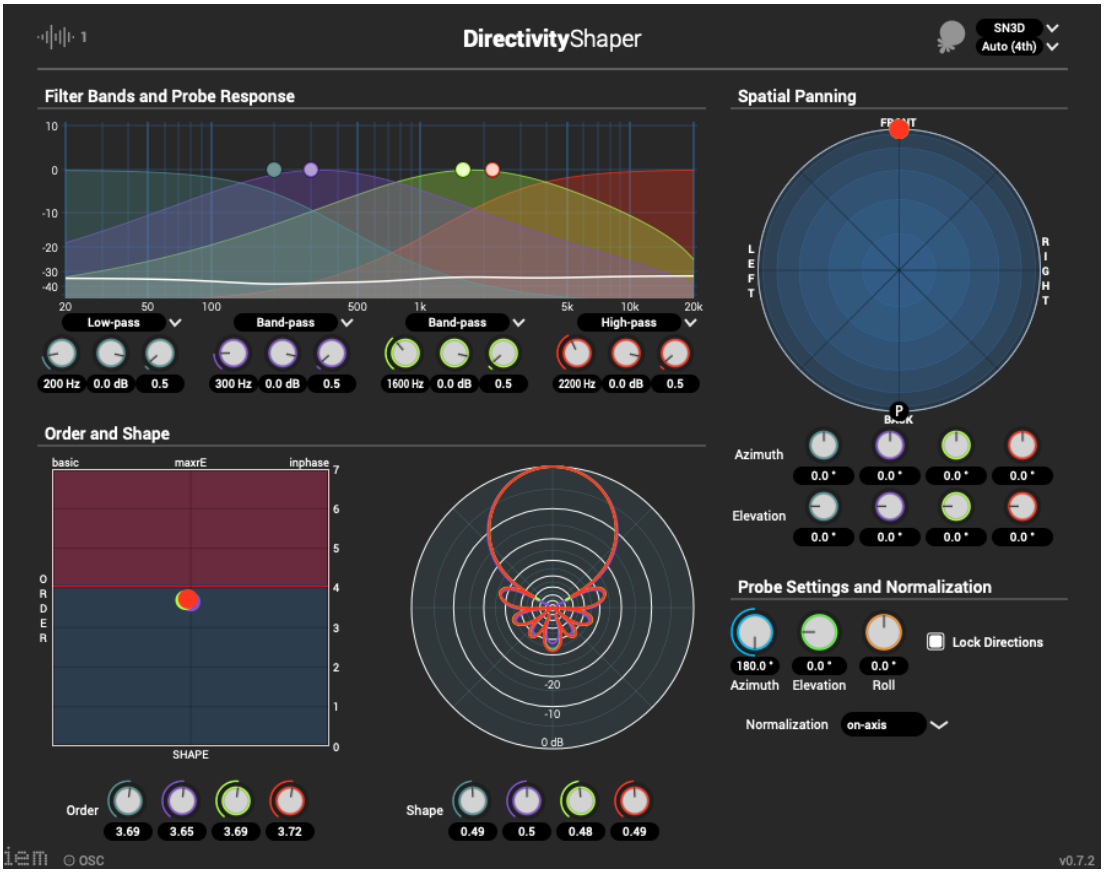


Figure 4.5: Screenshot of Directivity Shaper VST Plugin by IEM.

Once a natural sounding spatial reverb (Chapter 3) establishes a reference to the listener, spatial audio effects can be applied to manipulate the sound's position. Copies of instruments can be made with spatial looping, trails of sound can be generated with spatial delays, room acoustics can be multiplied by introducing feedback, and the instrument bodies can be manipulated with spatial compression thus extending techniques for musical performance with these new XR technologies.

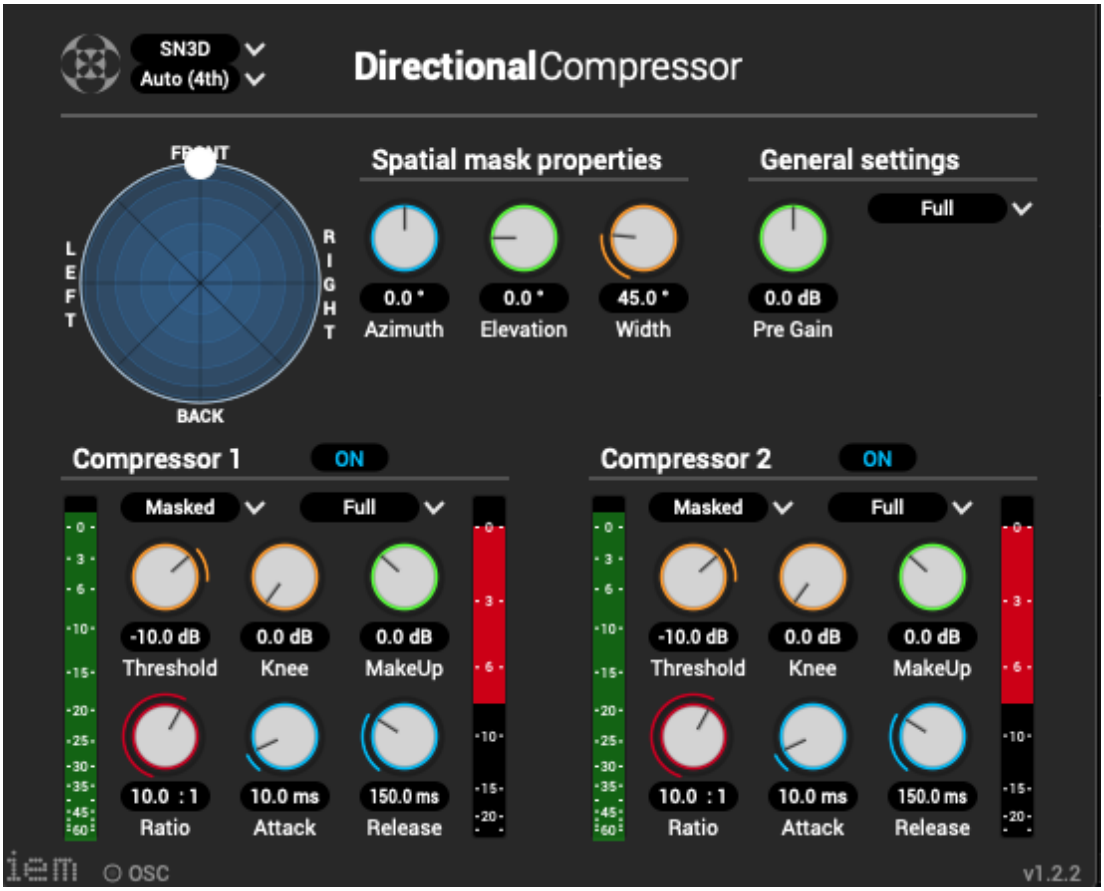


Figure 4.6: Screenshot of Directional Compressor VST Plugin by IEM.

Chapter 5

Extended Reality Audio Concert

Figure 5.1: Spherical Sound Search concert flyer.

This chapter will conclude with a reflection on the concert, *Spherical Sound Search*, that was performed on Saturday, February 29, 2020 in the Conrad Prebys Music Center (CPMC) Experimental Theater at the University of California San Diego featuring the XR musical instruments designed for this research (Figure 5.1). Although the XR musical instruments were

designed for performers wearing headphones, the concert showcased an XR audio performance for loudspeaker reproduction using an ambisonics decoder designed for the multi-channel arrangement of the theater. The MaxMSP patches created for the concert are available in github [113].

5.1 Composing for XR Musical Instruments

The concert was an exploration of XR musical instruments. The first piece, *VR Singing Kite Concerto*, was composed for VR with a single stationary performer. The second piece, *Between a Log and a Pluck Place*, was an MR composition where a single performer moves around the stage. The third piece, *Push Pull*, was composed for an AR electric guitar and AR electric drumset. The three pieces composed for the concert began with emulation of natural sounding acoustics and then used the spatial audio effects from Chapter 4 to augment the spatial reverb in a way that was no longer natural, but musically motivated.

5.1.1 *VR Singing Kite Concerto*

The VR Singing Kite app used in the pilot study was extended to enable the creation of many copies of the kite. The performer is oriented in the same direction as the audience, towards the projection screen. Screen sharing of the performer's HMD allowed the audience to see the VR world of the performer. Ambisonic rendering of the audio did not include the head rotation of the performer so that the positions of the kites remained static with respect to the venue.

The piece begins with the performer picking up the handle of the kite using the right controller of the Oculus Quest. Playback of the kite sound begins once the kite is off the ground. The performer controls the direction of the kite by the direction of the kite handle and controls the distance of the kite by pressing the left controller's index trigger. Flying the kite 360^{o} at different distances gives the audience a sense of the space in which the kite is being "flown". Next, the

performer engages the spatial feedback by pressing the hand trigger on the left controller. If the right hand kept the kite stationary then the feedback would resonate quickly. To diffuse the resonance, the performer increased the distance of the kite to reduce the sound of the kite. When the feedback was engaged while the kite was moving, even at close distances, different resonance frequencies increased and diminished in gain but never grew out of control. The performer then begins to create copies of the kite by gesturing in the direction of the kite and pressing a button on the left controller. A static visual copy and a spatial audio looped copy[1] of the kite are generated when the button is pressed. The piece concludes after the performer has filled the virtual space with many copies and explored spatial feedback at the same time.

5.1.2 *Between a Log and a Pluck Place*

The musical instruments used in this piece were both augmented virtually and appeared physically on stage exemplifying an MR performance scenario. This piece began with spatial looping of an electric guitar plucked string and a teponazli mallet hit. It was a minuet[2] with a guitar pluck on count one and percussion hits on counts two and three. The performer utilized the 6DoF HMD to track their position as they walked around the guitar and drum. In *VR Singing Kite Concerto*, the orientation of the audio scene was fixed with respect to the loudspeaker rendering, whereas in this piece, the scene orientation followed that of the performer's head as they moved around the two sound sources. A humorous prop was utilized to inform the audience that they were being aurally localized at the performer's position by having the performer wear over-sized plastic ears on their head. The performer's movements explored the directivity and timing of the two sound sources. The distance to each instrument controlled the rate of playback of each loop so that only at the midpoint were the two instruments in sync for the minuet. Another effect of this distanced-based control was that the pitch of the instruments decreased as the distance of the

[1] Spatial looping for multiple copies was implemented using the poly~ object.
[2] In triple time.

performer to the instrument decreased.

5.1.3 *Push Pull*

This piece featured the AR electric guitar and AR drumset. The drumset consisted of a snare, hi-hat, crash, clap, and kick. The guitar signal was transmitted wirelessly to the computer using a UHF transmitter so that the performer was free to move all the way around the drummer. For this piece, the orientation of the loudspeaker playback was fixed so that the sounds from the instruments were heard as if the listener was centered in the seating area looking straight at the drummer for the entire piece.

Areas of interaction were defined by the radius from the drummer to the guitar player. At a farther radius, the guitar's position pulled the directivity of the drumset towards the guitar. The artistic intent was to manipulate the directivity of the drumset and increase the directivity in the direction of the guitar player when the volume of the guitar signal increased. The feedback level increased as the guitar moved closer to the drumset. When the feedback became unstable, the drummer was able to clear the feedback by hitting the crash. Thus, the crash punctuated rising musical envelopes in the piece initiated by the feedback. In combination with the AR drumset, the spatial compressor was used in a directivity-ducking manner in which the directivity of the AR guitar was notched in the direction of the AR drumset.

The score for *Push Push* is shown in Figures C.1 and C.2. The AR electric guitar part draws from a Spanish guitar folk melody. Spatial delay is notated by dashed arrows. The guitar's directivity was simulated by mapping the guitar player's rotation to the left/right axis. This rotation was notated by clockwise/counterclockwise circular arrows with an "R" in the center. Movement around the drummer was notated by similar arrows, but with "A_1" and "A_2" in the center for the areas of interaction with the AR electric guitar performer.

5.2 Setups for XR Musical Performance

5.2.1 Rehearsals

The design and evaluation of the author-created XR musical instruments were done using headphones (Chapter 3). To rehearse for a performance over loudspeakers, a five-channel Ambisonics decoder was created to simulate, in a smaller rehearsal room, what the audience would experience in the larger concert venue. The drummer was placed in the center of the four loudspeakers as Ambisonics rendering has a sweetspot [50] where the spatial audio effect is optimal. The three compositions in Section 5.1 were evaluated with this setup.

VR Singing Kite Concerto was rehearsed casting the VR view of the performer to the audience by casting the video through WiFi to a Google Chromecast. It was found during the soundcheck that the Chromecast was incompatible with the projector of the venue. Therefore, in the performance, the video from the Oculus Quest was displayed on a laptop using a USB cable connection and then through HDMI from the laptop to the projector.

The 6*x*6m rehearsal room was too small to practice the full range of movements prescribed for the performance of the second and third pieces. *Between a Log and a Pluck Place* needed a wide room so that the performer could walk completely around the guitar and the log drum. Similarly, in *Push Pull*, the guitarist could not move all the way around the drumset in the outer area of interaction.

5.2.2 Concert Setup

The CPMC Experimental Theater offers an octophonic loudspeaker layout for playback. Figure 5.2 is a modified seating chart of the venue showing the relative locations of the drumset, seats, and loudspeakers. The guitar was free to move in the performance space and utilized different areas for interaction with other instruments.

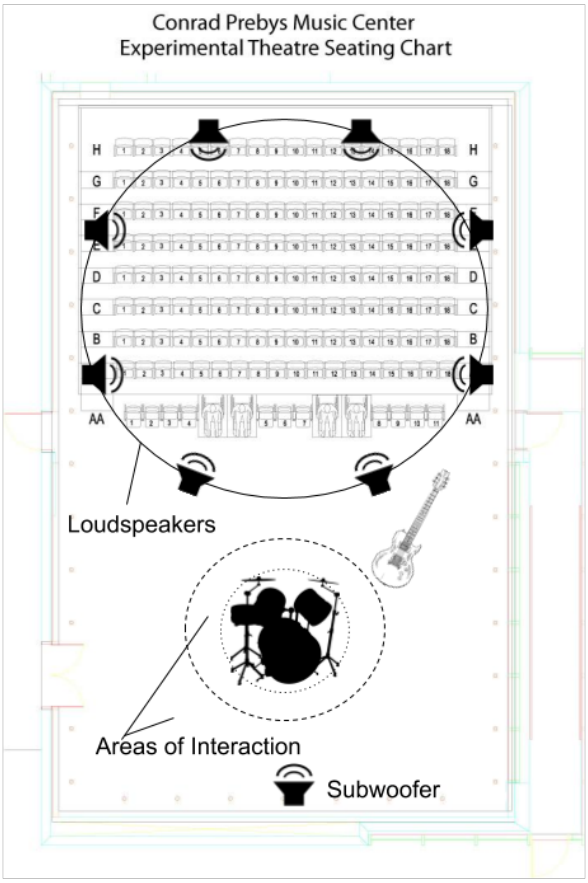


Figure 5.2: Conrad Prebys Music Center Experimental Theater seating chart showing the positions of the guitar, drums, and octophonic loudspeaker layout for playback.

5.2.3 Ambisonics Decoder Design

An Ambisonics decoder was required in order to play Ambisonics content over loudspeakers. The Pyramix DAW has an integrated ambisonics decoder for standard loudspeaker layouts [105]. Ambisonics decoders, made by tools such as the Ambisonics Docoder Toolbox [86], can be created for customized arrangements of loudspeakers. IEM plugins also provide the ability to create an Ambisonics decoder with the AllRADecoder plugin (Figure 5.3). A minimum of four loudspeakers must be used with corresponding azimuth and elevation coordinates. The decoder can be exported for Ambisonic order 1-7 to a JSON file which is read by the SimpleDecoder plugin (Figure 5.4).

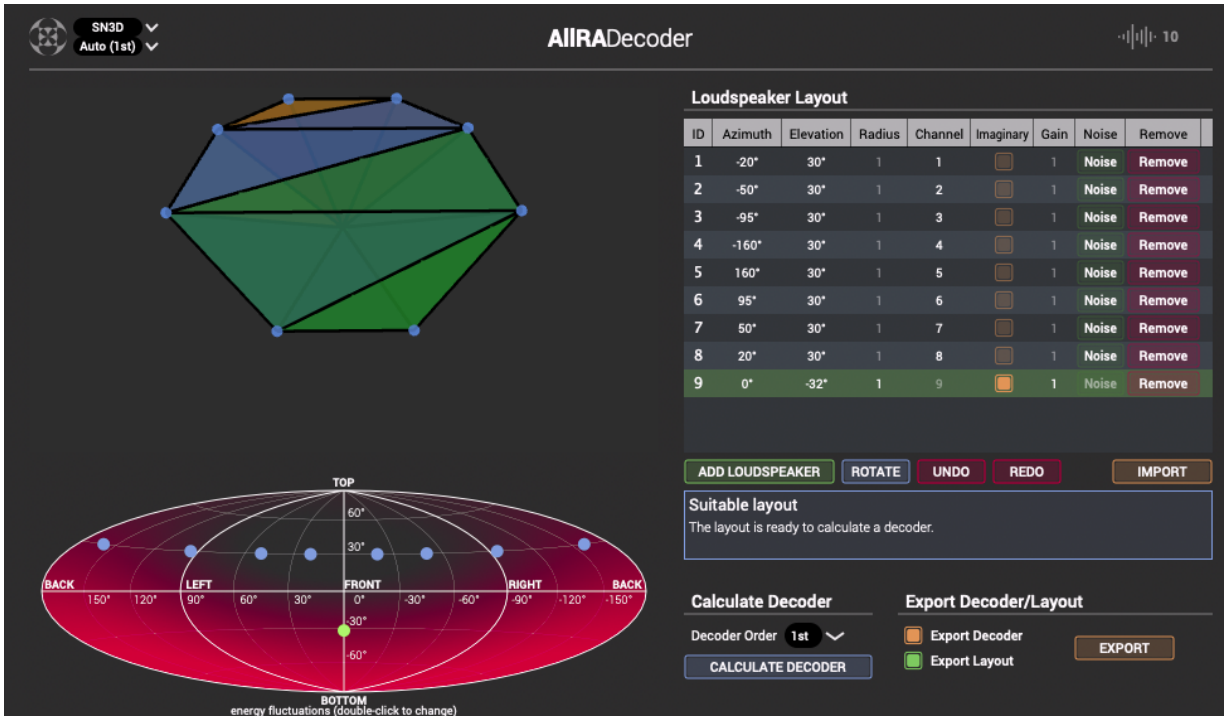


Figure 5.3: Screenshot of the IEM AllRADecoder plugin.

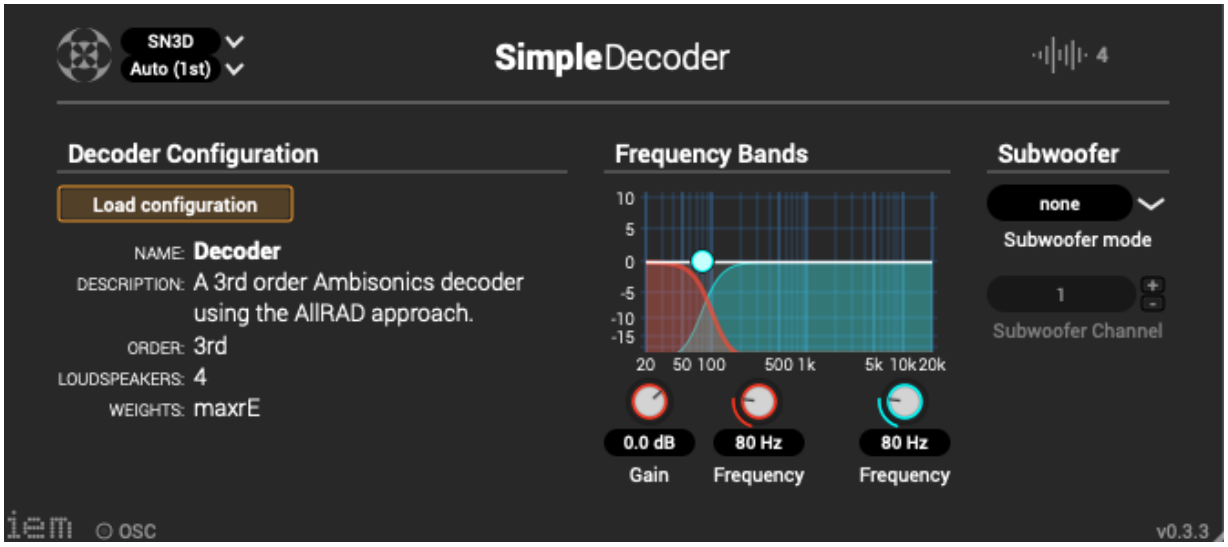


Figure 5.4: Screenshot of the IEM SimpleDecoder plugin.

5.2.4 XR System

Although the Oculus Quest, a 6DoF VR HMD, was used for all three pieces, other HMDs could have been used. A 3DoF HMD could be used for the first piece. The second piece requires

a 6DoF VR HMD which could be tethered to a computer. The first two pieces could be set up as art installations with playback over headphones. The third piece requires 6DoF tracking and a wireless electric guitar. The tracking in the concert was done by wearing the Oculus Quest like a backwards baseball cap, but it could have also been performed with a 6DoF AR HMD.

Limitations of the Oculus Quest's guardian (6DoF tracking area) were discovered during the soundcheck. The first step in setting up the Oculus Quest in the venue was creating the guardian by scanning the room and marking the perimeter using the controller. It was found that the maximum area of the guardian was only three-quarters of the stage. The perimeter was marked with tape on the floor to guide the performer because tracking data was lost when they moved outside of the guardian area. This restriction is due to the memory available on the HMD and may increase in future versions/models. The guardian was sensitive to low light and changes in light which resulted in loss of 6DoF tracking. This sensitivity was discovered when the lights were lowered to a concert setting. The HMD lost tracking and took some time to readjust to the new light level. This was not an issue during the performance, but it is unknown how the Oculust Quest would handle performances with changing light cues.

5.3 Reflection

Video recordings of the *Spherical Sound Search* XR Concert are available for viewing [111, 112, 114]. Adjustments to the Ambisonic playback over loudspeakers were made during the soundcheck at the venue. The Ambisonics decoder (Figure 5.3) initially designed for the octophonic loudspeaker system of the CPMC Experimental Theater set the elevation of the loudspeakers to match the actual height of the eight loudspeakers. It was difficult to localize sounds on the horizontal plane using this decoder, but improved when the decoder was re-configured with elevation set to 0^{o} for the loudspeakers. Since the Ambisonic decoder assumes that the speakers are equidistant to the audience, further adjustments were made to compensate

for the actual distances in the venue. The IEM Distance Compensator plugin (Figure 5.5) was used to adjust the gain and time of arrival from each loudspeaker to the center of the venue's audience.

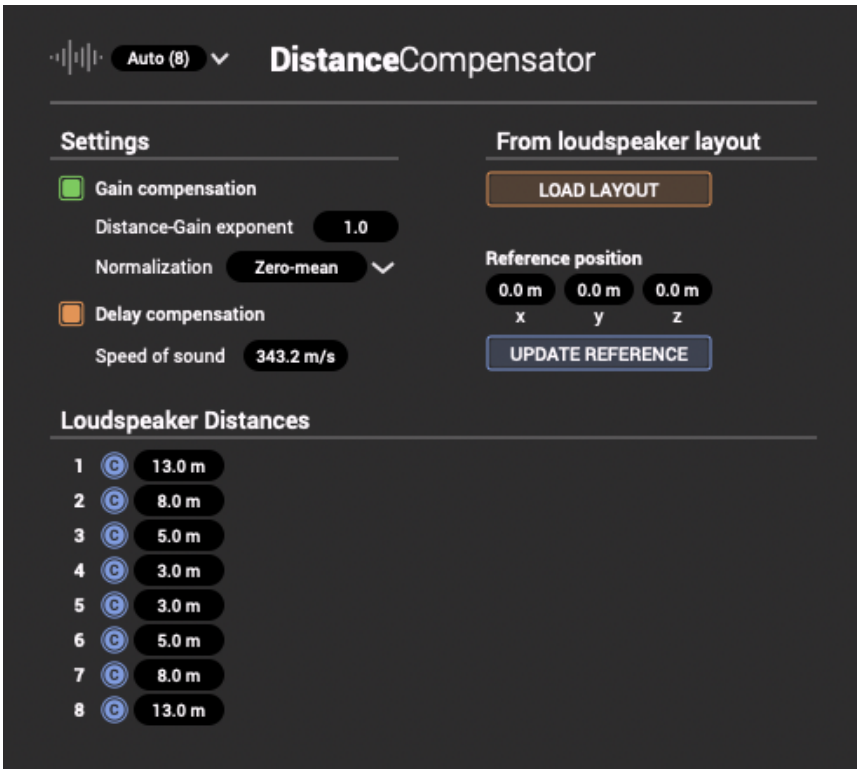


Figure 5.5: Screenshot of the IEM DistanceCompensator plugin.

Performers are typically able to monitor the audio by using loudspeakers placed on the floor of the stage facing the performers or in-ear headphones. In this particular venue, floor monitors were not used because the performers were able to hear what was output from the front two loudspeakers located above the center of the stage. Future concerts may explore the use of binaural in-ear monitoring to provide the performers with personalized spatial cues of their performance. The performers could choose to monitor the audio from the audience's perspective or from their own. A separate computer could be used to create the binaural monitoring mixes and use a low latency two-channel transmitter to send the mixes to the performers' in-ear headphones.

Connections between music notation and choreography were discovered in composing for

this concert. In *VR Singing Kite Concerto*, triggering the spatial feedback by squeezing the left hand controller was an example of a gesture that became part of the piece even though it was not notated in the composition. On the projector this gesture was seen as the left hand opening and closing which was similar to a conductor asking for more from an instrumentalist. The score of *Between a Log and a Pluck Place* was purely choreography as the performer was instructed to circle the electric guitar and the log drum to explore their sounds. Composing for the AR electric guitar in *Push Pull* blurred the line between notating musical instrument technique like string bowing, and choreography because movements of the AR electric guitar player were programmed to affect the sounds of the XR musical instruments. The AR electric guitar player's movements were specified in the score to synchronize the spatial audio effects with the percussion.

The concert illuminated difficulties in exploring sound source directivity. To perceive the directivity of the guitar and log drum in *Between a Log and a Pluck Place*, the performer would have had to move around the instruments in a perfect circle because the instrument's volume was a function of both directivity and distance. In *Push Pull*, the distance dependent behavior of directivity resulted in changes between the time the concept was originally tested using headphones, and the final performance over an octophonic loudspeaker layout because the drumset's directivity was lost when the listener was placed farther away. Using the IEM DirectivityShaper plugin in combination with the hybrid spatial reverb did not provide the amount of sound expansion/contraction envisioned due to the reduced loudness the listener's position was placed at the location of the audience. The listener's position in the initial design of the spatial compression effect was at the drummer's seat placing the drums roughly a meter away from the listener. For the performance, the pushing and pulling effect of the drumset was achieved by shifting the virtual location of the drumset based on the position of the guitar. In the outer area of interaction, the guitar pulls the drumset by setting the position of the drumset equal to that of the guitar player. In the inner area of interaction, where the guitar repels the drumset, the position of the drumset was moved 180^{o} from the guitar with the drummer's seat as the origin. The notes

for the AR electric guitar part were strictly notated, but it was difficult to do the same for the performer's position and rotation. It was left up to the performer to choose the speed and timing of movement. Methods for notating movement were out of the scope of this research[3].

[3]Please refer to [135] for an AR example of 6DoF notation.

Chapter 6

Conclusion

The XR concert, *Spherical Sound Search*, demonstrated advanced musical expression using a novel spatial audio reverb system with re-contextualized spatial audio effects on author-created XR musical instruments. Spatial looping was essential in creating an orchestra from a single kite sample for *VR Singing Kite Concerto*. In the MR piece, *Between a Log and a Pluck Place*, spatial looping enabled the superposition of electric guitar and log drum samples with their corresponding instruments. Spatial feedback illuminated the kite solo in *VR Singing Kite Concerto* and created the desired sense of tension between the AR electric guitar and the AR electric drumset in *Push Pull*.

The concert also highlighted challenges of integrating spatial audio into XR musical instrument design. The spatial compression effect was intended to showcase position-based changes in source directivity, but did not perform as well in the venue when compared to the rehearsal space as the effect was tuned in the rehearsal space where the listening position was close to the sound-emitting object. That distance between listener and sound source increased in the concert venue such that the lower gain (due to distance attenuation) made it difficult to perceive changes in source directivity. Modifications to the spatial compression effect are required to perform well in future 6DoF music compositions. Another challenge that arose in preparation

for the concert was related to compute power. Convolution reverb requires a significant amount of CPU. To prevent audio artifacts due to maxing out the CPU[1], the three pieces were separated into three MaxMSP patches and the buffer size was increased to 512 samples. For this performance the drummer had to consciously compensate their timing to account for the latency due to the buffer size and the latency of triggering the drum samples. Dedicated hardware for each XR musical instrument, similar to what has been implemented in other smart instruments [147], could mitigate such inconveniences in future performances.

Improvements to the hybrid spatial reverb created for this research are left for future work. The current hybrid spatial reverb system was not evaluated under situations where the source and listener moved away from the positions where the DRIR was measured. Conditions in the pilot test constrained the listener's position to the drummer's seat when evaluating the naturalness of the AR electric guitar and AR electric drumset. Conditions in testing with a 6DoF VR musical instrument were not as constrained as the listeners were allowed to move freely, however, it was observed that the listeners chose to stay in one place. As such, a corner case scenario, where the listener and sound source are displaced to opposite sides of the room, was not tested. The participants were not given precise instructions on where to move because it was thought that freedom of motion would result in a more realistic experience and allow the participants to focus on the task of evaluating naturalness. In the XR concert setting, the audience remained in a fixed position which afforded the use of the hybrid spatial reverb system without modifications. Improvements to spatial reverb will become more important when XR HMDs are used more commonly and the audience expects personalized binaural audio.

A formal study of spatial reverb would illuminate the role that spatial reverb plays in the naturalness of XR musical instruments. Several of the seven spatial reverbs compared in the pilot study were perceived differently when processing different frequencies. For example, some sounded natural when playing the crash which produces higher frequencies, but not when playing

[1] On a 3.6 GHz Intel Core i3.

the kick which produces lower frequencies and vice versa. The drumset was evaluated as a whole in the pilot test, but should be broken down into each individual drum to provide more accurate results on the listener's frequency-dependent preference. The formal study should also include more auralization models and update the raytracing model to replicate the room under test with as much detail as possible.

As XR technologies become more widely adopted, it will be exciting to see how improvements in spatial audio reverb, like the one presented in this dissertation, push the musicality of XR musical instruments beyond the video game market and onto the concert stage.

Appendix A

DRIR Plots

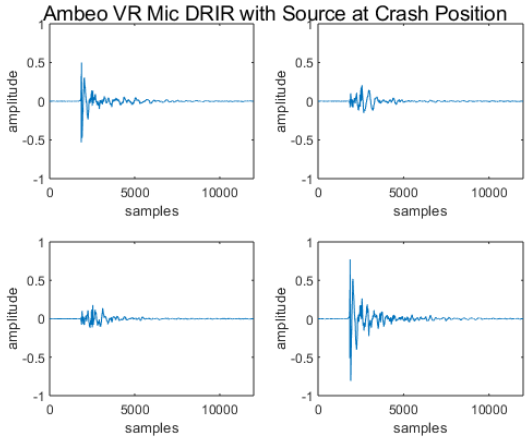


Figure A.1: DRIR recorded by Ambeo VR Mic with the source at the position of the crash. Top left is the omni channel, top right is the X channel, bottom left is the Y channel and bottom right is the Z channel.

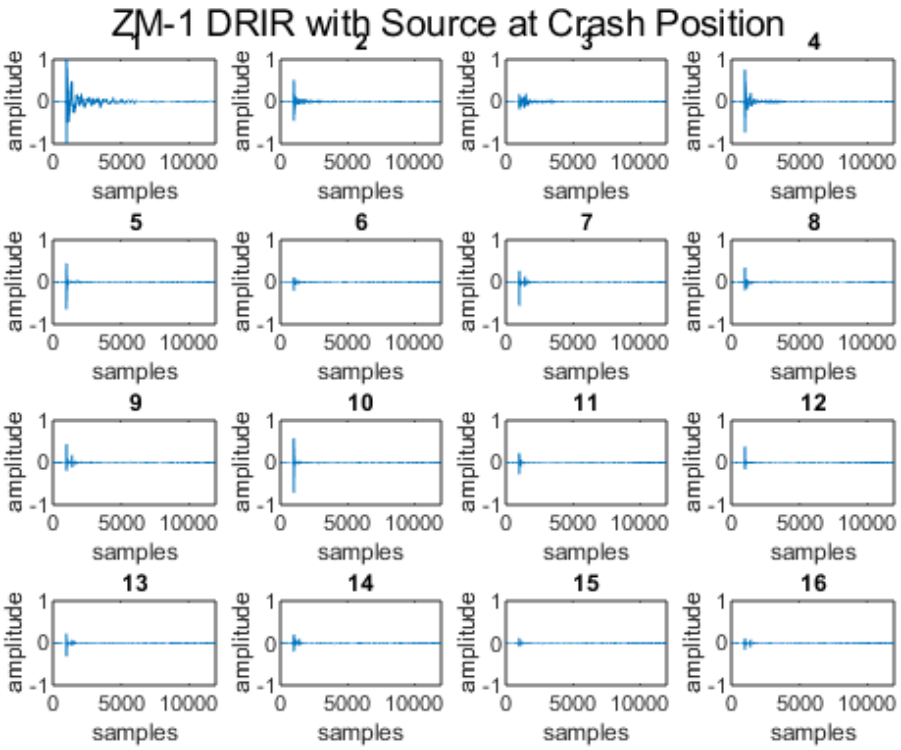


Figure A.2: DRIR recorded by the ZM-1 with the source at the position of the crash. Numbers 1 to 16 correspond to the Ambisonics Channel Numbering.

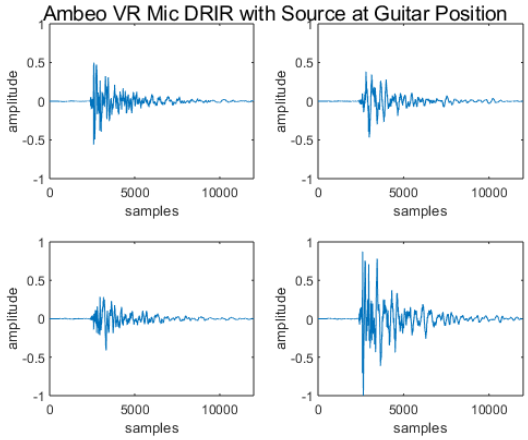


Figure A.3: DRIR recorded by Ambeo VR Mic at the drummer's seat and the source at the center of the room. Top left is the omni channel, top right is the X channel, bottom left is the Y channel and bottom right is the Z channel.

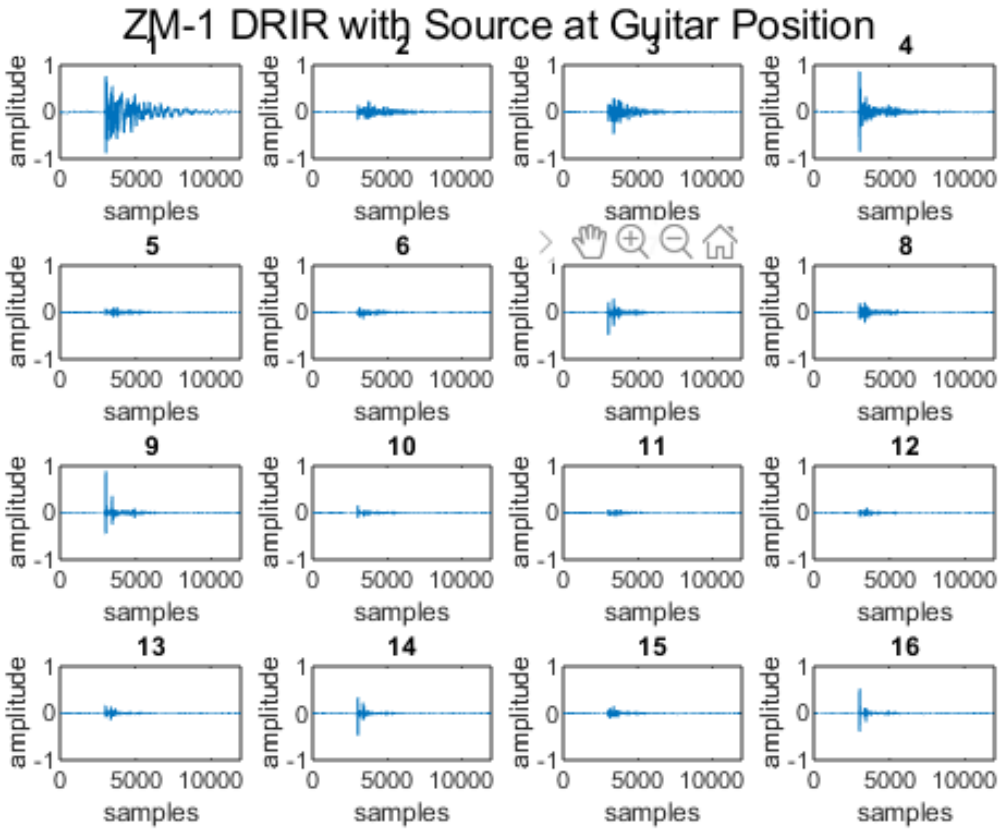


Figure A.4: DRIR recorded by the ZM-1 at the drummer's seat and the source at the center of the room. Numbers 1 to 16 correspond to the Ambisonics Channel Numbering.

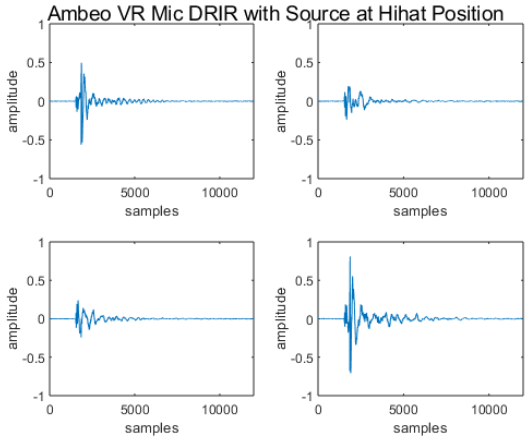


Figure A.5: DRIR recorded by Ambeo VR Mic with the source at the position of the hihat. Top left is the omni channel, top right is the X channel, bottom left is the Y channel and bottom right is the Z channel.

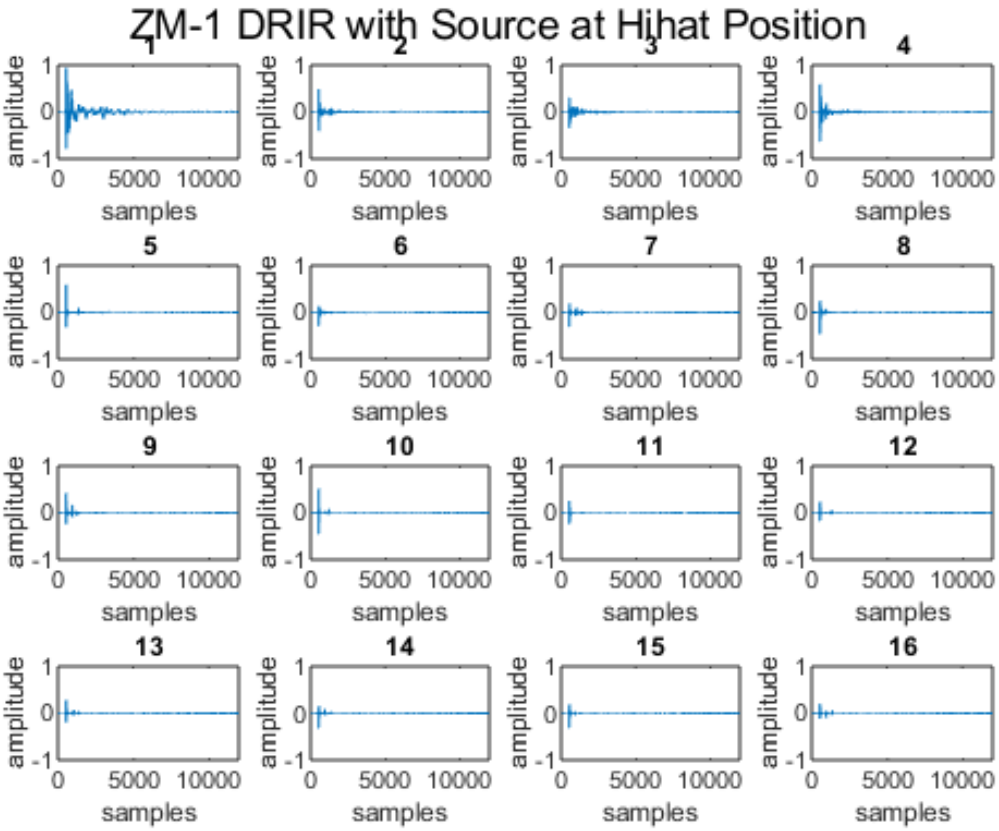


Figure A.6: DRIR recorded by the ZM-1 with the source at the position of the hihat. Numbers 1 to 16 correspond to the Ambisonics Channel Numbering.

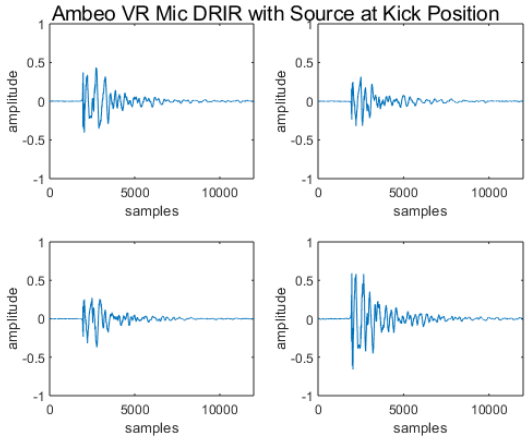


Figure A.7: DRIR recorded by Ambeo VR Mic with the source at the position of the kick. Top left is the omni channel, top right is the X channel, bottom left is the Y channel and bottom right is the Z channel.

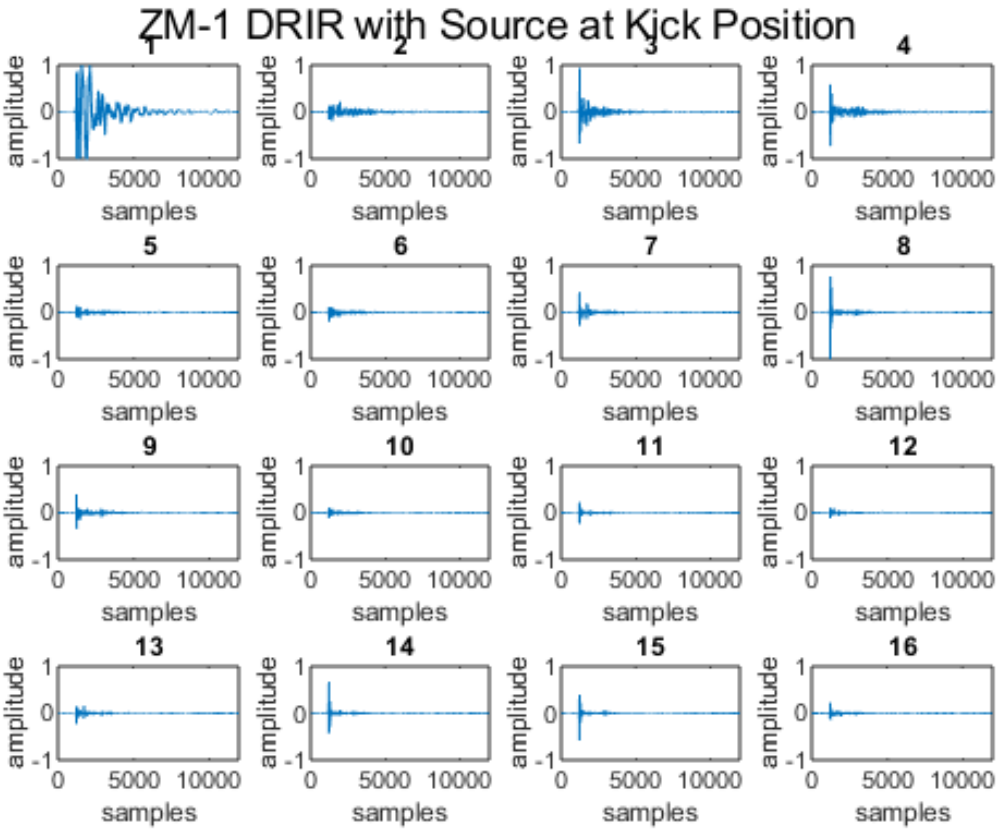


Figure A.8: DRIR recorded by the ZM-1 with the source at the position of the kick. Numbers 1 to 16 correspond to the Ambisonics Channel Numbering.

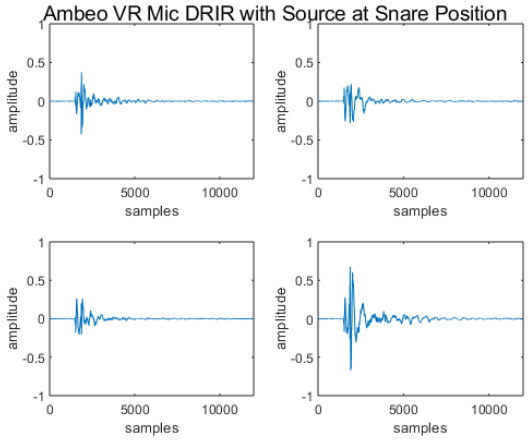


Figure A.9: DRIR recorded by Ambeo VR Mic with the source at the position of the snare. Top left is the omni channel, top right is the X channel, bottom left is the Y channel and bottom right is the Z channel.

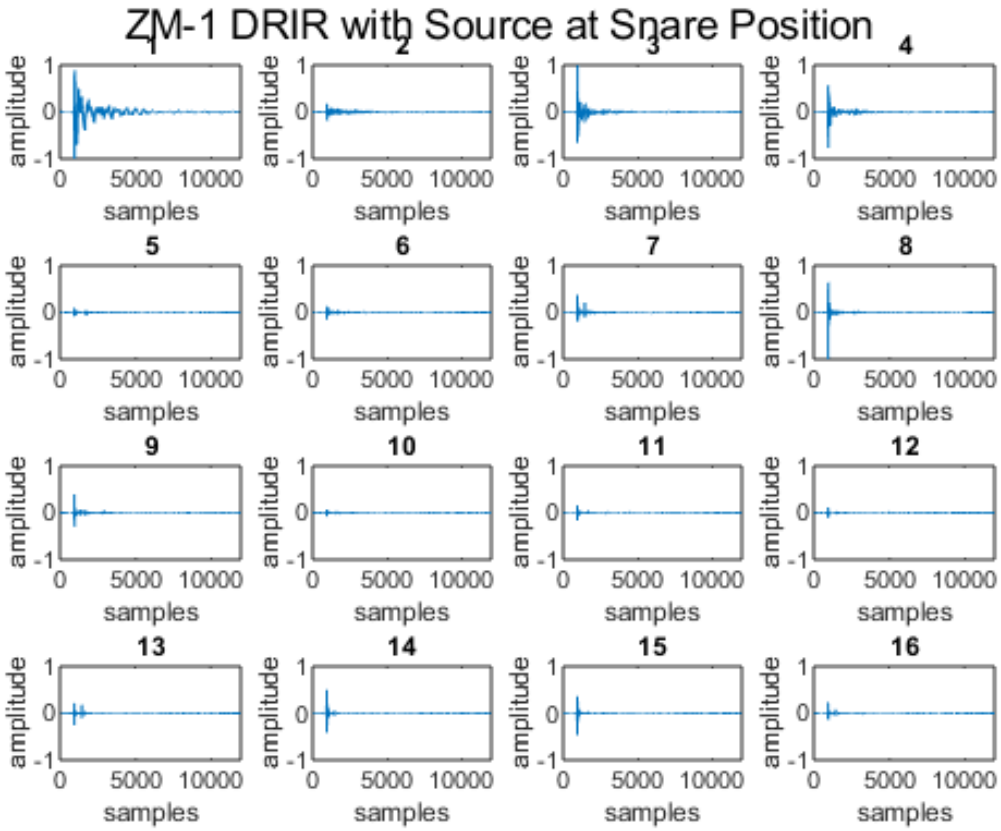


Figure A.10: DRIR recorded by the ZM-1 with the source at the position of the snare. Numbers 1 to 16 correspond to the Ambisonics Channel Numbering.

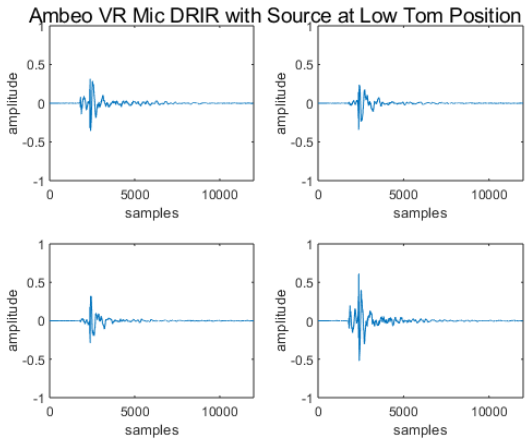


Figure A.11: DRIR recorded by Ambeo VR Mic with the source at the position of the low tom. Top left is the omni channel, top right is the X channel, bottom left is the Y channel and bottom right is the Z channel.

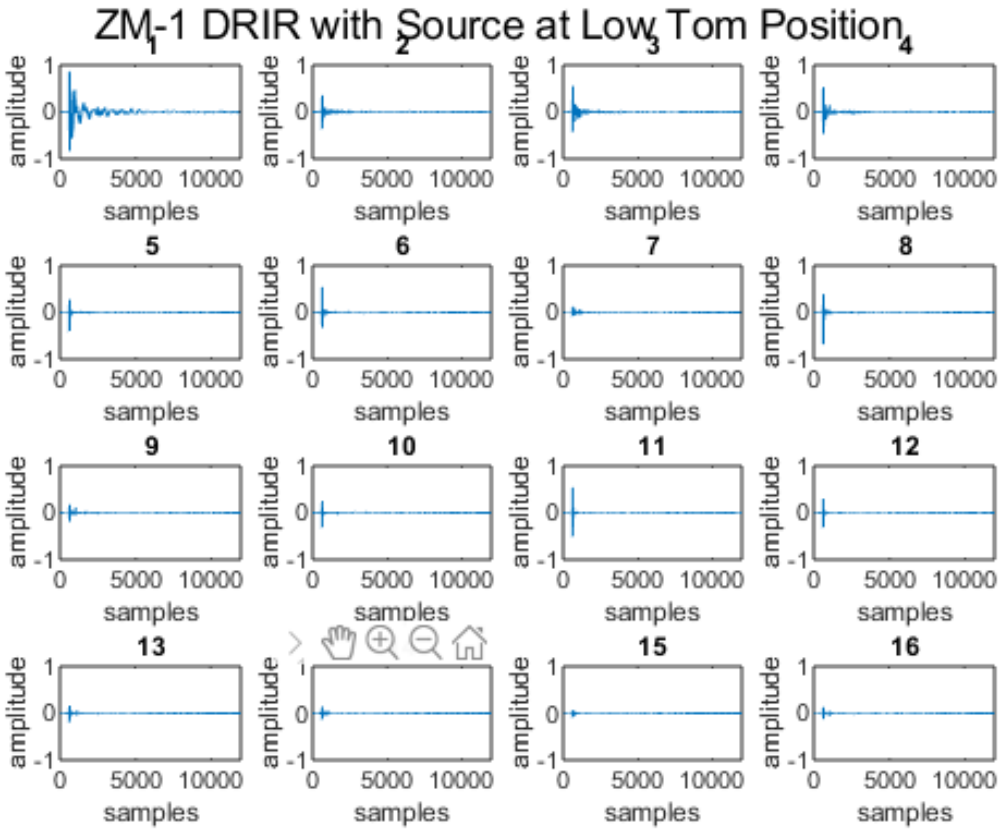
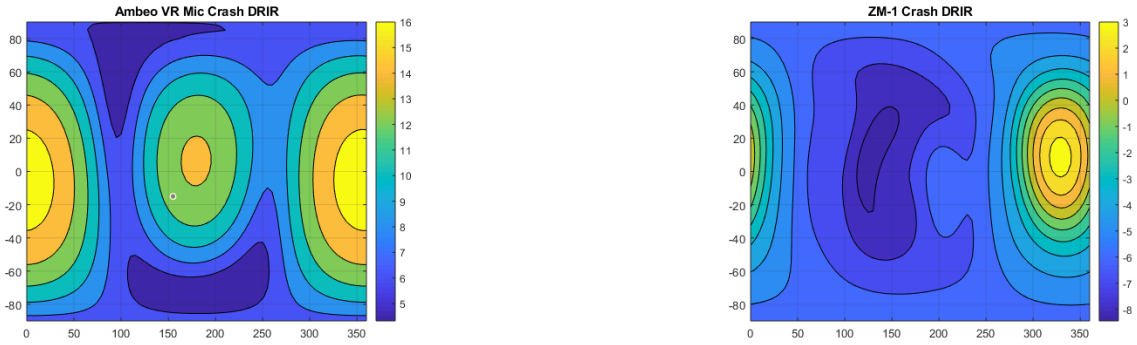


Figure A.12: DRIR recorded by the ZM-1 with the source at the position of the low tom. Numbers 1 to 16 correspond to the Ambisonics Channel Numbering.

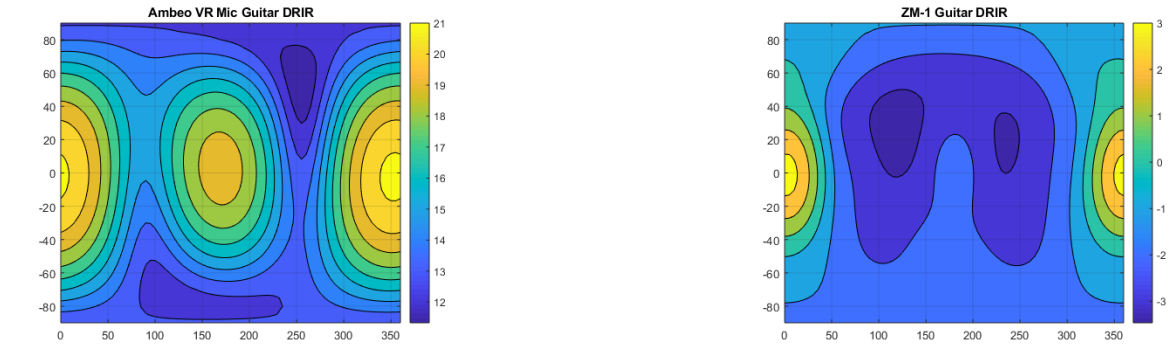
Appendix B

Energy Plots of Measured DRIRs



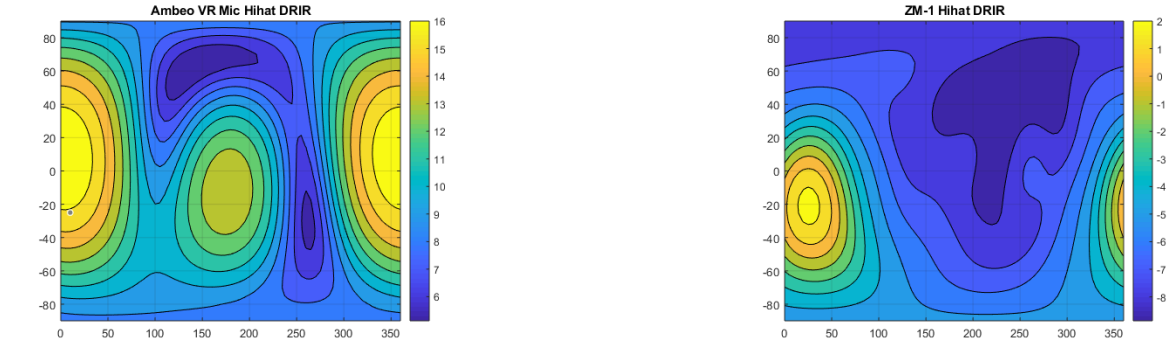
(a) (b)

Figure B.1: Energy plots of Ambeo VR Mic and ZM-1 for the DRIR source at the crash position. The x-axis is azimuth where 0&360 are facing forward and 180 is facing back. The y-axis is elevation where 0 is facing forward, 90 is facing straight up and -90 is facing down at the ground. The color scale is energy where warmer colors indicate higher energy and cooler colors indicate lower energy.



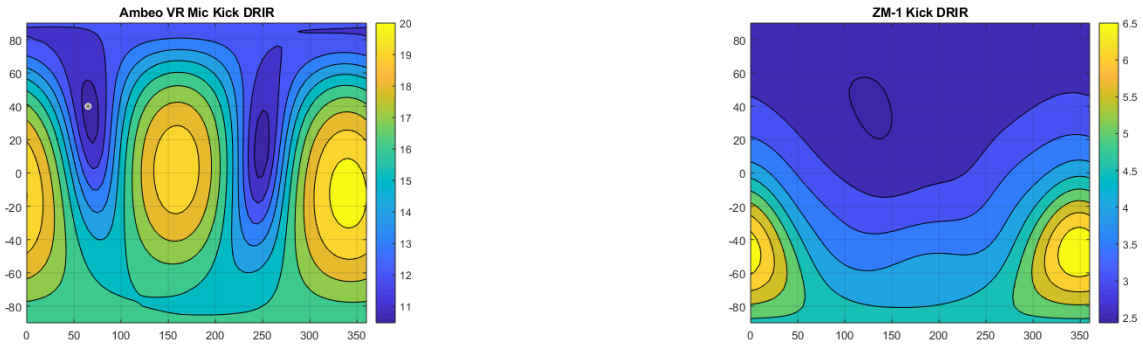
(a) (b)

Figure B.2: Energy plots of Ambeo VR Mic and ZM-1 for the DRIR source at the center of the test room. The x-axis is azimuth where 0&360 are facing forward and 180 is facing back. The y-axis is elevation where 0 is facing forward, 90 is facing straight up and -90 is facing down at the ground. The color scale is energy where warmer colors indicate higher energy and cooler colors indicate lower energy.



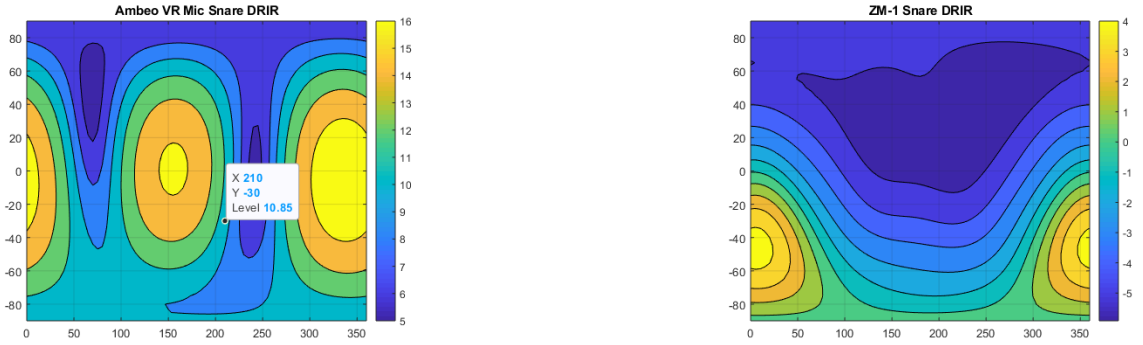
(a) (b)

Figure B.3: Energy plots of Ambeo VR Mic and ZM-1 for the DRIR source at the hihat position. The x-axis is azimuth where 0&360 are facing forward and 180 is facing back. The y-axis is elevation where 0 is facing forward, 90 is facing straight up and -90 is facing down at the ground. The color scale is energy where warmer colors indicate higher energy and cooler colors indicate lower energy.



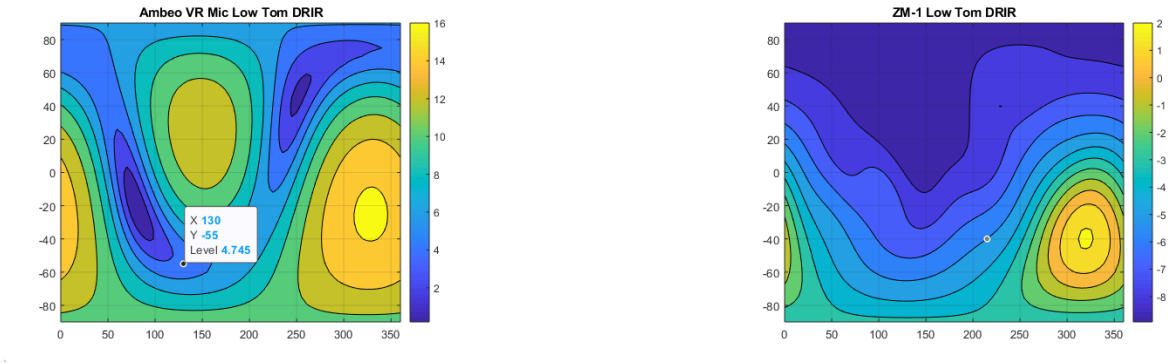
(a) (b)

Figure B.4: Energy plots of Ambeo VR Mic and ZM-1 for the DRIR source at the kick position. The x-axis is azimuth where 0&360 are facing forward and 180 is facing back. The y-axis is elevation where 0 is facing forward, 90 is facing straight up and -90 is facing down at the ground. The color scale is energy where warmer colors indicate higher energy and cooler colors indicate lower energy.



(a) (b)

Figure B.5: Energy plots of Ambeo VR Mic and ZM-1 for the DRIR source at the snare position. The x-axis is azimuth where 0&360 are facing forward and 180 is facing back. The y-axis is elevation where 0 is facing forward, 90 is facing straight up and -90 is facing down at the ground. The color scale is energy where warmer colors indicate higher energy and cooler colors indicate lower energy.



(a) (b)

Figure B.6: Energy plots of Ambeo VR Mic and ZM-1 for the DRIR source at the low tom position. The x-axis is azimuth where 0&360 are facing forward and 180 is facing back. The y-axis is elevation where 0 is facing forward, 90 is facing straight up and -90 is facing down at the ground. The color scale is energy where warmer colors indicate higher energy and cooler colors indicate lower energy.

Appendix C

Push Pull Score

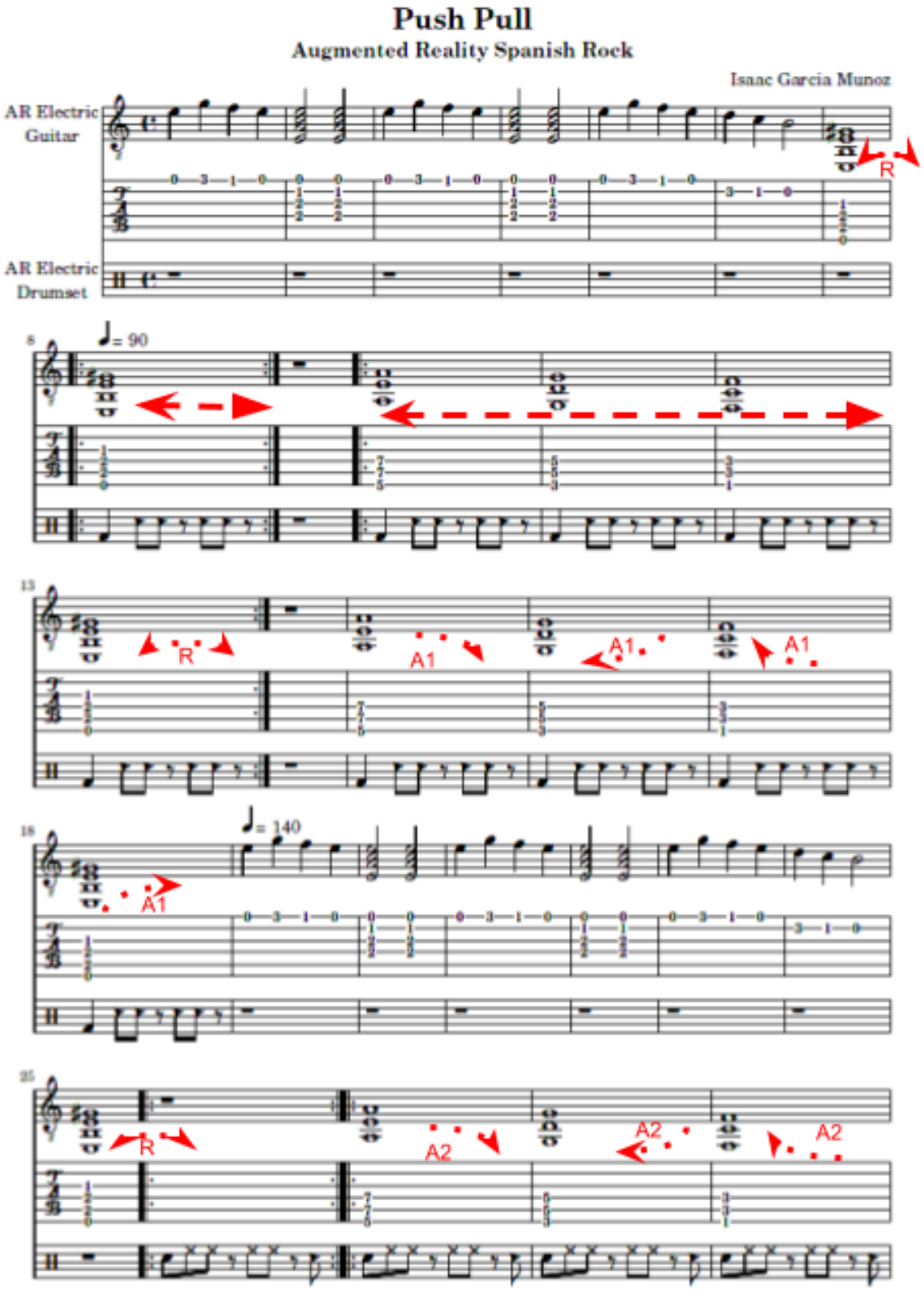


Figure C.1: Page 1 of the score for *Push Pull*.

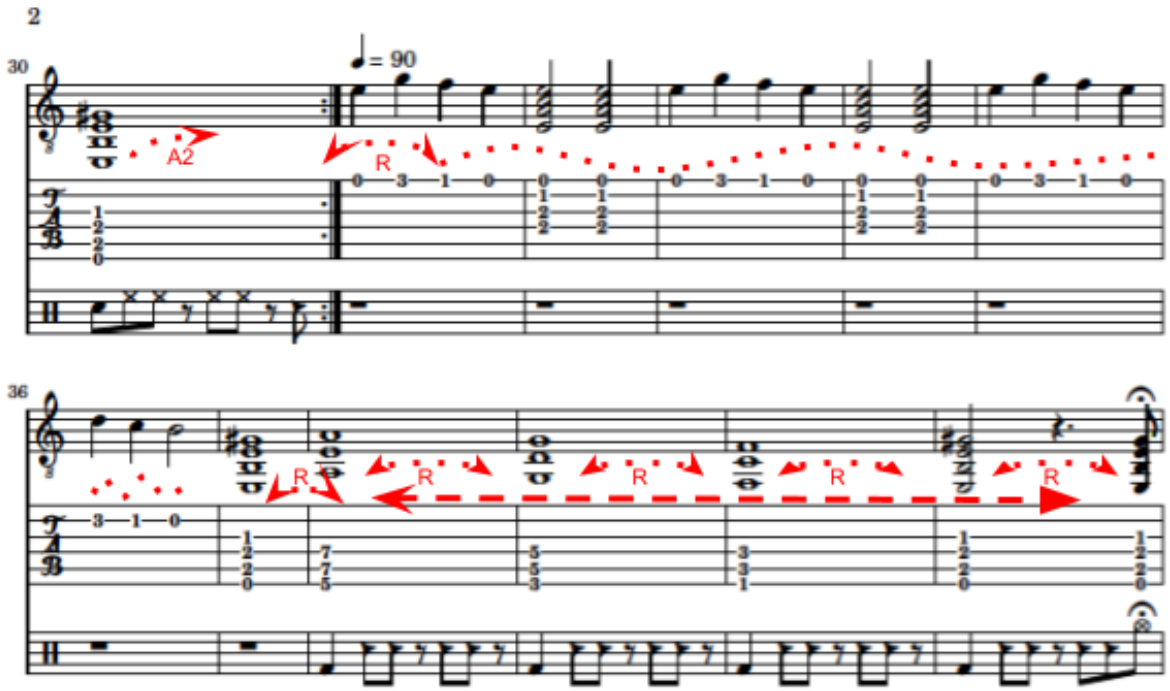


Figure C.2: Page 2 of the score for *Push Pull*.

Bibliography

[1] Aax. http://apps.avid.com/aax-portal/.

[2] Ableton push. https://www.ableton.com/en/push/.

[3] Ambeo a-b format converter. https://en-us.sennheiser.com/ambeo-abconverter.

[4] Ambeo vr mic. https://en-us.sennheiser.com/microphone-3d-audio-ambeo-vr-mic.

[5] Arcore for unity. https://github.com/google-ar/arcore-unity-sdk/releases.

[6] Arcore supported devices. https://developers.google.com/ar/discover/supported-devices.

[7] Beat saber. https://beatsaber.com/.

[8] Blender. https://www.blender.org/.

[9] Caffe. https://caffe.berkeleyvision.org/.

[10] Evertims. http://evertims.github.io/.

[11] Facebook. https://www.facebook.com/.

[12] Fear unsound ar. https://play.google.com/store/apps/details?id=com.floatingmusic.fearunsoundAR.

[13] Fmod. https://www.fmod.com/.

[14] Google glass. https://developers.google.com/glass/develop/gdk.

[15] Iem binauraldecoder plugin. https://plugins.iem.at/docs/plugindescriptions/#binauraldecoder.

[16] Iem directivityshaper plugin. https://plugins.iem.at/docs/directivityshaper/.

[17] Iem plugin descriptions. https://plugins.iem.at/docs/plugindescriptions/.

[18] Iem plugins. https://plugins.iem.at/.

[19] Iem roomencoder plugin. https://plugins.iem.at/#tab-RoomEncoder.

[20] Ir measurement toolbox. https://cycling74.com/projects/ir-measurement-toolbox.

[21] Leap motion. https://www.leapmotion.com/.

[22] Listen inc. soundcheck. https://www.listeninc.com/products/soundcheck-software/measurements/.

[23] Maxmsp. https://cycling74.com/products/max/.

[24] Microsoft hololens. https://www.microsoft.com/en-us/hololens.

[25] Neumann ku 100. https://en-de.neumann.com/ku-100.

[26] Oculus quest. https://www.oculus.com/quest/.

[27] Oculus rift. https://www.oculus.com/rift-s/.

[28] Oculus spatializer. https://developer.oculus.com/audio/.

[29] Project acoustics. https://docs.microsoft.com/en-us/azure/cognitive-services/acoustics/what-is-acoustics.

[30] Roli seaboard. https://roli.com/products/seaboard.

[31] Snapchat. https://www.snapchat.com/.

[32] Sound and violence.

[33] Soundparticles. https://soundparticles.com/.

[34] Space 3d. https://sonicarts.ucsd.edu/research/space-3d.html.

[35] Spatial audio designer. https://newaudiotechnology.com/products/spatial-audio-designer/.

[36] Steam audio. https://valvesoftware.github.io/steam-audio/.

[37] Steinberg vst. https://www.steinberg.net/en/products/vst.html.

[38] Tensorflow. https://www.tensorflow.org/.

[39] Unity3d. https://unity.com/.

[40] Unreal. https://www.unrealengine.com/en-US/.

[41] https://resonance-audio.github.io/resonance-audio/.

[42] Vuforia. https://www.vuforia.com.

[43] Wwise. https://www.audiokinetic.com/products/wwise/.

[44] Zylia 6dof sdk. https://www.zylia.co/zylia-6dof.html.

[45] Zylia ambisonics converter. https://www.zylia.co/zylia-ambisonics-converter.html.

[46] Zylia zm-1 microphone. https://www.zylia.co/zylia-zm-1-microphone.html.

[47] 3382-1, I. Acoustics—measurement of room acoustic parameters—part 1: Performance spaces, 2009.

[48] ABEL, J. S., BRYAN, N. J., HUANG, P. P., KOLAR, M., AND PENTCHEVA, B. V. Estimating room impulse responses from recorded balloon pops. In *Audio Engineering Society Convention 129* (2010), Audio Engineering Society.

[49] ACKERMANN, D., ILSE, M., GRIGORIEV, D., LEPA, S., PELZER, S., VORLÄNDER, M., AND WEINZIERL, S. A ground truth on room acoustical analysis and perception (grap).

[50] AHRENS, J., AND SPORS, S. An analytical approach to sound field reproduction with a movable sweet spot using circular distributions of loudspeakers. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing* (2009), IEEE, pp. 273–276.

[51] ALGAZI, V. R., DUDA, R. O., THOMPSON, D. M., AND AVENDANO, C. The cipic hrtf database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)* (2001), IEEE, pp. 99–102.

[52] BACILA, B. I., AND LEE, H. 360° binaural room impulse response (brir) database for 6dof spatial perception research. In *Audio Engineering Society Convention 146* (2019), Audio Engineering Society.

[53] BAUMGARTNER, R. Time domain fast-multipole translation for ambisonics. *Master's thesis, University of Music and Performing Arts-Institute of Electronic Music and Acoustics, Inst. Electron. Music Acoust., Graz 8010* (2011).

[54] BEGAULT, D. R., AND TREJO, L. J. 3-d sound for virtual reality and multimedia. *NASA* (2000).

[55] BERDAHL, E., NIEMEYER, G., AND SMITH, J. O. Using haptics to assist performers in making gestures to a musical instrument. In *NIME* (2009), pp. 177–182.

[56] BERDAHL, E., STEINER, H.-C., AND OLDHAM, C. Practical hardware and algorithms for creating haptic musical instruments. In *NIME* (2008), vol. 8, pp. 61–66.

[57] BERNSCHÜTZ, B., PÖRSCHMANN, C., SPORS, S., AND WEINZIERL, S. Sofia sound field analysis toolbox. In *Proceedings of the International Conference on Spatial Audio (ICSA)* (2011), pp. 7–15.

[58] BROWN, C. t60.m. Tech. rep., MATLAB Central File Exchange., Retrieved February 21, 2020. 2020.

[59] BRUNGART, D. S., SIMPSON, B. D., MCKINLEY, R. L., KORDIK, A. J., DALLMAN, R. C., AND OVENSHIRE, D. A. The interaction between head-tracker latency, source duration, and response time in the localization of virtual sound sources. In *Proceedings of ICAD 04- Tenth Meeting of the International Conference on Auditory Display* (2004), Georgia Institute of Technology.

[60] CAKMAKCI, O., BÉRARD, F., AND COUTAZ, J. An augmented reality based learning assistant for electric bass guitar. In *Proc. of the 10th International Conference on Human-Computer Interaction, Crete, Greece* (2003).

[61] CARPENTIER, T., NOISTERNIG, M., AND WARUSFEL, O. Twenty years of ircam spat: looking back, looking forward.

[62] CHENG, C. I., AND WAKEFIELD, G. H. Introduction to head-related transfer functions (hrtfs): Representations of hrtfs in time, frequency, and space. In *Audio Engineering Society Convention 107* (1999), Audio Engineering Society.

[63] CHENG, W. *Sound Play: Video Games and the Musical Imagination*. Oxford University Press, 2014.

[64] COOK, P. R. *Music, cognition, and computerized sound: An introduction to psychoacoustics*. The MIT press, 1999.

[65] CORREA, A. G. D., DE ASSIS, G. A., DO NASCIMENTO, M., FICHEMAN, I., AND DE DEUS LOPES, R. Genvirtual: An augmented reality musical game for cognitive and motor rehabilitation. In *Virtual Rehabilitation, 2007* (2007), IEEE, pp. 1–6.

[66] CUEVAS-RODRÍGUEZ, M., PICINALI, L., GONZÁLEZ-TOLEDO, D., GARRE, C., DE LA RUBIA-CUESTAS, E., MOLINA-TANCO, L., AND REYES-LECUONA, A. 3d tune-in toolkit: An open-source library for real-time binaural spatialisation. *PloS one 14*, 3 (2019).

[67] DAVIS, G., ANDRE, S., MUNOZ, I., AND PETERS, N. Perceptual evaluation of personalized brirs and headphone compensation. In *Audio Engineering Society Conference: 2019 AES INTERNATIONAL CONFERENCE ON HEADPHONE TECHNOLOGY* (2019), Audio Engineering Society.

[68] DEAN, R. T., ET AL. *The Oxford handbook of computer music*. OUP USA, 2009.

[69] DENG, X., AND TANG, Z.-A. Moving surface spline interpolation based on green's function. *Mathematical geosciences 43*, 6 (2011), 663–680.

[70] ETSI. Speech and multimedia transmission quality (stq); methods for reproducing reverberation for communication device measurements. https://www.etsi.org/deliver/etsi_ts/103500_103599/103557/01.01.01_60/ts_103557v010101p.pdf, 2018.

[71] EYRING, C. F. Reverberation time in "dead" rooms. *The Journal of the Acoustical Society of America 1*, 2A (1930), 217–241.

[72] FARINA, A. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Audio Engineering Society Convention 108* (2000), Audio Engineering Society.

[73] FLETCHER, H., AND MUNSON, W. A. Loudness, its definition, measurement and calculation. *Bell System Technical Journal 12*, 4 (1933), 377–430.

[74] FLOATINGMUSIC. Fear unsound. Google Play Store, March 2017.

[75] GARCIA-MUNOZ, I. Transforming the teponaztli. *UCSD Master's Thesis* (2013).

[76] GARDNER, W. G. Efficient convolution without input/output delay. In *Audio Engineering Society Convention 97* (1994), Audio Engineering Society.

[77] GARDNER, W. G. *Transaural 3-D audio*. Citeseer, 1995.

[78] GARDNER, W. G. 3d audio and acoustic environment modeling. *Wave Arts, Inc 99* (1999).

[79] GORZEL, M., ALLEN, A., KELLY, I., KAMMERL, J., GUNGORMUSLER, A., YEH, H., AND BOLAND, F. Efficient encoding and decoding of binaural sound with resonance audio. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio* (Mar 2019).

[80] GOULD, R., AND LANE, C. Let's test: 3d audio spatialization plugins. *Designing Sound. The Art & Technique of Sound Design* (2018).

[81] GRIESINGER, D. Explanation of clarity versus reverberation in concert acoustics. *The Boston Musical Intelligencer* (2010).

[82] GROUP, K. Opensl es. https://www.khronos.org/opensles/, 2020.

[83] GUPTA, R., RANJAN, R., HE, J., AND WOON-SENG, G. On the use of closed-back headphones for active hear-through equalization in augmented reality applications. In *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality* (2018), Audio Engineering Society.

[84] HARKER, A., AND TREMBLAY, P. A. The hisstools impulse response toolbox: Convolution for the masses. In *Proceedings of the international computer music conference* (2012), The International Computer Music Association, pp. 148–155.

[85] HARMA, A., JAKKA, J., TIKANDER, M., KARJALAINEN, M., LOKKI, T., AND NIRONEN, H. Techniques and applications of wearable augmented reality audio. In *Audio Engineering Society Convention 114* (2003), Audio Engineering Society.

[86] HELLER, A. J., BENJAMIN, E. M., AND LEE, R. A toolkit for the design of ambisonic decoders. In *Linux Audio Conference* (2012), pp. 1–12.

[87] HENDERSON, D. M. Euler angles, quaternions, and transformation matrices for space shuttle analysis.

[88] HERRE, J., ET AL. From joint stereo to spatial audio coding-recent progress and standardization. In *Sixth International Conference on Digital Audio Effects (DAFX04), Naples, Italy* (2004).

[89] HIEBERT, G., GORDON, R., PEACOCK, D., AND VOGELSANG, C. Openal 1.1 specification and reference. *Creative Labs* (2006).

[90] HUANG, W., ALEM, L., AND LIVINGSTON, M. A. *Human factors in augmented reality environments*. Springer Science & Business Media, 2012.

[91] INSTITUTE OF TECHNICAL ACOUSTICS, RWTH AACHEN UNIVERSITY. Virtual Acoustics - A real-time auralization framework for scientific research. http://www.virtualacoustics.org/, 2018. Accessed on 2020-02-09.

[92] IRCAM. Spat. https://forum.ircam.fr/projects/detail/spat/.

[93] JARRETT, D., HABETS, E., THOMAS, M., AND NAYLOR, P. Rigid sphere room impulse response simulation: Algorithm and applications. *The Journal of the Acoustical Society of America 132*, 3 (2012), 1462–1472.

[94] JORDÀ, S. The reactable: tangible and tabletop music performance. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. 2010, pp. 2989–2994.

[95] JOT, J.-M., AND LEE, K. S. Augmented reality headphone environment rendering. In *Audio Engineering Society Conference: 2016 AES International Conference on Audio for Virtual and Augmented Reality* (2016), Audio Engineering Society.

[96] KALTENBRUNNER, M., JORD, S., GEIGER, G., AND BENCINA, R. The reactable. In *Proc. ICMC* (2005), vol. 5, pp. 579–582.

[97] KAYSER, H., EWERT, S. D., ANEMÜLLER, J., ROHDENBURG, T., HOHMANN, V., AND KOLLMEIER, B. Database of multichannel in-ear and behind-the-ear head-related and binaural room impulse responses. *EURASIP Journal on Advances in Signal Processing 2009* (2009), 6.

[98] KIEFER, C., AND CHEVALIER, C. Towards new modes of collective musical expression through audio augmented reality. In *Proceedings of the International Conference on New Interfaces for Musical Expression* (2018).

[99] KOLARIK, A. J., MOORE, B. C., ZAHORIK, P., CIRSTEA, S., AND PARDHAN, S. Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, & Psychophysics 78*, 2 (2016), 373–395.

[100] KRUIJFF, E., SWAN, J. E., AND FEINER, S. Perceptual issues in augmented reality revisited. In *Mixed and Augmented Reality (ISMAR), 2010 9th IEEE International Symposium on* (2010), IEEE, pp. 3–12.

[101] KRYTER, K. D. Noise and behavior. *Journal of Speech and Hearing Disorders 15* (1950).

[102] LINDEMAN, R. W., NOMA, H., AND DE BARROS, P. G. Hear-through and mic-through augmented reality: Using bone conduction to display spatialized audio. In *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality* (2007), IEEE Computer Society, pp. 1–4.

[103] LUCIER, A. I am sitting in a room (1969). *UBUWEB: SOUND website, accessed 11th January* (2015).

[104] MENZIES, D., AND AL-AKAIDI, M. Nearfield binaural synthesis and ambisonics. *The Journal of the Acoustical Society of America 121*, 3 (2007), 1559–1563.

[105] MERGING TECHNOLOGIES INC. *Pyramix Digital Audio Workstation User Manual*, 2019.

[106] MERIMAA, J., AND PULKKI, V. Spatial impulse response rendering i: Analysis and synthesis. *Journal of the Audio Engineering Society 53*, 12 (2005), 1115–1127.

[107] MOORE, F. R. A general model for spatial processing of sounds. *Computer Music Journal 7*, 3 (1983), 6–15.

[108] MOTOKAWA, Y., AND SAITO, H. Support system for guitar playing using augmented reality display. In *Mixed and Augmented Reality, 2006. ISMAR 2006. IEEE/ACM International Symposium on* (2006), IEEE, pp. 243–244.

[109] MOUSTAKAS, N., FLOROS, A., AND GRIGORIOU, N. Interactive audio realities: An augmented/mixed reality audio game prototype. In *Audio Engineering Society Convention 130* (2011), Audio Engineering Society.

[110] MUNOZ, I. G. Cumbia leap. https://www.youtube.com/watch?v=MAjY42Oafdg, March 2013.

[111] MUNOZ, I. G. Between a Log and a Pluck Place – Mixed Reality Minuet. In Department of Music Concert Archive. UC San Diego Library Digital Collections. https://doi.org/10.6075/J0DZ06P8, 2020.

[112] MUNOZ, I. G. Push Pull – Augmented Reality Rock. In Department of Music Concert Archive. UC San Diego Library Digital Collections. https://doi.org/10.6075/J0959FZ9, 2020.

[113] MUNOZ, I. G. Spherical sound search maxmsp patches. https://github.com/floating-music/SphericalSoundSearch, February 2020.

[114] MUNOZ, I. G. VR Singing Kite Concerto – Virtual Reality Concerto. In Department of Music Concert Archive. UC San Diego Library Digital Collections. https://doi.org/10.6075/J0JM2814, 2020.

[115] NOISTERNIG, M., KATZ, B. F., SILTANEN, S., AND SAVIOJA, L. Framework for real-time auralization in architectural acoustics. *Acta Acustica United with Acustica 94*, 6 (2008), 1000–1015.

[116] OCULUS. Thumbs up: Hand tracking on oculus quest this week. https://www.oculus.com/blog/thumbs-up-hand-tracking-now-available-on-oculus-quest/, December 2019.

[117] OLIVER, J., AND JENKINS, M. The silent drum controller: a new percussive gestural interface. In *ICMC* (2008).

[118] OLIVIERI, F., PETERS, N., AND SEN, D. Scene-based audio and higher order ambisonics: A technology overview and application to next-generation audio, vr and 360 video. *EBU Tech* (2019).

[119] PAPAYIANNIS, C. Models for learning reverberant environments.

[120] PARSEIHIAN, G., KATZ, B. F., AND CONAN, S. Sound effect metaphors for near field distance sonification. Georgia Institute of Technology.

[121] PATRICIO, E., RUMINSKI, A., KUKLASINSKI, A., JANUSZKIEWICZ, L., AND ZERNICKI, T. Toward six degrees of freedom audio recording and playback using multiple ambisonics sound fields. In *Audio Engineering Society Convention 146* (2019), Audio Engineering Society.

[122] PEREZ-LOPEZ, A. ambiscaper documentation, 2018.

[123] PERRETT, S., AND NOBLE, W. The contribution of head motion cues to localization of low-pass noise. *Perception & psychophysics 59*, 7 (1997), 1018–1026.

[124] PETERS, N., SEN, D., KIM, M.-Y., WUEBBOLT, O., AND WEISS, S. M. Scene-based audio implemented with higher order ambisonics (hoa). In *SMPTE 2015 Annual Technical Conference and Exhibition* (2015), SMPTE, pp. 1–13.

[125] PETERS, N., SEN, D., KIM, M.-Y., WUEBBOLT, O., AND WEISS, S. M. Scene-based audio implemented with higher order ambisonics. *SMPTE Motion Imaging Journal 125*, 9 (2016), 16–24.

[126] PINCHBECK, D. Shock, horror: First-person gaming, horror, and the art of ludic manipulation. *Horror Video Games: Essays on the Fusion of Fear and Play* (2009), 79–94.

[127] POIRIER-QUINOT, D., KATZ, B., AND NOISTERNIG, M. Evertims: Open source framework for real-time auralization in architectural acoustics and virtual reality.

[128] POLITIS, A., ET AL. Microphone array processing for parametric spatial audio techniques.

[129] POUPYREV, I., BERRY, R., BILLINGHURST, M., KATO, H., NAKAO, K., BALDWIN, L., KURUMISAWA, J., ET AL. Augmented reality interface for electronic music performance. In *Proceedings of HCI* (2001), pp. 805–808.

[130] POUPYREV, I., BERRY, R., KURUMISAWA, J., NAKAO, K., BILLINGHURST, M., AIROLA, C., KATO, H., YONEZAWA, T., AND BALDWIN, L. Augmented groove: Collaborative jamming in augmented reality. In *ACM SIGGRAPH 2000 Conference Abstracts and Applications* (2000), p. 77.

[131] PULKKI, V., AND MERIMAA, J. Spatial impulse response rendering ii: Reproduction of diffuse sound and listening tests. *Journal of the Audio Engineering Society 54*, 1/2 (2006), 3–20.

[132] QUALCOMM. Qualcomm helps make your mobile devices smarter with new snapdragon machine learning software development kit. https://www.qualcomm.com/news/releases/2016/05/02/qualcomm-helps-make-your-mobile-devices-smarter-new-snapdragon-machine, May 2016.

[133] ROSSING, T. D., MOORE, F. R., WHEELER, P. A., AND ROSSING-MOORE-WHEELER... *The science of sound*, vol. 3. Addison Wesley San Francisco, 2002.

[134] RUBINE, D., AND MCAVINNEY, P. Programmable finger-tracking instrument controllers. *Computer music journal 14*, 1 (1990), 26–41.

[135] SANTINI, G. Linear (live-generated interface and notation environment in augmented reality). In *Proc. TENOR* (2018), pp. 33–42.

[136] SARUWATARI, H., KURITA, S., TAKEDA, K., ITAKURA, F., NISHIKAWA, T., AND SHIKANO, K. Blind source separation combining independent component analysis and beamforming. *EURASIP Journal on Advances in Signal Processing 2003*, 11 (2003), 569270.

[137] SCHÖRKHUBER, C., ZAUNSCHIRM, M., AND HÖLDRICH, R. Binaural rendering of ambisonic signals via magnitude least squares. In *Proceedings of the DAGA* (2018), vol. 44, pp. 339–342.

[138] SCHROEDER, M. R. New method of measuring reverberation time. *The Journal of the Acoustical Society of America 37*, 6 (1965), 1187–1188.

[139] SEN, D., PETERS, N., KIM, M., AND MORRELL, M. Efficient compression and transportation of scene-based audio for television broadcast. In *Audio Engineering Society Conference: 2016 AES International Conference on Sound Field Control* (2016), Audio Engineering Society.

[140] SERAFIN, S., ERKUT, C., KOJS, J., NILSSON, N. C., AND NORDAHL, R. Virtual reality musical instruments: State of the art, design principles, and future directions. *Computer Music Journal* (2016).

[141] SILVA, E. S., DE ABREU, J. A. O., DE ALMEIDA, J. H. P., TEICHRIEB, V., AND RAMALHO, G. L. A preliminary evaluation of the leap motion sensor as controller of new digital musical instruments. *Recife, Brasil* (2013).

[142] STADE, P., BERNSCHÜTZ, B., AND RÜHL, M. A spatial audio impulse response compilation captured at the wdr broadcast studios. In *Proc. of the VDT International Convention* (2012).

[143] STAN, G.-B., EMBRECHTS, J.-J., AND ARCHAMBEAU, D. Comparison of different impulse response measurement techniques. *Journal of the Audio Engineering Society 50*, 4 (2002), 249–262.

[144] SUDOL, J. D. Classical and contemporary cambodian music and dance. http://www.newmusicbox.org/articles/classical-and-contemporary-cambodian-music-and-dance/, May 2016.

[145] TACKETT, J. Iso 226 equal-loudness-level contour signal. MATLAB Central File Exchange. https://www.mathworks.com/matlabcentral/fileexchange/7028-iso-226-equal-loudness-level-contour-signal, February 2020.

[146] TEACHERS, T. Zylia zm-1 review. https://thevocaliststudio.com/zylia-zm-1-review/, October 2019.

[147] TURCHET, L., MCPHERSON, A., AND FISCHIONE, C. Smart instruments: Towards an ecosystem of interoperable devices connecting performers and audiences. In *Proceedings of the Sound and Music Computing Conference* (2016), pp. 498–505.

[148] TYLKA, J. G. *Virtual Navigation of Ambisonics-Encoded Sound Fields Containing Near-Field Sources*. PhD thesis, Princeton University, 2019.

[149] VENNERØD, J. The hard case - improving room acoustics in cuboid rooms by using diffusors: Scale model measurements.

[150] WALLACH, H., NEWMAN, E. B., AND ROSENZWEIG, M. R. A precedence effect in sound localization. *The Journal of the Acoustical Society of America 21*, 4 (1949), 468–468.

[151] XIANG, N. Evaluation of reverberation times using a nonlinear regression approach. *The Journal of the Acoustical Society of America 98*, 4 (1995), 2112–2121.

[152] ZAHORIK, P. Auditory display of sound source distance. In *Proc. Int. Conf. on Auditory Display* (2002), pp. 326–332.

[153] ZALLES, G., KAMEL, Y., ANDERSON, I., LEE, M. Y., NEIL, C., HENRY, M., CAPPIELLO, S., MYDLARZ, C., BAGLIONE, M., AND ROGINSKA, A. A low-cost, high-quality mems ambisonic microphone. In *Audio Engineering Society Convention 143* (2017), Audio Engineering Society.

[154] ZHAPAROV, M., AND ASSANOV, U. Augmented reality based on kazakh instrument" dombyra". In *Application of Information and Communication Technologies (AICT), 2014 IEEE 8th International Conference on* (2014), IEEE, pp. 1–4.

[155] ZOTTER, F., AND FRANK, M. All-round ambisonic panning and decoding. *Journal of the audio engineering society 60*, 10 (2012), 807–820.