UNIVERSITY OF CALIFORNIA SAN DIEGO

Creating, capturing and conveying spatial music: an open-source approach

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

in

Music

by

Gabriel Zalles Ballivian

Committee in charge:

Professor Shahrokh Yadegari, Chair
Professor Tom Erbe
Professor Justin Roberts
Professor Tamara Smyth
Professor Rand Steiger

2023

The dissertation of Gabriel Zalles Ballivian is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

DEDICATION

Para mi madre, quien amo de todo corazón.

EPIGRAPH

The music is in the making.

—Miller Puckette

TABLE OF CONTENTS

SUPPLEMENTAL FILES

1. zalles_6seasons.pdf

2. zalles_alternate_spaces_binaural.mp4

3. zalles_chaqu_binaural.m4v

4. zalles_HOA_array.pdf

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank my committee for all their guidance and wisdom, my family for supporting me, my friends at UCSD for their passion and curiosity, and all my other mentors such as Lei Liang, who made me feel at home. I also want to thank the Whale Acoustics Laboratory, the staff at the UCSD Department of Music, anybody responsible for maintaining the fabrication labs on campus, and the various scholars who have inspired my research path. I thank every single undergraduate student who believed in me and collaborated with me, I appreciate your support. Finally, I want to thank all the programmers and artists who share their code, documentation, media, and research with their world allowing me to learn, teach and adapt these materials.

VITA

2016	B.A. in Music, UC San Diego
2018	M.A. in Music, New York University
2023	Ph.D. in Music, UC San Diego

ABSTRACT OF THE DISSERTATION

Creating, capturing and conveying spatial music: an open-source approach

by

Gabriel Zalles Ballivian
Doctor of Philosophy in Music

University of California San Diego, 2023

Professor Shahrokh Yadegari, Chair

This dissertation deals with the topic of spatial audio in the context of Free and Open Source Software (FOSS). It is separated into three main sections: creating, capturing, and conveying spatial audio. One additional introductory chapter discusses the historical significance of this medium, as well as the psychoacoustic principles exploited. The final chapter summarizes the contributions, motivations, and outline of this dissertation presented in Chapter 1. During my graduate education in New York, I was first exposed to spatial audio as a research topic. During my time in San Diego as a Ph.D. student, I was exposed to the world of open-source software. This dissertation is my attempt at marrying both worlds, describing a framework and methodology rooted in social consciousness and scientific principles for art-making and research. Part of my motivation for writing this dissertation stems from my background as a citizen of an underdeveloped nation: Bolivia. I hope this manuscript can serve as a roadmap for artists and researchers in underprivileged communities interested in the field of spatial audio, and in particular ambisonics.

Chapter 1

Introduction

1.1 Motivation

The motivation behind this dissertation is quite simple: we would like to demonstrate how the socioeconomic barriers in the field of spatial audio can be overcome by employing open-source software and hardware. This dissertation serves as a guide for artists and researchers interested in spatial audio, and in ambisonics in particular, who would like to adopt an open-source approach to their creative and scientific practice. In the context of underprivileged communities, such as those in my home country of Bolivia, this approach allows educators to engage with students of varying economic means while still providing a 21st-century education focused on cutting-edge technologies and complex scientific concepts.

There exist other motivations as well for this model of art-making and research. Reproducibility in science is the idea that an experiment needs to be able to be reproduced by other scholars for the results to be validated. Open-source software reduces the barrier to entry to reproducibility allowing this concept to be a reality. Furthermore, when researchers seek to extend or expand the scope of an experiment if open materials are used, this task becomes much simpler. In contrast, if one decides

instead to use proprietary software, financial barriers might limit the advancement of these scientific pursuits.

Reproducibility is not a problem with written music for acoustic instruments. However, when it comes to computer music, the problem is more complex. The operating system (OS), the hardware and software dependencies, and even the network protocols have to be considered if we wish for our work to stand the test of time. Adapting and expanding musical works created with this framework is much more efficient. Even if the original code is not available, using open software signals to one interested in reproducing the musical oeuvre, perhaps from just a description, that this task is possible with limited means[Puc01].

Both my father and mother are economists who dedicated their lives to the alleviation of poverty. This dissertation, and my practice in general, consider the realities of the people in my country. Education is one of the great equalizing forces capable of lifting people out of destitution and giving them an opportunity for a better life. The synthesis of my knowledge presented here, I hope, can set the foundation for future courses that I may teach in Bolivia to marginalized populations. In some respect, this work is meant to honor the legacy and contributions of my parents and acknowledge their sacrifices.

1.2 Contribution

This dissertation is separated into three main categories: creating, capturing, and conveying spatial audio. In the first of these sections, I will discuss the principles of ambisonics, the main spatial audio technology I employ in my art and research. This first section will also discuss the development of various artistic works employing this technique. In total, I will discuss three works: *Chaqu*, *Alternate Spaces*, and *Six Seasons*, the latter being a collaboration with composer Lei Liang. These pieces

were performed with the support of the NYCEMF[1], SMC[2] conference, and ArtPower group, respectively. For all these works, we sought to adopt a strict open-source methodology. In certain cases, we employ proprietary, free, cross-compatible software (this is software that is not open-source but can be used in all major OSs including Linux distributions).

In the second section of this document, we will describe the development of an ambisonic microphone array. The uniqueness of this device lies in its open-source design and use of Microelectromechanical Systems (MEMS) - in this case, microphones. These microphones are extremely cheap making them ideal for fabricating microphone arrays on a budget. The contributions include a:

1. 3D models which can be 3D printed using relatively cheap Fused Deposition Modeling (FDM) printers,

2. Set of MATLAB routines for calibrating, encoding, and equalizing the raw microphone signals, and,

3. A document describing the fabrication and deployment of the array in real-world conditions.

In the final section, we will address the topic of conveying spatial audio using WebXR technologies. This final chapter will introduce the reader to various artistic projects which employ Javascript (JS) libraries to distribute spatial music on the browser. This section will address the benefits and limitations of this technology, which, in contrast to other methods of spatial music dissemination is far more accessible, sustainable, and flexible. In particular, this section will discuss the development and execution of three works: *Bits*, *Pigments of Imagination (WebXR)*, and *Continuum*. All of these projects were created using FOSS and feature three

[1]New York City Electroacoustic Festival
[2]Sound and Music Computing

different kinds of 3D audio JS libraries. We should note, the second piece in this list is ongoing work by Tim Gmeiner, who has continued developing the project using the Unreal engine; this initial prototype of the work exists on the browser and is powered by the WebVR library A-frame. All three works in one way or another have also been submitted to calls for music at conferences worldwide.

1.3 Outline

As aforementioned, this dissertation will be separated into three main parts corresponding to the topics of synthesizing, recording, and distributing spatial music using FOSS. In particular, for the first of these sections, we will be splitting the chapter further into three main sections: encoding ambisonics, decoding ambisonics, and selected works. As expected, the first of these subsections will address the concept of ambisonic encoding, otherwise known as ambisonic synthesis, which entails duplicating a signal and multiplying the copies by a set of coefficients to derive the spherical harmonics - representing the sound pressure in 3D space. The second subsection will introduce the reader to the topic of decoding, which comprises using a linear combination of these basis functions to produce (virtual or physical) transducer feeds. The overview of binaural audio will be reserved for Chapter 4, which relies on this decoding process to deliver spatial music via the browser. The third and final subsection of Chapter 2 will focus on the aforementioned selected works which make use of these technologies.

Chapter 3 will discuss the development of an ambisonic microphone array in three parts. The first part will serve as a literature review on the topic, helping us frame our project in the scope of similar projects. In particular, we would like to focus on four projects we consider highly relevant to the discussion: the SpHEAR array [LL19], the MSMA [GPL18], the HOSMA array [MDLP20], and the Middlicott array [MW19]. The second section of this chapter will discuss the design and evaluation of

a First Order Ambisonic (FOA) array. Objective measurements of this system have already been published in the Proceedings of the Audio Engineering Society [Zal19]. The concepts and discussions arising from this topic will inform and support the final section which details the fabrication and implementation of a Higher Order Ambisonic (HOA) array.

Chapter 4, much like the preceding chapters, will also be divided into three sections: WebXR, binaural audio, and selected works. In the first of these subsections, we will talk about VR and AR. We will discuss hardware and software limitations and affordances of various platforms as well as the promise of WebXR. In the subsequent subsection we will address the topic of binaural decoding, which comprises the technique of filtering speaker feeds with transfer functions simulating the human head. Finally, we will discuss, in the final part of this chapter, about three selected works that implement binaural decoding via WebXR technologies as examples of FOSS-based compositions, demonstrating the viability of this format and its shortcomings.

As a means of framing this work, the introductory chapter of this text will also cover two related topics: the history of spatial music, and the psychoacoustics of spatial music. By spatial music, we mean music that features sound localization cues as a distinct character of the work. These include compositions such as Rand Steiger's *Triton's Rising*, in which musicians were displaced from the stage, residing instead in balconies around the performance area, Shahrokh Yadegari's *Becoming*, which uses sophisticated GPU-based physical modeling algorithms to spatialize sound or any of the selected works described in Chapters 4 and 2. In our discussion of spatial music, we exclude telematic material[3], which could also be considered to fall within this umbrella term.

[3]Music which uses network protocols allowing performers in two remote locations to perform together.

1.4 History of Spatial Music

Space as a parameter of music-making is a feature that has been explored by composers for hundreds of years. Unfortunately, most musicians have a limited understanding of the possibilities that modern systems provide them and continue to operate on the basis of two-dimensional sound[4]. Despite great efforts by the audio industry to expand commercial systems from stereo to more sophisticated formats, two-channel audio reproduction systems have remained the *de facto* playback method in most homes due to their accessibility and simplicity.

Before the rise of electro-acoustic composers, many artists experimented with space as a parameter of composition by including within their scores: the placement of musicians in different parts of the concert hall; choreographed trajectories for musicians; or using height by placing musicians in balconies. Many challenges that existed then remain now: synchronicity between players, architectural changes between venues, and, providing a consistent experience for all audience members regardless of their seats.

Avant-garde composers in the 20th century pushed the envelope further by taking advantage of the technological developments of their era. With the advent of transmitted and recorded sound, *disembodied sound* - a musicological term referring to the displacement in space-time of musician and sound - found a permanent role in the acousmatic works of Schaeffer, Boulez, Stockhausen, Xenakis, and Cage, to name a few[5]. These acousmatic works meant for reproduction over loudspeakers made great use of spatial audio technologies such as the *quadraphonic* or *octophonic* sound systems, which were being developed at the time[6].

Unfortunately, despite great scientific leaps, many of these older works cannot

[4]Left/right panning plus distance modeled via amplitude changes, or the addition of reverberation (a three-dimensional model would include height).

[5]Acousmatic typically refers to musical material where the musicians are not present, we only experience pre-recorded sounds in these types of works.

[6]Four and eight-channel sound systems, respectively.

be fully appreciated by the general public since these high-end sound reproduction systems remain protected by privileged institutions. Commercial movie theatres, which harbor similar systems, have no financial incentive to reproduce these works. The primary motivation of this thesis is to explore accessible techniques for the creation, documentation, and dissemination of works that employ spatial sound as a primary component of the compositional process. This section is intended to provide background information by outlining a few important events and people in the field of spatial music. This context, we hope, will convey the importance that space has had in the practice of musical composition, much before VR existed.

1.4.1 Acoustic Music

Antiphonal Music

Antiphonal music is perhaps the oldest spatial music tradition. The practice of antiphonal music, also known as *call and response*, can be traced back to Biblical Times, with evidence of its existence as far back as the Roman Catholic Church in the 4th century. In a call and response system, the composer writes melodic lines using tension and resolution having an independent choral group assigned different parts of the melody. The composer might create tension by using a dissonant note as the final note of the calling phrase. A different group, located in a different location on the stage, would resolve the melodic phrase - usually ending in the tonic of the scale, while the harmony resolves using some traditional cadence.

Baroque Era

Many of these early works are hard to replicate today due to the lack of documentation. In the 16th century, however, printed works in spatial music would surface. Some of these initial innovative pieces were written by the Flemish composer Adrian Willaert, for example, who exploited the *Basilica de San Marco's* two

organs in antiphonal compositions featuring two separate choirs and multiple instrumental groups[7] [Arn59]. The technique *cori spezzati*, or separated choirs, was re-introduced in his 1550's piece titled *Vespers*, which itself featured multi-part arrangements and echo effects. Willaert's pupil, Andrea Gabrieli, an Italian composer of the late Renaissance, would later continue his teacher's work, cementing spatial music as a hallmark of the Venetian musical practice.

Figure 1.1: Basilica de San Marco [Cle]

As a result of Willaert's work the practice of spatial music quickly spread to other parts of Europe where it was quickly adopted by composers such as Thomas Tallis in England. *Spem in alium*, one of Tallis's most famous pieces, was composed for Queen Elizabeth upon her 40th birthday, in 1573, and featured 40 vocal parts arranged in eight 5-voice choirs. The high point of spatial music during the Baroque era, however, was Orazio Benevoli's *Festival Mass*, written in 1628 for the Salzburg

[7]Adrian Willaert was a Netherlandish composer of the Renaissance. The Basilica de San Marco is a cathedral church of the Roman Catholic Archdiocese of Venice, northern Italy.

Cathedral[8]. It called for 16 vocal parts, 34 instrumental parts, two organs, and a *basso continuo*[9] [Zvo99].

The Romantic Period

Following the Baroque era, interest in spatial music subsided until the beginning of the Romantic period. A key distinction of this era is the use of spatial music for theatrical effect. The Romantic period is often characterized by the rise and popularity of operas by composers such as Hector Berlioz, a French Romantic composer, and conductor (11 December 1803 – 8 March 1869), who wrote *Requiem* in 1837, and Gustav Mahler, an Austro-Bohemian Romantic composer (7 July 1860 – 18 May 1911), who wrote *Symphony No.2* in 1895[10][Ein48]. These composers not only spaced instruments for their performances but also choreographed the entrance and departures of musicians as they played, creating some iconic musical moments.

The 20th Century

In the 20th century, experimentalists such as Charles Ives, an American modernist composer (October 20, 1874 – May 19, 1954), and Luigi Russolo, an Italian Futurist composer and the author of the manifesto *The Art of Noises* (30 April 1885 – 6 February 1947), inspired by the tumultuousness of industrial life, refined the art of musical collage using spatial sound [Jon91] - Ives's 1908 *The Unanswered Question* called for offstage strings. Later, Henry Brant, inspired by Ives's music, would go on to create *Antiphony I* (1953) which called for five spatially separated orchestras,

and, *Voyage Four* (1963) which called for three conductors to direct: percussion and brass on stage, violins on one balcony, violas and celli on another, basses on the

[8]Benevoli was a Franco-Italian composer born in Rome (19 April 1605 – 17 June 1672).

[9]Figured bass where the most common combination is harpsichord and cello for chord and bass-line respectively.

[10]Mahler is considered one of the leading conductors of his generation.

floor level at the rear, woodwinds at a rear balcony, and several performers in the audience[11] [Zvo99].

Brant's *Windjammer* (1969), likely inspired by the opera composers of the baroque times, featured a static horn soloist and several wind players that moved along prescribed routes as they performed, in a choreographed manner. These works of acoustic music compel us to consider how spatial music can be created without technological means. Today we wonder how these works would have sounded in situ, at the time they were being performed. Technologies such as the ones described in Chapter 3 might become increasingly necessary for documenting the spatial music works of today. Composers might imbue significance to sound positions, which are lost if flattened - imagine motifs corresponding to forlorn lovers, getting closer physically as the music progresses.

Henry Brant

Henry Brant is one of the most successful contemporary composers of acoustic spatial music. Harley [Har97] provides a rich analysis of four of Brant's most important works in the field of spatial music. Brant emphasized the need for differing timbres to be represented in his music since he believed the heterogeneity would aid in the clarity of the musical representation. Brant summarized his main observations with regard to spatial composition in a 1967 article paraphrased as follows:

1. **Spatial separation clarifies texture**: when multiple musicians play different musical phrases in the same octave range, separating them spatially helps give clarity to each stream.

2. **Separate groups are difficult to coordinate**: exact rhythms might be difficult to accomplish due to the distance it takes for sounds to arrive from

[11] Henry Brant: Canadian-born American composer who composed numerous orchestral spatial works (September 15, 1913 – April 26, 2008).

one playing position to the next.

3. **Spatial separation is equivalent to range separation**: it is possible to enhance the texture of a melodic phrase simply by separating musicians playing the same material.

4. **Spatial arrangements must allow flexibility**: the specific architectural demands of the works cannot always be met, and alternatives should be provided whenever possible.

Brant is perhaps one of the most prolific composers of spatial acoustic works with a repertoire of 67 spatial pieces. Brant also wrote extensively on the subject of spatial music and experimented at length with form, rhythm, and perceptual experience in this context. In contrast to other composers of his era, Brant's music dealt exclusively with acoustic means, which he creatively articulated to provide depth, envelopment, and movement. These principles, albeit articulated for acoustic composition, are some worthwhile ideas for electroacoustic music composers to consider in works for multichannel systems with musicians, such as *Six Seasons* described in Chapter 2.

1.4.2 Electro-acoustic Music

This section is intended to show how technological advancements have a profound influence on artistic pursuits. Much like composers in the past, modern artists gravitate toward new tools in their quest to create an indelible mark in history. The tools we have at our disposal shape our curiosity and creations. New technologies always harbor inspiration for those who can afford them. Our goal is to demonstrate how these technologies can be made more accessible through education, and inclusive-minded design. These preceding technologies are those which galvanized some of the most important electroacoustic composers of the 20th century. Their works raise many questions: How will the spatial instruments of the future look? How will we

document and present this music? And, will virtual and physical worlds ever blend seamlessly?

The Theremin

Along a parallel branch of history, there exist a number of musical experiments conducted in the 20th century by avant-garde musicians and composers typically categorized as *electro-acoustical*. The composers of this space and time were all activated by the development of sound recording, radio, and telephony, and used them purposefully in their works. With the possibility of sound as a medium displaced from its source, dozens of new forms of music were created. In 1923, Leon Theremin introduced a new instrument to the public; the *theremin*, named after its inventor, used two antennae to control the pitch and volume of a synthesizer. The user must position their hands in space to control the instrument and produce the desired sound. Theremin later formed ensembles in which multiple theremins were used in one of the first public displays of multi-channel loudspeaker music ever.

Phonographs and Music for Magnetic Tape

The 20th century also saw the rise of phonographs as musical instruments[13]. Paul Hindemith, a prolific German composer (16 November 1895 – 28 December 1963), began this practice in the 1920s and 30s nearly 80 years before DJing practices became popular [Man13]. The American avant-garde composer John Cage (September 5, 1912 – August 12, 1992) also adopted the use of phonographs compositionally in *Imaginary Landscape No. 1* (1939), which featured multiple turntables and test tones, and in *Imaginary Landscape No. 4* (1951), where he used: 12 radios, 24 performers, and a conductor. In addition to exploring the use of record players as

[12] Alexandra Stepanoff was one of Léon Théremin's first theremin students in the United States [20].

[13] Colloquially known as record players.

Figure 1.2: Alexandra Stepanoff[12]playing the theremin on NBC Radio, 1930 [The]

instruments, Cage also exploited radio broadcasts and tape in his creative practice.

By separating the original performer from the playback these composers were playing not just with space, but also time. Cage, along with a group of experimental composers called "Project for Music for Magnetic Tape", would go on to write four pieces for tape during the 50s [Cag61]. The most famous piece that emerged from the group was likely *Williams Mix* (1952) which called for 8 tape machines, each played back from its own speaker, and hundreds of sounds carefully spliced together[14]. This was one of Cage's first uses of chance in musical composition. The project also resulted in works by Earle Brown(*Octet*) and Morton Feldman (*Intersection*), both using 8 tape players and speakers[15].

[14]Tom Erbe created a Pd version of William's Mix. (access: Jan 7, 2021)

[15]Brown was an American composer who pioneered the use of graphic scores (December 26, 1926 – July 2, 2002). Feldman was an American composer best known for his extended works which could last up to six hours (January 12, 1926 – September 3, 1987).

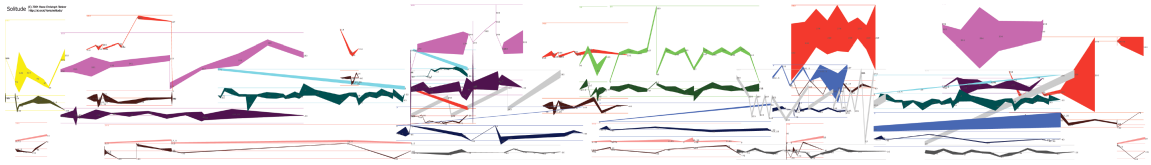


Figure 1.3: Hans-Christoph Steiner's graphic score for Solitude, created using Pure Data's data structures. An example of a graphic score, such as those popularized by Earle Brown. [wik20b]

Musique Concrète

Cage, Brown, and Feldman were greatly influenced by Pierre Schaeffer, a French engineer at Radiodiffusion-Television Francaise (RTF), who, in 1948, presented the first musical works created with disk recorders. These were the first examples of *musique concrète*, a style of music that is defined by the use of recorded sounds interpreted as musical material. Schaeffer would later go on to collaborate with Pierre Henry, another notorious French composer, creating a repertoire of works for tape including, most notably *Symphonie pour un Homme Seul*, which translates to "Symphony for one man alone". (1950).

In their 1950 piece, four channels were arranged in a tetrahedral configuration (e.g., a three-sided pyramid) with two front speakers, one back speaker, and a final overhead speaker making this one of the first examples of *periphonic* music[16]. Schaeffer also helped develop the *potentiomètre d'espace*, one of the first *spatial audio* controllers which he manipulated during live performances to modify the amplitude of speaker feeds. In Chapter 4 we will talk about a composition inspired by musique concrète, in which we sample 10,000 audio files and represent them in a virtual scene.

[16]Periphonic refers to sound with height, while *pantophonic* is used to denote sound systems with horizontal-only reproduction.

WDR

In Germany, another composer by the name of Karlheinz Stockhausen was making his mark in *electro-acoustic* music. Stockhausen is perhaps the most ambitious composer of *spatial music* in this era. Inspired by tape music, Stockhausen would travel to RTF, the birthplace of tape-based works, to learn about the technology. Soon after his time at RTF, he would return to Westdeutsche Rundfunk's (WDR) Studio fur Elektronische Musk to create his most prolific tape pieces[17]. *Gesang der Jünglinge* (1956) is considered by some as the first piece for multi-track tape, using a 4-track machine plus a fifth mono tape player for the fifth channel[18]. The premiere featured a number of speakers arranged *panoramically* onstage [Zvo99] - Stockhausen later remixed this piece for a quadraphonic sound system.

In 1960, Karlheinz would go on to create *Kontakte*, his first truly quadraphonic piece. Stockhausen used a turntable system with a rotating speaker and four microphones to create the illusion of spinning sounds. Stockhausen also explored spatial attributes of sound in his acoustic compositions written for multiple orchestras and multiple choruses. His most ambitious and bizarre work, however, is perhaps *Helikopter-Streichquartett* (1993) - completed much later. In this piece, the four members of a string quartet perform from four helicopters, which fly independent routes. Each has its own audio-visual system, along with a sound engineer. The sound and video are transmitted to the concert hall and the sound from the flying helicopters forms part of the piece [Sto96].

While quadraphonic systems were very popular in the '60s and '70s, today many spatial music concerts around the world use much larger speaker systems. The price of hardware has made it possible for Universities, theatres, and cinemas to deploy large-scale speaker systems for this type of work. Unfortunately, composing for such places may feel impossible without the appropriate infrastructure. Chapter 2 will

[17]Westdeutsche Rundfunk translates to: "studio for electronic music".
[18]The title translates to "Song of the Youths".

describe some accessible, cross-compatible, and open-source tools that can partially remediate this. This makes it possible for someone with limited means to compose for elite spaces that provide such facilities.

World Expos

In 1958, another one of the most important examples of spatial electronic music would be performed. Edgard Varèse's[19] *Poème Électronique* was featured at the Brussels World Fair (Expo 58), a major international event with artists from all over the world. The festival attracted up to two million visitors! For the exposition of this piece, the Philips Corporation set up 15 tape recorders and over 400 loudspeakers [MM95]. This is still one of the largest spatial music compositions that have ever been accomplished.

Xenakis, another giant of electro-acoustic music, designed the Philips Pavilion under the supervision of renowned Swiss-French architect Le Corbusier (1887-1965). Xenakis's *Concret PH*, created using recordings of burning charcoal, and modified using tape techniques, also played at the Expo, as an interlude between performances of Varèse's piece, which was the main attraction[VTL10]. Expo 58 also featured *Vortex*, a series of thirteen programs developed by Jordan Belson and Henry Jacobs for the Morrison Planetarium in San Francisco. The program featured music by Stockhausen among many others [Zvo99].

EXPO 70 in Osaka, Japan, also featured a number of groundbreaking spatial works. There, Xenakis presented a 12-channel tape composition titled *Hibiki Hana Ma*, which translates to "Reverberation Flower Interval", which was composed using the UPIC system. The UPIC[20] system was an instrument that allowed the composer

[19]French-born composer (December 22, 1883 – November 6, 1965). Varèse coined the term "sound mass" to describe his music.

[20]Unité Polyagogique Informatique CEMAMu (Centre d'Etudes de Mathématique et Automatique Musicales).

Figure 1.4: The Philips Pavilion [wik20a]

to draw scores that would be interpreted by a machine[21]. The graphical input device was used in conjunction with orchestral recordings, biwa[22] recordings, and snare drum recordings [Ian].

For the performance at the Japanese Steel Pavilion, the sound system featured 800 speakers situated above, around, and even under the seats. At the same time, in the German Pavilion, Stockhausen, along with 20 soloists, was performing in a spherical auditorium featuring 55 speakers and six small balconies. The auditorium had an acoustically transparent grid that allowed sound to emanate from below the seating area which seated 600 people [Zvo99]. This highlights how much has already been accomplished in the field of electroacoustic spatial music.

The Pepsi Cola Pavilion was another space created for EXPO 70 - this time designed by an American group called: "Experiments in Art and Technology" (EAT). The domes' 37 loudspeakers were arranged in a rhombus and could be driven with 16 monaural tape recorders and 16 microphone preamps for a total of 32 inputs.

There were also 8 signal processing channels with amplitude modulation (AM),

[21]The software iannix is the modern equivalent of the UPIC system.
[22]Japanese plucked instrument resembling a lute or guitar.

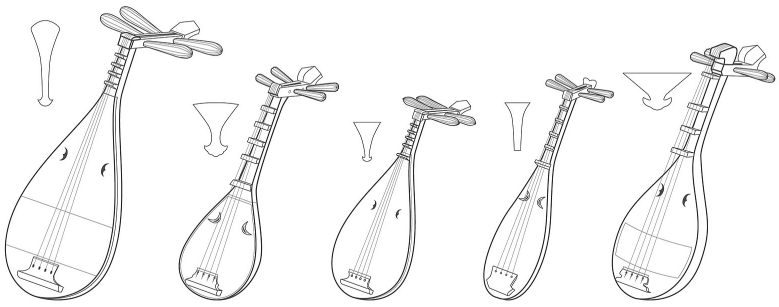


Figure 1.5: Types of Biwa [Fild]

frequency modulation (FM), and various filters, through which these sources could be routed. Handsets were also distributed which could pick up sound in each of the 11 zones. The dome was outfitted with laser beams and fog machines, and the walls of the dome itself were made of specular surfaces, for added dramatic effect. The dome was designed as a modular instrument, which could be reconfigured to fit the range of artists' visions [Ber12].

Classifying Electro-Acoustic Spatial Music

Given the wide range of creative options for spatialization of sound [Zvo00] provides a list of the different types of electro-acoustic works in relation to their associated techniques - along with some influential composers representative of each method - to help us understand their differences:

1. Live performance or "diffusion" of sound (Pierre Henry).
2. Environmental multi-channel soundscape (John Cage).
3. Classic studio multi-track tape composition (Karlheinz Stockhausen).
4. Automated location control (Edgard Varèse).

In this **live performance** setting, Pierre Henry created a repertoire of tape works that would be *panned* around in real-time according to the desired trajectories manipulated by the composer. These works made use of Schaeffer's aforementioned *Potentiometre d'espace* created in 1951, which translates to spatial potentiometer[23]. Henry also worked on **diffusion works** where stereo (or mono) tapes would be played back over large loudspeaker systems.

Examples of Cage's **environmental multi-channel soundscape** include *HPSCHD*[24](1960) used 58 channels - seven for harpsichord soloists and 51 for computer-generated tapes. The result of all these sound sources was a dense microtonal sound mass. There were also 80 slide projectors, seven film projectors, and a 340-foot circular screen. The piece was 5 hours in duration, however, participants were encouraged to enter and leave the environment as they desired.

Classic studio multi-track compositions refer to short oeuvres, possibly with musicians, which are performed with a multi-channel tape player. **Automated location control** refers to using predefined sound trajectories to control the movement of sound. In the time when these compositions were first being developed the movement of sound sources was printed on tape recordings, and the amplitude of various channels would be modulated by the output of these other tape players. Today, we use digital methods to accomplish the same results with far greater ease, or we can use randomness to control trajectories in real-time.

Chowning

John Chowning, famous for his discovery of frequency modulation (FM), as a musical technique, at Stanford, also published and composed spatial pieces still relevant

[23]A potentiometer is an electrical component which allows one to manually change the resistance between two points in a circuit.

[24]Short for harpsichord. In collaboration with Lejaren Hiller, another American composer famous for seminal work in algorithmic composition and founder of the Experimental Music Studios at the University of Illinois at Urbana-Champaign.

today. "The Simulation of Moving Sound Sources" [Cho71], published by Chowning in 1971, is considered a seminal piece of technical literature by a pioneering computer musician detailing his spatial audio algorithms. Chowning describes a system for the synthesis of spatial sound relying on FM and reverb to describe moving sources. FM in this context was used to simulate the *Doppler shift*.

His piece *Turenas* (1972), which makes use of a quadraphonic layout and this algorithm, was composed at CCRMA[25], an important institution he was a founding member of. It should be noted here that the substantial difference in speaker count between *Poème Électronique* and *Turenas* had no impact on the historical importance of these works. It is easy to believe that a higher speaker count is crucial for great spatial music - Chowning's work demonstrates that this is not always the case.

Part of the reason for the lower speaker count in some of these early works can be explained by the exorbitant cost of computers capable of handling these processes at the time. In 1981 at IRCAM[26] the 4X synthesizer used to perform Pierre Boulez's[27] *Répons* (1985) had a cost of $100,000. Today it is possible to recreate the synthesized elements of this piece using a personal computer [Zvo00].

CARL

Roger Reynolds in the 70s also created a number of important works. Reynolds used voice recordings in his *Voicespace* pieces such as *Still* (1975) and *Eclipse (Voicespace III)* (1979). The pieces were created using analog equipment including recorders, mixers, reverbs, and voltage-controlled spatial location systems. *The Palace (Voicespace IV)* (1980), used a quadraphonic sound system and recording of singer Philip Larson, which were analyzed and processed to emphasize the harmonic content. A

[25]Center for Computer Research in Music and Acoustics (CCRMA) at Stanford is an important computer music research facility.

[26]Institut de Recherche et Coordination Acoustique/Musique. Or, Institue for Research and Coordination Acoustical/Musical. It is organizationally linked to the Centre Pompidou in Paris.

[27]French composer perhaps best known for his large orchestral works featuring live electronics (26 March 1925 – 5 January 2016). Founding member of the aforementioned IRCAM.

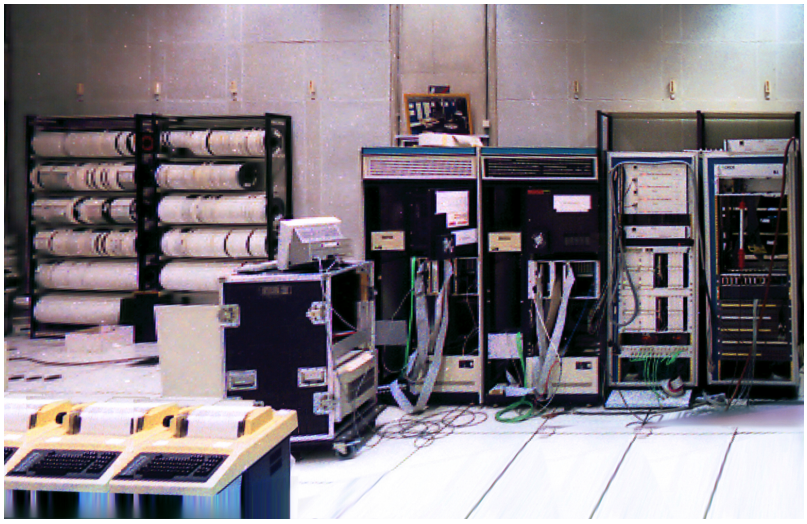


Figure 1.6: IRCAM 4X [Guy]

reverberation algorithm also gave the illusion of an impossibly huge space [Zvo99].

Reynolds continued working with sound analysis in *Transfigured Wind II* (1984). The piece features quadraphonic tape, a solo flute, and an orchestra. There is also analysis and resynthesis of the flute using software created by IRCAM. *Watershed IV* (1996), for solo percussion and computer, is another example of his spatial works, this time featuring 6 speakers on stage and 2 additional surround speakers. The piece was performed by Steve Schick (1954), the American percussionist and conductor.

The Red Act Arias (1997), another work by Reynolds, is based on the Greek tragedy of Clymenestra and Agamemnon. The piece features orchestra, chorus, and octophonic processed sounds using the CMUSIC "space" unit generator. CMUSIC was developed in 1980 by Richard Moore at the Computer Audio Research Laboratory (CARL) of the Center for Music Experiment at the University of California at San Diego (UCSD)[28][Moo82]. *Justice* (2000), by Reynolds, featured a soprano, actress, percussionist, tape, and computer spatialization. This demonstrates the legacy of spatial music that our University has, providing a reason for this research, which we

[28]It was later ported to Pd by Shahrokh Yadegari and further expanded into a GPU ray-tracing algorithm.

believe could be valuable for our institution.

Acousmonia

In Montreal in 1999, ACREQ[29] reproduced Pierre Henry's *L'Apocalypse de Jean* (1968) through a 24.6 sound system using the commercial CD as the sound source. Henry is one of the pioneers of this *diffusion* practice, in which mono or stereo signals are played back over a multitude of speakers. Sound systems such as the ones proposed by Henry are sometimes referred to as an *acousmonia*[30]. Here the sound system becomes an instrument itself, to be performed by the composer using, traditionally, musique concrète. François Bayle designed the first acousmonium in 1974 which was used by Groupe de Recherches Musicales (GRM)[31].

The Gmebaphone and Cybernéphone at the Institut International de Musique Eléctroacoustique de Bourges (IMEB)[32], and the Birmingham[33] ElectroAcoustic Sound Theatre (BEAST) are a few other examples of acousmonia. Notable American acousmonia include composer Stan Shaff's *Audium: A Theatre of Sound-Sculpted Space* and the Recombinant Media Lab both in San Francisco. These are all historical examples of multi-channel sound systems, since then several institutions have developed their own diffusion systems for spatial music performance. These historically relevant systems have also evolved over the years from diffusion systems to arrays supporting modern spatialization algorithms.

[29]Association pour la Création et la Recherche Electroacoustiques du Québec. The first non-university-affiliated organization in Canada exclusively dedicated to electro-acoustic music.

[30]Acousmonium being singular.

[31]Institute created by Pierre Schaeffer in 1958 in France.

[32]Central France.

[33]Second largest city in England.

Figure 1.7: Pierre Schaeffer Presenting the Acousmonium [Filc]

Spectromorphology

Denis Smalley is another pioneering composer who has worked and continues to work, extensively with space in his acousmatic compositions. Smalley has created his own language to describe a sound which he describes as spectromorphological. Smalley describes [Sma97]:

> Spectromorphology is not a compositional theory or method, but a descriptive tool based on aural perception. [...] Although spectromorphology is not a compositional theory, it can influence compositional methods since once the composer becomes conscious of concepts and words to diagnose and describe, then compositional thinking can be influenced, as I am sure my own composing has been.

Smalley has created an extensive dictionary to describe the way he envisions and articulates his music. [Bla11] pictorialized some of these concepts in order to help the reader better understand Smalley's language. Here we will only present written definitions of some of the most popular concepts defined by Smalley.

1. **Source-bonded space:** refers to the spatial zones and mental images produced by or inferred from, sounding sources, and their causes (if there are any).

2. **Source bonding:** related to source-bonded space, is the *natural* tendency to relate sounds to *supposed* sources and causes and to relate sounds to each other because they appear to have shared or associated origins.

3. **Perspectival space:** relating to our perspective, can be further subdivided into various perspectives.

   (a) **Prospective space:** frontal image, which extends to create panoramic space.

   (b) **Panoramic space:** the breadth of the prospective space, extending the limits of the listener's peripheral view.

   (c) **Circumspace:** the extension of prospective/panoramic space so that sound can move around, above, and through the space occupied by the listener.

4. **Spectral spatiality:** impression of space and spaciousness evoked by occupancy and motion within the range of audible frequencies.

This list forms only a small part of the language created by Smalley to describe his music. *Pentes* (1974), *Empty Vessels* (1997), *Wind Chimes* (1987), and *Valley Flow* (1991–92), are a few of Smalley's most notorious compositions in which he attempts to sonically articulate all these different concepts. O'Callaghan [O'c11] provides a closer examination of these works. According to O'Callaghan, many of Smalley's works utilize mimetic properties: the sound design aims to simulate real-world sounds using electro-acoustic means, much like *Riverrun* (1986) by Barry Truax, which utilizes granular synthesis to emulate the sound of a flowing river.

All of these various ideas about composing with space show how pertinent this work is for modern composers, given Smalley's prominence and recency, and provide us with a framework for composing our own spatial music. This final example brings us to the conclusion of the history of spatial music. As we can see, while many contemporary composers might attribute the development of spatial music to composers of the 20th century, the use of spatial elements in composition has a longstanding tradition that spans hundreds of years.

While many composers today, in well-funded Universities, have the pleasure to explore spatial sound through the use of multi-channel systems, much of their detailed work commonly gets *down-mixed*[34] to stereo formats prior to distribution. If the spatial attributes of a particular sound are indeed important to the composition, this ultimate representation leaves much to be desired. Luckily, many free and open-source solutions exist for audio engineers to create and re-distribute works of this variety preserving the auditory elements pertinent to the works; our intention is to highlight some of the best low-cost solutions available, giving artists greater freedom in the music-making process.

In addition to discussing technologies available for the creation, recording, and presentation of 3D audio - in a musical context - we will discuss some of the technical challenges inherent in these productions in later chapters. Namely, we will consider the difficulty of 3D sound reproduction based on the computer: storage, speed, and changing operating system (OS); we hope this knowledge proves useful for any composer interested in the minutiae of the science, and art, of *spatial music*.

1.5 Psycho-acoustics of Spatial Sound

Before we can understand the mathematical details behind different 3D audio technologies, we should discuss some of the psycho-acoustic principles on which all

[34]Projected down from a multi-track format to stereo, usually for commercial purposes.

spatial audio technologies rely. Many authors have written extensively about the subject, with entire books having been dedicated to the subject (Blauert's 1997 [Bla97] being perhaps the most popular). Here we will give only a short account of the many principles that have been defined in this domain in order to inform the reader of some of the most salient features of the auditory system that inform our perception of sounds. The purpose of this section is to explain and inform the reader of the underlying biological mechanisms that permit this type of experience.

1.5.1 Sound Localization

From the psycho-acoustic perspective, the most important feature of the auditory system is our ability to localize sound, this is perhaps why so much attention has been devoted to the subject in the psycho-acoustic community. Having a sensitive auditory localization system poses an evolutionary advantage: being able to detect the presence of a predator, before it can reach us, is likely why we have developed the ability to localize sound. Similarly, being able to detect prey, before it can escape, could be considered evolutionarily advantageous.

A number of perceptual mechanisms work in conjunction and amalgamate into a single sensory experience informing our understanding of a sound origin (directionally speaking). Not only are we conditioned to use acoustic cues to draw these conclusions, but memory and vision also form a part of this complex model [Ken95]. Our subconscious contains a mental map of the prevailing origin of different sounds based on prior experience which informs our predictions of source orientation, velocity, and direction.

Our auditory model of sound localization is quite complex - especially for multiple sources in spaces. When sound waves travel through the air, if there is more than a single source, the pressure waves constructively and destructively interfere making localization estimates harder. Additionally, if the space is very reverberant (e.g., if

there are many reflections from walls, such as in a chapel) these predictions might become even more complicated to make.

It is useful to define two common conditions for sound sources: those in *free-fields* and those in *diffuse-fields*. Free-field refers to an environment in which there is no reverberation; in other words, there are no surfaces upon which the sound may reflect. This condition is seldom found in nature, however, *anechoic chambers*, rooms with no echoes, such as the one in Figure 1.8, have been designed specifically with this characteristic in order to conduct acoustic experiments. This distinction will become important in Chapter 3 when we discuss the measurement and calibration process of our microphone array.

In contrast to free-field conditions, a diffuse field refers to an environment in which prominent reflections are present, ideally with equal energy arriving from all directions. In architectural acoustics, diffusion will often be introduced intentionally, such as in concert halls, in order to spread energy evenly across the entire listening space. In between these two extremes, there exists a continuum of acoustic spaces, each with its own acoustic characteristics, such as the: *reverberation time* of the room or the shortest reflection path from a source to a destination - whether that be a person or a receiver.

1.5.2 HRIRs

An important aspect of spatialization in the musical domain, in addition to the horizontal and vertical direction of sounds, is the distance of the event in relation to the listener. For these three critical localization parameters, various associated auditory mechanisms exist in the acoustic domain which facilitates our ability to place sources in space. Any acoustic space can be considered a linear time-invariant (LTI) system and can be characterized by a single IR (impulse response) much like any digital or analog filter. Subtle changes in air pressure as a function of temperature

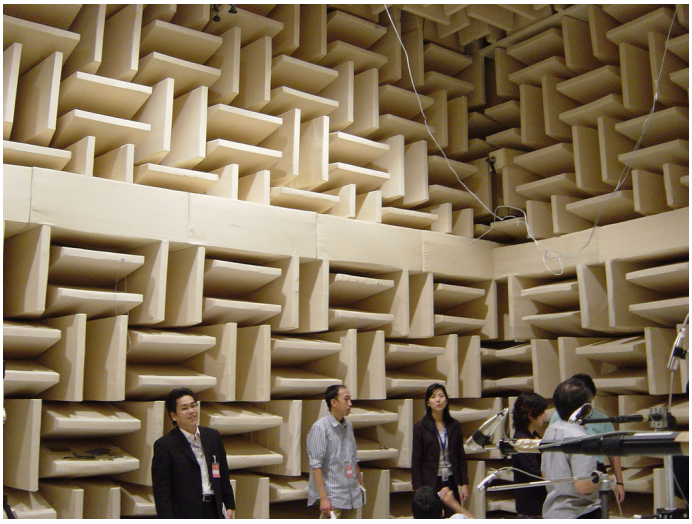


Figure 1.8: IBM Anechoic Chamber [Fila]

can affect sound propagation, but for all intents and purposes, this theory holds true.

In modern spatialization systems, we are not concerned with the IR of a single omnidirectional acoustic sensor but rather with the two impulse responses that can be captured at the listeners' ears. For this method there exist various devices used to capture the two IRs commonly referred to as *binaural impulse responses* (BIRs) or *Head Related Impulse Responses* (HRIRs) - in the time domain[35]. When a *Fourier transform* is applied to these IRs they are referred to as HRTFs, or *Head-Related Transfer Functions*.

The most common technique for HRIR capture involves placing two small microphones, or *in-ear* microphones, inside a person's ears, and taking snapshots at as many positions as possible by sending *exponential sinusoidal sweeps*(ESS) from speakers surrounding the listener. The recordings are then *de-convolved* in order to obtain the effect only of the listener's head and body on the two recordings as a function of direction and distance.

When the excitation signal is perfectly known, as in this case, de-convolution

[35]When these are taken in diffuse-field conditions they are called Binaural Room Impulse Responses (BRIRs) since the effects of the room are included in the transfer function.

yields an impulse response that captures the effects of the: room, speaker, and microphone. A full-spectrum signal is required for a good IR to be obtained. We tend to use *flat* frequency microphones and speakers, and an anechoic chamber, which has no response, to isolate the effects of the head and torso on the incoming signals.

Individual acoustic characteristics of the HRIRs have been assigned specific names. *Inter-aural level difference* (ILD) refers to the overall energy difference between the two IRs, *inter-aural time difference* (ITD) refers to the time offset, or phase delay, between measurements. Based on the *diffraction* and *reflection* effects caused by the geometry of the human head, the HRIRs will also exhibit spectral differences, which can be used to discern directions in situations when ITDs and ILDs are not sufficient.

Diffraction refers to sounds' ability to bend around objects. This is especially true for low frequencies. Reflection refers to sounds' ability to bounce off objects. This is especially true for high frequencies. ILDs are typically caused by the *shadowing effect* of the head and are therefore frequency dependent [CRPGT+19]. In other words, the ILD will be greater for higher frequencies since lower frequencies diffract, or bend, around the head - while high frequencies reflect off the head. ITDs are a prominent localization cue at low frequencies and ILDs are the prominent localization cue for mid and high frequencies.

This concept is known as the *Duplex Theory of Localization* and was proposed by Lord Rayleigh in 1907. These mechanisms and concepts will become important in Chapter 4, wherein we present a number of different WebXR experiences that rely on this technology to disseminate spatial music over the browser.

1.5.3 Perception of Distance

Perception of distance is more complex in some ways than that of orientation. Zahorik [ZBB05] provides, a comprehensive overview of auditory distance perception

research, including some cognitive science research that examines brain response to acoustic stimuli. As aforementioned, there are several cues that affect our perception of distance - the most salient one being overall energy. A few other unmentioned cues, however, also affect our perception of distance. Among these is the absence or presence of high-frequency content in a signal. High-frequency signals tend to dissipate faster in the air than low-frequency ones, because of the additional friction between molecules inherent to these acoustic events.

Auditory parallax, the effect whereby the position or direction of an object appears to differ when heard from different positions, has also been shown to improve distance estimations. Finally, as mentioned before, familiarity tends to affect our perception of sound origin. A sample of a whispering voice, virtually placed far away, will appear closer than it really is. Adversely, a shouting voice, placed close, will appear to be further away than reality. It should be noted that this parallax is not limited to the subject's directional or rotational motion, the movement of the sound source can also help inform our perception of distance.

Finally, *interaural coherence* (IC), the measured statistical coherence of signals received at each ear, aids in distance estimation. In a *diffuse field*, IC is low, because the scattering of sounds off walls interacts in unpredictable ways, which results in some amount of randomness in the signals when compared to each other (binaurally speaking). Therefore, IC can be used as a statistical estimate of distance, since we can assume that in most real-world conditions, sound sources will be diffused acoustically - lowering the overall IC for distant sources more than those proximal. This effect was exploited in the design of an instrument for spatial music discussed in Chapter 2, which featured the calls of various marine mammals.

1.5.4 Precedence Effect

Summing localization is a phenomenon - based on ITDs - which affects of perceived direction-of-arrival of sounds. When a broadband signal - a signal with a broad range of frequencies present - is played from two directions, with a delay of less than 1 ms, a single event is perceived in the direction between sources [HDSC$^+$17]. The perceived location shifts towards the first played source, as the delay increases. This *fusion* is one of three characteristics of the *precedence effect*, also known as the *Law of First Wavefront* or *Haas effect*, described in 1949 by Helmut Haas.

When the delay is between 1 and 5 ms, a single event close to the leading source can be heard [HDSC$^+$17]. The lagging source - the delayed copy of the sound - can be perceived due to the change in timbre, and is perceptually equivalent to a feedforward comb filter where the delayed copy is played back over a second discrete channel. In this context, the *echo threshold*, refers to the amount of delay that must be introduced to the copy before we begin perceiving the two sounds as independent non-fused events. This threshold is dependent on the nature of the signal. For a short broadband click, 5 ms might be enough to create an echo. For music and speech signals, the threshold is longer and can be as high as 20 ms.

The precedence effect can be summarized by the following three phenomena:

1. **Localization dominance**: the direction of an auditory event depends predominantly on the leading source.

2. **Fusion**: a single auditory event is perceived when two sound events are below the *echo threshold*.

3. **Lag Discrimination Suppression**: the direction of the lagging sound is suppressed.

This phenomenon is particularly important for spatial music represented via loudspeakers. If an audience member is sitting too close to a speaker, a source's direction

might be erroneously perceived - relative to the composer's intention. It is for this reason that surround sound systems require large spaces, and can't always accommodate the largest crowds.

1.5.5 Doppler Shift

Doppler shift is another auditory cue, related to *auditory parallax*, which can aid in our estimation of source distance and direction. Doppler shift is the phenomenon we experience when we perceive the pitch of an ambulance's siren or any other moving source, change as it approaches or moves away from us. Specifically, as the ambulance siren moves towards us, the pitch seems to be higher than when stationary, and as it moves away, it seems to be lower than when stationary.

This has to do with the nature of longitudinal waves, which contract and expand in relation to the speed of the object. LaValle [LaV16] provides a whole chapter on the auditory mechanisms involved in Virtual Reality (VR). The book also covers the fundamentals of VR optics, interaction, and tracking, and is recommended for anybody interested in Extended Reality (XR). Equation 1.1, from Chapter 11, describes our perceived frequency of sound as a function of its velocity. By combining this auditory cue with distance-based intensity dampening we can realistically simulate moving sources.

$$f_r = \left(\frac{s + v_r}{s + v_s} \right) f_s \tag{1.1}$$

Where:

f_r is the resulting frequency.

s is the speed of sound (343.2 m/s roughly at C 20°).

v_r is the velocity of the receiver (negative if the receiver is moving away).

v_s is the velocity of the source (negative is the source is moving away).

f_s is the original frequency of the source.

Doppler shift is an important cue to add realism to virtual worlds. In the context of music making, however, sometimes we have often opted out of applying Doppler shift to our sources since changing the pitch shifts their relationship to other harmonically related material. This underlines the distinction between using these technologies for simulating real environments versus implementing only some features to create experiences that are unrealistic yet musical.

1.5.6 Binaural Synthesis

Of particular importance to our work is the use of binaural synthesis for the simulation of acoustic spaces. Binaural synthesis allows us to experiment with *spatial music* without the need for sophisticated and often inaccessible loudspeaker systems. In binaural synthesis, sound sources are dynamically *convolved*[36] with HRIRs to simulate surround sound experiences. Binaural synthesis has become increasingly popular with the growth of *extended reality* (XR) systems. However, several problems, such as non-personalized HRTFs, externalization errors, and localization issues; persist.

Personalized HRTFs are believed to be part of the key to solving front-back confusions sometimes experienced in binaural synthesis [Joh19]. Unfortunately, these tend to be too laborious for most people to obtain. An additional problem was the lack of a standardized format for sharing HRTF data. Recently the Spatially Oriented Format for Acoustics (SOFA) format was designed to alleviate this issue [MIC+13]. Johansson [Joh19] also notes two additional problems with binaural synthesis that warrant more exploration:

1. Coloration artifacts might be the result of poor HRTF measurement techniques.

[36] As noted, convolution of a source with an impulse response is equivalent to applying a filter in LTI theory.

2. The lack of "dynamic HRTFs", which he describes as those that model head movement independently of shoulder movements.

Poor *externalization*[37], the degree to which listeners perceive sources as originating inside their head, can be mitigated by the use of a head-tracking system. The systems, based on *inertial measurement units* (IMU), monitor the orientation of the listener's head and adjusts the sound field accordingly. When no rotation adjustment is included, our auditory system assumes that sources must be originating from inside our head, as this is the only rational conclusion for sources without dynamic filtering.

The subtle spectral changes from position to position, provided by head-tracking, provide a way to disambiguate between source positions. The addition of diffuse field modeling, in the form of a reverberation unit, also supports improved *externalization*([SGK76]). In the past, we have built head-trackers using Arduino Micro or Teensy, both of which proved successful and useful for composing music with spatial attributes with a limited budget[38].

During the compositional process undertaken for the works featured in Chapter 2 we implemented the Nvsonic head-tracking system. The design by Tomasz Rudzki from York University allows us to send IMU (inertial measurement unit) data to a miniature development board (Arduino Micro) which then pipes the data to the computer via a serial bus line. The data is then sent to a router via OSC [Wri05] via a local IP allowing an external application, in our case Reaper, to receive and apply the data.

This concludes our introduction to the subject of the psycho-acoustics of spatial sound. It should be noted, once again, that for the sake of brevity, we have only covered the basic principles of this domain. The body of literature on this subject is simply too large to address in this chapter. This information is meant to complement

[37]Also sometimes referred to as Inside-the-Head Locatedness (IHL).
[38]See Nvsonic head tracker or Teensy head tracker.

the musical oeuvres which accompany this dissertation, giving the reader greater insight into the scientific principles exploited in the compositional process.

1.6 Conclusion

In this chapter, we have presented the reader with different perspectives from which to begin appreciating the complexity of the field of spatial music, and some background knowledge that should provide clarity to the proceeding chapters. Namely, we touched briefly on the compositional, psycho-acoustic, and engineering aspects of this domain. We talked about the history of spatial music from an acoustic and electro-acoustic perspective and provided the reader with a brief summary of the biological systems that create the illusion of spatial music commonly experienced in movie theatres and more recently in VR. The aim of the historical overview has been to provide a background that will shape the lens through which works in later sections are viewed. The psycho-acoustic primer was provided for completeness and to engender a greater appreciation for the richness of this musical domain. In the proceeding chapter, we will explore the technique of ambisonics, which can be used to synthesize music with spatial attributes. Namely, we will discuss the theory of ambisonic encoding and decoding, and present three works that make use of this technology.

Chapter 2

Creating Spatial Music

2.1 Introduction

Over the last century, a number of composers have explored the use of space as a key parameter of musical composition. Recently, with the growth of *Extended Reality* (XR), an even greater number of artists have begun exploring this medium, as a way to push this musical domain to new extremes. XR environments, such as *Virtual Reality* (VR) or *Augmented Reality* (AR), constitute a new field rich with creative possibilities for art-making. In this chapter, we explore the latest technological developments vis-à-vis tools that can be used to create spatial music which can be incorporated into these virtual environments.

In our research, we found that state-of-the-art development in this domain has shifted towards *Virtual Studio Technologies* (VSTs). Two institutions in particular, Aalto and IEM, have been developing *Free and Open Source Software* (FOSS) solutions, which are intended for *Digital Audio workstations* (DAWs) like Ardour or Reaper[1]. Other computer music languages, such as MAX/MSP[2] also support

[1]These particular DAWs have multi-track support and are either free or low-cost.

[2]Pd also has support for VSTs, see vstplugin~

VSTs, and offer support for multi-channel music, making them ideal for incorporating these tools into unquantized real-time compositional performances or interactive experiences.

This dissertation focuses primarily on the adoption and utilization of Free and Open Source Software (FOSS) in artistic and scientific contexts. The rationale behind this methodology stems from a desire to create an equitable and accessible practice that can be emulated with few financial restrictions. This is also a strategy that can allow me to teach concepts I have learned abroad to a wider base of students in my country of origin, Bolivia, who might not have the same access to resources as their international counterparts. This particular chapter of the dissertation focuses on artistic endeavors and the related practice that is enabled by ambisonic software created by several scholars at several Universities worldwide. We hope this deep dive into the mechanics of these effects will inspire sound engineers to develop their own plug-ins, as well as inform composers about the internal nature of these applications.

2.2 Ambisonic Encoding

This chapter of the dissertation focuses on the principle of ambisonic encoding or synthesis. The word encoding can have many different meanings in spatial audio which differ depending on the context. The word encoding should be understood as applying a transform to a signal or set of signals, thus converting the data to an intermediary format, presumably for later decoding using an inverse of the original encoding transform. In this particular chapter, we will explain how an arbitrary signal, in this case, a monophonic signal, can be encoded into ambisonics.

Ambisonics is a spatial technique relying on basis functions in 3D space entitled spherical harmonics (SH). These basis functions can be found in chemistry, physics, and optics. The underlying mathematics behind these basis functions can be quite complex. Much like sinusoidal waves can be used as basis functions for the de-

composition of a complex sound into several pure tones via an FFT (Fast Fourier Transform), spherical harmonics are used to deconstruct sound propagation in 3D space - in other words, spherical harmonics are a natural extension of sinusoidal patterns, simply expanded to one additional dimensional space.

When working in ambisonics, the first determination is the ambisonic order (n). The ambisonic order determines the spatial resolution that we can achieve with our data. The higher the ambisonic order the greater the spatial resolution, but also the greater the computing and memory requirements of the system. First-order ambisonics (FOA) has a total of 4 spherical harmonics, second-order ambisonics (2OA) has 9, and so forth (e.g., $(n+1)^2$). To encode a monophonic signal into ambisonics, we calculate a set of floating-point coefficients using the real-valued spherical harmonic equation. The equation gives us the magnitude of the spherical harmonic for a pair of angles θ and ϕ which correspond to the elevation and azimuth (e.g., angular direction of the signal on the horizontal plane) of the signal correspondingly.

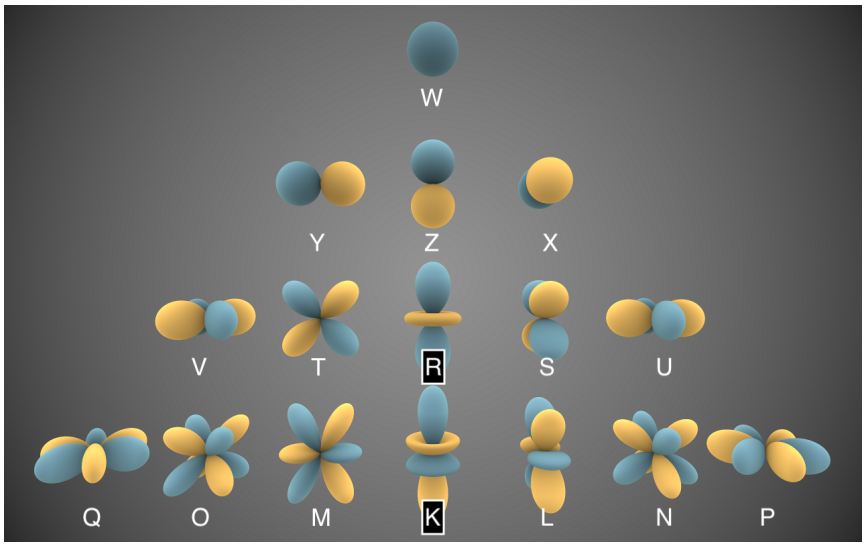


Figure 2.1: Spherical Harmonics [Ini14]

The easiest case is the harmonic W which is an omnidirectional pressure signal, it returns the value of 1 regardless of the values of θ and ϕ since the pressure is identical

at all points on the surface of the sphere. In contrast, the third harmonic (in ACN ordering called the Z harmonic), shaped like the figure-8 microphone pattern, returns 0 for any signal with an elevation equal to 0[3]. Intuitively, we can notice that this SH is seeking to encode information above and below the horizontal plane, and assigning a particular polarity to the signal based on its position in space. In Section 2.3 we will explain in further detail how these signals are decoded to an arbitrary loudspeaker array. The basic premise is that a linear combination of these signals can be used to generate cardioid (for FOA) virtual microphones pointing in different directions in space. The virtual microphones form the speaker feeds which will be distributed to our speaker array or sometimes these are convolved with HRTFs for binaural synthesis, which we will cover in Section 4.4.

In order to synthesize our monophonic signal into an ambisonic signal, sometimes referred to as B-format[4], we use the appropriate number of SH equations, based on the order of the system, and calculate the set of coefficients. For FOA we only need three additional functions seen in Equation 2.1. In real-time, assuming the leading coefficient r is equal to 1, we calculate the various values and multiply four copies of our original signal with the resulting values. The resulting signals are then said to be encoded in first-order ambisonic, ready for playback in an arbitrary speaker system[5]. These signals contain the X, Y, and Z pressure gradients, as well as one omnidirectional signal, which describes the fluctuation in pressure in space over time. The signals are not meant to be played back directly but decoded either binaurally or using a loudspeaker array.

[3]In certain libraries the index starts at 0, which would make the Z harmonic, harmonic number 2.

[4]For some authors B-Format is reserved for FOA encoded audio.

[5]The more speakers, however, the greater the ambisonic order is required. The available number of SHs must be greater than or equal to L, the number of loudspeakers.

Ambisonic Degree and Order

In mathematics and physics, the order (m) and degree (l) of spherical harmonics is the opposite of the ambisonic order (n) and ambisonic degree (m). In other words, the ambisonic order is the mathematical degree of the spherical harmonic, and the ambisonic degree is the mathematical order of the spherical harmonic[6]. Figure 2.1 shows a classic representation of spherical harmonics of increasing *degree* (l), where the top harmonic has a degree of 0 and the bottom harmonic has a degree of L. The *order* (m), in the mathematical sense, spans from negative to positive values and includes 0. The image was modified to show the letters associated with each harmonic, according to ACN ordering. To avoid confusion we will use the term *ambisonic order* rather than order, along with the corresponding notation[7]. For example, in Figure 2.1, the ambisonic order (n) increases from 0 to N, where the top harmonic has an ambisonic order of 0, and the bottom harmonic has an ambisonic order of N. The ambisonic degree (m) spans negative and positive values and includes 0 at the center column.

Coordinate System

An important consideration for ambisonic research is the 3D coordinate system. In spatial audio research, the coordinate system used is unfortunately not always consistent. This is especially true when we switch over to discussions of different spatial audio methods such as VBAP[8] or OBA[9]. For example, in SpatDIF [PLS12], the azimuth angle increases clockwise, and the X, and Y, axes are switched relative to the Ambix convention. In this text, when talking about ambisonics, we will use ϕ to denote the azimuth angle, increasing in the anti-clockwise direction, and θ to denote

[6]We are not really sure how this came to be, it simply is.
[7]Unless mathematical order is implied.
[8]Vector Based Amplitude Panning.
[9]Object Based Audio

the elevation angle, increasing upwards to 90°. This has become the standard way of writing ambisonic equations. The X, Y, and Z axes point towards the front, left, and up respectively. The conversion between spherical and Cartesian coordinates is defined as [GAK+19]:

$$\begin{aligned} x &= \|\vec{r}\| \cos(\phi) \cos(\theta) \\ y &= \|\vec{r}\| \sin(\phi) \cos(\theta) \\ z &= \|\vec{r}\| \sin(\theta) \end{aligned} \tag{2.1}$$

where $\|\vec{r}\|$ is the norm of the vector r, which is the distance of our source from the origin of our sphere[10]. In Figure 2.2, point P can be defined by the Cartesian vector with variables x, y, z, or a spherical coordinate vector with elements ϕ and θ.

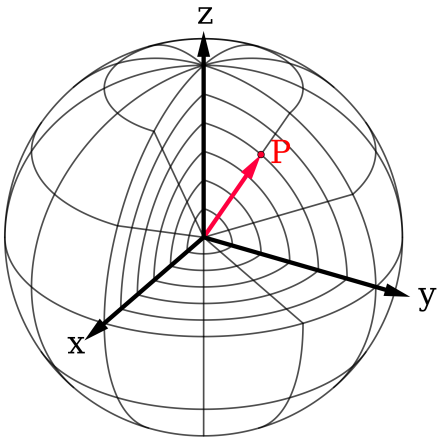


Figure 2.2: Spherical Coordinate System [Sha]

2.2.1 Normalization

Normalization refers to the practice of weighting the different ambisonic channels such that the SH vectors will not only be *orthogonal* to each other but also *orthonor-*

[10]The norm is the length of the vector, which is the square root of the sum of the squares.

mal[11]. Orthonormal bases have unique properties, desired for ambisonic systems, such as linear independence. In order to maintain a proper balance between diffuse (W) and direct (X, Y, Z) data, normalization is often applied. The normalization must be matched between the encoder and decoder, otherwise, an excess of a particular signal might be introduced into each speaker channel, distorting the balance between direct and diffuse sound.

The N3D normalization format results in a set of orthonormal basis vectors, which have useful mathematical properties. However, the preferred normalization today is SN3D, which forces SH above W to not exceed the level of the 0th ambisonic order harmonic [Kro14]. There are various standards for *ordering* ambisonic channels and their *normalization*. The most common are the *Furse-Malhalm* (FuMa) [CRPGT+19] ordering scheme and the *ACN*[12] ordering scheme. Table 2.1 shows these two ordering schemes up to 3OA[13].

At this point, FuMa, and the *maxN* normalization scheme it used are mostly legacy standards[14]. Developers are instead focused on the Ambix format [NZDS11], which consists of the ACN ordering scheme and the *Schmidt semi-normalized* harmonics format (SN3D) for normalizing. Equation 2.2 shows the SN3D normalization, according to Gorzel's 2019 paper on the implementation of Resonance[15][GAK+19].

$$N_{n_{SN3D}}^{|m|} = \sqrt{(2-\delta_m)\,\frac{(n-|m|)!}{(n+|m|)!}} \tag{2.2}$$

where $\delta_m = 1$ for $m = 0$, and 0 otherwise. In the Ambix specification paper [NZDS11], one should note there is a $1/4\pi$ term that is ignored in our formula. Ultimately, we can treat this term as a scalar, since it is independent of *ambisonic*

[11]Mathematically this means that the SH matrix multiplied by its transpose yields the identity matrix.

[12]Which stands for Ambisonic Channel Number.

[13]A third, more obscure ordering called SID by Daniel [DM04], was left out. 3OA stands for third-order ambisonic. SID is used in Pd in the iem_ambi library.

[14]In other words, they are no longer used.

[15]Resonance is an open-source Google product for ambisonic manipulation.

Table 2.1: Ambisonic Ordering in FuMA and ACN

Ambi Ordering (3OA)		
	FuMa	ACN
W	0	0
Y	2	1
Z	3	2
X	1	3
V	8	4
T	6	5
R	4	6
S	5	7
U	7	8
Q	15	9
O	13	10
M	11	11
K	9	12
L	10	13
N	12	14
P	14	15

order (n), or *ambisonic degree* (m) - thus, it can justifiably be omitted.

Another normalization standard is the N3D normalization scheme. The aforementioned N3D normalization standard is defined by the following equation from [PPQ16]:

$$N_{n_{N3D}}^{|m|} = \sqrt{(2n+1)\frac{(n-|m|)!}{(n+|m|)!}}. \tag{2.3}$$

This equation is included here simply for completeness. Most authors and companies use SN3D as the default normalization today in ambisonic development. Conversion between the two is also discussed in [PPQ16].

Real-Valued SH

As an introduction to the topic of ambisonics, we also present equation 2.4, which is used to calculate real-valued *spherical harmonics*.

$$Y_n^m(\phi,\theta) = N_n^{|m|} P_n^{|m|}(\sin(\theta)) \begin{cases} \cos(|m|\phi) & \text{if } m \geq 0 \\ \sin(|m|\phi) & \text{if } m < 0 \end{cases} \tag{2.4}$$

$N_n^{|m|}$ corresponds to the normalization term discussed earlier. $P_n^{|m|}$ are the *associated Legendre functions*, in this case, evaluated for $\sin(\theta)$[16]. Depending on the ambisonic degree, a suitable right-side argument is chosen to determine our solution. These functions are used to encode mono signals into SH representations and are also necessary to encode/decode signals recorded by SMAs (Spherical Microphone Arrays).

Associated Legendre Functions

The associated Legendre functions are solutions to differential equations used to describe wave functions in the spherical domain. Most authors use numerical libraries, such as Julia or MATLAB, to calculate these. However, if we wanted to calculate these ourselves, we could also use the recurrence relation and the first few given terms to manually calculate these for any l and m we desire[17].

Equation 2.5 shows the recurrence equation including the *Condon-Shortley* phase term [Wei].

$$(l-m)P_l^m(x) = x(2l-1)P_{l-1}^m(x) - (l+m-1)P_{l-2}^m(x) \tag{2.5}$$

In our case, we can calculate $P_n^{|m|}(\sin(\theta))$ by using the initial conditions:

[16] In accordance to our coordinate system.

[17] Note here the use of l, m rather than ambisonic order and ambisonic degree.

$$
\begin{aligned}
P_0^0(\sin\theta) &= 1 \\
P_1^0(\sin\theta) &= \sin\theta \\
P_1^1(\sin\theta) &= -\cos\theta
\end{aligned}
\tag{2.6}
$$

[Wei] shows these three initial conditions for $x = cos(\theta)$. We can use the general equations for $P_n^{|m|}(x)$ to find the first three solutions to $P_n^{|m|}(\sin(\theta))$. The Pythagorean identity $\sin^2(\theta) + \cos^2(\theta) = 1$ shows that $P_1^1(\sin\theta) = -\cos\theta$. Equation 2.7 from [Wei] shows the general formula for associated Legendre polynomials for the first three terms, allowing us to set up our recurrence relation. E.g.

$$
\begin{aligned}
P_0^0(x) &= 1 \\
P_1^0(x) &= x \\
P_1^1(x) &= -\left(1 - x^2\right)^{1/2}
\end{aligned}
\tag{2.7}
$$

Condon-Shortley Phase One final consideration pertains to the polarity of the spherical harmonic, which is determined by the *Condon-Shortley* phase term - defined as $(-1)^m$, where m is the mathematical order[18]. Equations 2.5 and 2.7 both include this Condon-Shortley phase term. Nachbar [NZDS11], who authored the Ambix specification, suggests removing the Condon-Shortley term in ambisonic software development - as it only really simplifies matters in the field of mechanics.

Our third initial condition from Equation 2.6 still satisfies the Pythagorean theorem if the Condon-Shortley phase gets removed[19]. To prove this set $P_1^1(x) = \pm\left(1 - x^2\right)^{1/2}$ equal to $\pm cos(\theta)$, and set $x = sin(\theta)$. The sign of $P_1^1(x) = \pm\left(1 - x^2\right)^{1/2}$ is irrelevant since squaring both sides will force both arguments to be positive.

In our numerical computing library, whichever we chose, we will have to make sure to compensate for this $(-1)^m$ term. Alternatively, we can choose to write our own function that excludes the term. In order to compensate for the extra $(-1)^m$,

[18]Note that m always corresponds to the orientation of the harmonic.

[19]The phase change only occurs for odd harmonics, excluding the first harmonic, which is a special case - since $(-1)^0 = 1$.

we simply add another $(-1)^m$ to our recurrence equation[20]. This will make all our harmonics of the same polarity, regardless of the value of m.

2.3 Ambisonic Decoding

Decoding of ambisonic signals refers to the process of taking B-format spherical harmonics and linearly combining these to produce speaker signals. Ambisonics is best suited for regular loudspeaker layouts. Regular speaker layouts in the spherical domain include those modeled after the five *platonic solids*. These include the tetrahedron (4 faces), cube (6 faces), octahedron (8 faces), dodecahedron (12 faces), and icosahedron (20 faces). Each of the faces of these *convex regular polyhedra* is represented by a speaker in our layout.

Unfortunately, placing listeners at the center of such geometries, which would correspond to the ideal listening location, is often difficult - since this requires speakers below and above the listener. In very sophisticated listening environments, an acoustically transparent grid is often employed which allows speakers to be placed below the listener. An additional layer of complexity arises from determining the ideal circumference of the listening area. We ideally need to find a balance between the size of the reproduction area, and the proximity to any individual speaker.

A lot of research has been undertaken in the domain of ambisonic decoding for irregular layouts [HLB08]. The general process used to decode ambisonic signals involves computing the real-valued spherical harmonics for the speaker positions ($\mathbf{L}$) and subsequently calculating the decoding matrix by finding the inverse of this matrix ($\mathbf{D}$). The resultant matrix D is multiplied by the matrix of ambisonic signals $\mathbf{B}$ yielding the speaker feeds ($\mathbf{G}$). The size of the decoding matrix becomes L x $(N+1)^2$, where L is the number of loudspeakers and N is the *ambisonic order*.

Heller [HLB08] provides a magnificent introduction to decoder design. One of the

[20] Or use an `abs()` function, whichever is faster.

points that Heller raises is that some poor decoding designs, such as ones where all frequencies are decoded exactly the same, produce *comb-filtering*[21]. Instead, notable improvements can be made if the user performs *velocity decoding* at low frequencies, and *energy decoding* at high frequencies. *Near-Field Compensation* (NFC) can also be applied to sources that one wishes to be perceived as being inside the listening area. NFC is a method for adjusting the distance of the source that considers the behavior of low and high frequencies over arbitrary distances, namely, since high frequencies dissipate more quickly in the real world, these filters adjust the signals accordingly to simulate this phenomenon.

Three additional factors Heller mentions that may affect decoder quality include:

1. Room acoustics.
2. Accuracy of speaker positioning and matching[22].
3. Timing errors in multi-channel *Digital to Analog Converters* (DACs).

According to Heller [HLB08], a decoder must meet three criteria for proper reproduction:

1. Velocity and energy vector directions are the same at least up to 4kHz, and substantially unchanged with frequency.
2. At low frequencies, the magnitude of the velocity vector is near unity for all reproduced azimuths.
3. At mid/high frequencies the energy vector magnitude is substantially maximized across as large a part of the sound stage as possible.

[21]Phase errors caused by two complex signals being played simultaneously with a small delay between them.

[22]The frequency and phase response measured at the center of the array should be identical.

This section was intended to inform the reader of the various trade-offs and considerations resulting from the chosen algorithm. In practice, we have found the AllRAD decoder by IEM to be suitable and easy to use. This information can be useful to advanced users who may want to choose a specific decoder based on the nature of the music, or the acoustical properties of the space in which sounds are decoded. It is also intended for readers interested in developing ambisonic tools such as instruments, recording devices, or displays.

2.4 Selected Works

Having discussed briefly the technical elements of reproducing music via ambisonics, we now present a short list of works by the author that have utilized this technique. The first two of these works were presented at conferences internationally using real-world speaker systems and a live audience. In these two cases, we were fortunate that ambisonics provided us with a way to modify our playback mechanism for the speaker configuration. By sending the organizers of the event a .JSON decoder specification file, it was possible to reconfigure their DAW to the prescribed space minutes. The last piece was performed at UCSD in the experimental theatre using a 16-channel sound system as part of an ArtPower event with the Mivos Quartet.

2.4.1 Alternate Spaces

Alternate Spaces is a fixed-media work I created in my second year during my Ph.D. as part of the SElectOr series I have been producing and organizing[23]. The second year the group was active I decided to produce a series of concerts, unfortunately, because of the pandemic, I was not able to produce the third concert. This concert took place in the winter of 2020 and in it, various spatial music works were

[23] ucsdselector.com

presented including this piece. Other artists and students involved in this event included Tiange Zhou, Qing Qing Wang, Nicholas Solem, Leslie Fisher, and Alex Tung (all UCSD affiliated either as graduate students, undergraduate students, or alumni).

The music was created with a proprietary DAW but using all free plug-ins - this was prior to our transition toward FOSS. Videos were collected from various sources and edited to create a multimedia experience. The visual elements of the piece show a series of increasingly opulent and lavish scenes. The music in contrast is dark and brooding. The desired effect was to question the true human cost of these luxuries. At the peak of the experience, we see a burning forest, a landfill, images of riots, and finally a graveyard. These images serve as juxtaposing forces to the first half of the piece, revealing the damage excess can have, outside our field of view. The final image, of an ultrasound, is meant to symbolize hope, that a new generation will change these destructive patterns and realize a better future.

This piece was submitted to the Sound and Music Computing Conference in 2022 and was played in a 24-channel sound system for the attendees. The piece was submitted as an ambisonic file in third order using the Ambix specification along with a JSON file specifying the decoder. A Reaper session was also provided along with instructions on how to set up the video player in Reaper and how to load the decoder into the IEM plug-in. There is a clear duplicity in the message and the medium[24], however, this piece does not require a 24-channel system to be played back. Any stereo system is capable of reproducing the sound file since ambisonic is isotropic (e.g., treats all directions equally). To give the reader a taste of what this might sound and look like, a HOAST [DMKH+20] instance was created allowing the reader to view and listen to this piece with binaural audio. To listen to this piece as

[24]Since the composition is about opulence and 24-channel systems are very exclusive.

well as Chaqu visit this link[25]. The player does not work on Safari[26] but it should work on modern mobile devices as well as HMDs such as Oculus Rift.

2.4.2 Chaqu

Chaqu was initially composed in 2019 while I was taking a MAX/MSP course at UCSD to fulfill the Computer Music curriculum. Later the elements of the piece were imported into Reaper where a spatial audio mix was created. I also sourced videos from the internet to create an accompanying visual experience for the piece. The goal was to combine traditional Bolivian instruments with orchestral sounds and electronic synthesis. A proprietary DAW was used to compose the piece using free plug-ins, and later Reaper was employed, along with the IEM Plug-in Suite to create the HOA sound field.

This piece was submitted and performed at the NYCEMF in 2022 using a 16-channel sound system. It features a chacarera rhythm, quenas, zampoñas, and several classical instruments, as well as synthesizers of many styles. The music is sad and solemn, it is intended to evoke respect and wonder for the natural world. The indigenous populations of Bolivia consider Mother Nature their God, this piece is attempting to conjure the mysticism and beauty of this philosophy.

This piece is likely one of the only examples of Bolivian instruments and rhythms combined with electronics and orchestral sounds in an ambisonic context. In addition to the fixed-media work, it is also possible to perform several of the parts of this composition in real-time with the Reaper session. Using the MIDI data from the original session, a PDF score of the piece was rendered using free and open-source composition software.

Both *Chaqu* and *Alternate Spaces* were performed remotely - the author did not

[25]It is possible the link will be changed in the future. The material should still be found on the author's site.

[26]Firefox is ideal for WebXR.

attend the performances. This raises another issue about creating spatial music, which is the cost of travel and fees associated with conferences. We are hopeful that more venues are accepting remote submissions as we believe this will greatly increase the number of international artists submitting their works to NYCEMF - for example. However, the fees associated with these events are still high, creating a barrier to entry for those in other countries who are less wealthy.

2.4.3 Six Seasons

Six Seasons is a composition by Lei Liang that relies on a computer system I designed which allows a computer musician to control the sounds of the ocean with a MIDI controller. Inspired by past performances by Caroline Louise Miller, a UCSD alumnus, I opted to design the system around a simple and affordable MIDI controller. The first version of the patch was written in MAX/MSP but later ported to Pd. The patch uses 5.1 sound files generated by the MIAP software developed by Zachary Seldes [Sel14]. The source material for these surround sound files are hydrophone recordings acquired by the Whale Acoustics Lab at UCSD's Scripps Institute of Oceanography.

The MIDI controller allows one to cycle between different seasons of the Inuit calendar. Each season corresponds to a different folder containing at most eight sound files which are loaded onto custom sound file players (SFPs). The SFPs first encode the surround sound files into ambisonics, giving us greater playback flexibility and allowing us to apply additional spatial transforms to the signals (e.g., changing the height of the surround sound layer). The SFPs are triggered by the computer musician who can also control the reverb applied to these signals - using a combination of gain and reverb one can faithfully simulate distance effects. The system also allows the user to apply a delay and feedback upon the musicians' signals, of which there are four; the piece is written for a quartet. Each of these delayed copies

of the sound is panned using Lissajous figures of random frequencies which change each time the system is engaged incorporating another degree of indeterminacy to the piece.

The work is a culmination of multiple years of work bridging scientific and artistic contributions by oceanographers, composers, performers, engineers, and sound artists. After the debut of this piece in the experimental theatre the MAX/MSP patch was ported to Pd so that institutions with limited funds might be able to mount the production. There are three versions of the Pd patch: a sixteen-channel version, an octophonic version, and a quadraphonic version. The premiere of this piece counted on the cooperation of the acclaimed Mivos Quartet who responded to the sounds of ice, mammals, and wind, captured in the Northern Alaskan waters. See this link for code in Pd and MAX/MSP.[27]

2.5 Conclusion

This chapter has covered the principles of ambisonic encoding for sound field synthesis using spherical harmonics as basis functions, various decoding methods that can be used to reproduce this intermediary format in arbitrary loudspeaker systems, and three musical oeuvres in which the author was involved, which made use of this technology. The first of these works presented at SMC22 was decoded in St Etienne using a 24-channel sound system. The piece offers a contrast between the opulence and luxury of certain environments and the wretched and neglected state of others. The second piece is an homage to mother nature, which features Bolivian rhythms and instruments, accompanied by electronics and orchestral elements in an ambisonic composition. It was played at NYCEMF22 using a 16-channel sound system at the Sheen Center Black Box Theatre. The final composition, belonging to

[27] A full paper was recently submitted to the journal of the International Computer Music Association ICMA, awaiting review.

Lei Liang, features an instrument I designed that allows the composer to manipulate hydrophone recordings in real-time. The sound was played back using a 16-channel sound system in CPMC122 in coordination with ArtPower. Two performances of the piece were mounted in which the Mivos quartet improvised along with these recordings decoded MAX/MSP but later ported via free and open source ambisonics software in Puredata for greater accessibility.

Chapter 3

Capturing Spatial Music

3.1 Introduction

During the 21st century, we have seen a renewed interest in the field of virtual spatial acoustics, as investments in *Extended Reality* (XR) systems have exploded. With the proliferation of *Head Mounted Displays* (HMDs) as entertainment systems for personal use, several composers and sound engineers have expectedly begun researching the use of *Spherical Microphone Arrays* (SMA) towards musical ends.

Generally, although not exclusively, SMAs are associated with a technology known as ambisonics, developed in the 1970s by Michael Gerzon et al. The initial formulations of the system denominated *First Order Ambisonics* (FOA), consisted of deriving three orthogonal figure-8 microphones to encode sound particle velocity in three-dimensional space. This spatial audio system is based on the theory of *spherical harmonics* (SHs), also used in various other fields including computer graphics and astronomy.

Spherical harmonics can be considered basis functions of 3D spherical phenomena, much the same way that simple trigonometric functions can be considered bases for 2D phenomena. The well-known *Discrete Fourier Transform* (DFT) is used

to decompose complex sound events into a sum of sin and cos waves of varying amplitude, phase, and frequency. Much the same way, spherical harmonic theory aims at decomposing spherical sound phenomena using a combination of spherical harmonics of increasing order.

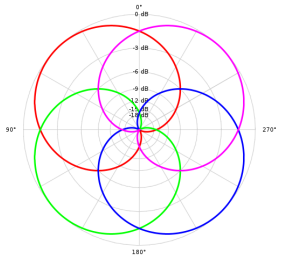
A lot of work has been done in the last 20 years to extend FOA into *Higher Order Ambisonics* (HOA), providing a larger sweet spot[1], and improved directionality. Various microphone designs have been documented and evaluated in academic publications, however, very few have been documented to the extent that these can be recreated by people without extensive technical backgrounds. In the commercial sector, a few FOA and HOA solutions exist, however, most of these remain economically unviable for most artists.

This work aims to present existing HOA microphone designs, considering principally those that favor low-cost and high-quality in the design process. The present author has already been involved in the process of constructing a low-cost FOA microphone using *Micro-Electronic Mechanical Sensors* (MEMS). Unlike traditional *Electret Condenser Microphones* (ECMs), MEMS capsules may come equipped with *Analog-to-Digital-Converters* (ADCs), allowing direct integration with *Micro-Controller Units* (MCUs). The reduced cost of these MCUs over studio ADCs makes these capsules ideal for anybody building an HOA array on a budget.

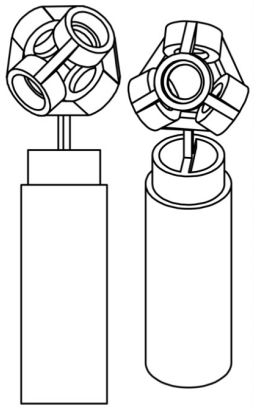
3.1.1 Sound-field Microphones

First Order Ambisonic (FOA) microphones were developed initially in the '70s by Michael Gerzon, from Oxford's Department of Mathematics, and Peter Fellgett, from the University of Reading [Ele91], under the supervision of the National Research Development Corporation (NRDC) in England. Gerzon and Fellgett were indubitably inspired by Alan Blumlein's work in the field of stereophony.

[1] Area within which audience members can sit and get a coherent sound image.



(a) FOA Patterns (Top View) [Net]



(b) SolidWorks Drawing - FOA Microphone

Figure 3.1: FOA Polar Patterns (Top View) and CAD Drawing of FOA Mic

Blumlein, another Englishman, was the pioneer of various audio technologies in the '70s. Chief among these was a sum-and-difference matrix which was capable of creating high-quality stereo recordings. This technique, denominated *Mid-Side Stereo*, derives two cardioid patterns from the combination of one omnidirectional microphone, closely positioned (e.g. coincidently positioned) with a figure-8 microphone. Gerzon and Felgett envisioned a 3D sound system that could be created by adding two additional figure-8 microphones, each representing a different axis in 3D space.

Instead of using three velocity microphones (e.g. figure-8 microphones), however, Gerzon and Fellgett designed a way to derive these three responses from four cardioid sensors mounted on the faces of a tetrahedron - using a sum-and-difference matrix akin to Blumlein's. Consider Figure 3.1; on the left, we can see the polar pattern of the FOA mic from a birds-eye view, on the right, we see a *Computer Assisted Design* (CAD) drawing of an actual FOA microphone[2].

[2]Created by Yigal Kamel from "The Cooper Union for the Advancement of Science and Art" in SolidWorks.

As we can see from this top view, three orthogonal figure-8 microphones can be derived from the sum and difference of these different cardioid capsules. For example, to derive the front-back microphone, we sum the front capsules and subtract the back capsules - the sides will cancel out and we will be left with a figure-8 response.

In addition to the three orthogonal[3] figure-8 microphone signals, an additional omni-direction channel, often denoted with the letter W, can be encoded by summing the four cardioid capsules - we call this a pressure signal. Once these four signals are recorded, they can be *decoded*[4] into any number of different reproduction systems. For example, the combination of signals can be used to yield a stereo representation with good mid/side separation[5]. To do this simply add the positive area of two figure-8 signals (X/Y) and combine the result with the W signal, then repeat for the other channel, changing signs accordingly.

Unfortunately, one of the main problems with ambisonics, or any surround sound reproduction system for that matter, is correct speaker placement. As Elen [Ele91] points out:

> "Most readers, I am sure, have visited numerous friends who keep their stereo loudspeakers in some very odd places – one channel behind the sofa and the other on top of the bookcase, for example. It's hard enough to get people to put two speakers in sensible places for stereo – what about four for surround sound?"

Luckily, many modern advancements in technology allow us today to record and reproduce ambisonic music seamlessly. Large speaker arrays can be substituted with binaural synthesis processes, reducing both the cost and complexity of ambisonic reproduction. Ambisonic microphones are a particularly attractive recording technology due to their simplicity compared to the encoding[6] of individual sources. A

[3]E.g. perpendicular.

[4]Decoding refers to using linear combinations of SHs to create speaker or headphone signals.

[5]The UHJ system, also known as "C-format", was an ambisonic system designed to be compatible with mono and stereo media.

[6]Encoding corresponds to transforming raw audio signals into spherical harmonics (SHs). The term encoding also refers to transforming raw HOA array signals into SHs.

good ambisonic microphone should only need to be gain-adjusted, for clear and realistic reproduction of spatial audio.

Capsule Ordering

A common problem with sound-field microphone design can be the confusing naming and labeling of capsules, which is crucial to proper encoding. In simple FOA microphones, various authors have adopted the naming: FLU, FRD, BLD, and BRU (where F/B corresponds to front/back, L/R corresponds to left/right, and U/D corresponds to up/down). For HOA, alternate naming systems must be developed and maintained throughout calibration and encoding. This can be especially complicated at higher orders, where designs might feature 64 capsules or more.

Harmony must also be maintained between the ambisonic signal *ordering* in the encoder and decoder. SH ordering refers to the indexing of SHs in the final B-format signal[7]. Meta-data can be used to determine if the sound file is ordered according to a particular standard and if other relevant processing has been introduced to the signals which the decoder might need to account for[8].

3.2 Literature Review

Middlicott et al. Middlicott et al. [MW19] recently published a paper regarding HOA microphone calibration approaches, in which they described the fabrication of a planar[9] ambisonic array. The main theme of the paper is not the fabrication of the microphone itself, but rather evaluating different calibration approaches for A-format signals in HOA. The A-format signals are subsequently encoded to determine how much the pre-encoding calibration improves the response of spherical harmonics.

[7] B-format is the name given to audio that has been encoded.

[8] For example, if *Near Field Compensation* (NFC) filters have been applied.

[9] It is cylindrical but with an open design, we call it planar because all capsules lie on a single plane.

A number of important procedures are elucidated by the authors. Like many other authors involved in this work, the first consideration is the room acoustics. For this experiment, a *hemi-anechoic* room was used (e.g. not perfectly anechoic). As a result, some processing will need to be applied to the measured IRs. Lopez-Lezcano [LL19] describes windowing applied to the measurements in order to reduce reflection effects from room surfaces. This entails: measuring the close surface to the microphone, calculating in samples that amount of time, trimming the IR, and using a windowing function to remove discontinuities - which might cause FFT artifacts. This process will be covered in more detail in Section 3.2.

The authors noted that only half of the impulse responses were necessary for analysis given the symmetry of polar patterns[10]. This effectively cuts the measurement time in half. The authors also noted the necessity to compensate for the speaker's response. For this, a flat frequency response microphone is used, and, as is common, the Nelson/Kirkeby inversion approach is implemented. One should also trim and window this impulse response to compensate for the room's acoustics. Equation 3.1 shows the Nelson/Kirkeby algorithm, used to generate the inverse filter (e.g. the compensation filter), including a frequency-dependent regularization parameter.

$$H_{inv}(\mathrm{k}) = \frac{H_{\mathrm{Target}}\,(k) \cdot \mathrm{Conj}(H(k))}{\mathrm{Conj}(H(k)) \cdot H(k) + \epsilon(k)} \tag{3.1}$$

where:

$H_{inv}(\mathrm{k})$ is the inverted filter,

$H_{\mathrm{Target}}\,(k)$ is the ideal frequency response,

$H(k)$ is the measured amplitude response, and

$\epsilon(k)$ is a regularization parameter.

[10]The same principle was used in our own work [Zal19]. Unfortunately, assuming symmetry can lead to errors.

The value, k is the frequency index (e.g. the bin), and the $conj()$ operator corresponds to the *complex conjugate*. Without the regularization parameter, the inversion of the filter is idealized, but this can lead to over-emphasis in certain regions of the spectra. The authors use a vector with various regularization parameters depending on the region of interest for A-format calibration but seem to suggest using no regularization for the speaker inversion. The resultant inverse filter is convolved with each A-format IR to negate the effect of the speaker. In this case, the target filter would most likely be the FFT of a delta function, which is perfectly flat. Another approach is using software such as DRC[11] to generate such a filter.

The remainder of the paper discusses four different approaches used for calibrating the capsules, which produce our A-format signals. These four approaches are called:

1. Calibration by 1/3rd Octave Average Gain Matching,
2. Calibration to a Specific Capsules On-Axis Frequency Response,
3. Calibration to a Flat Frequency Response, and
4. Calibration by Diffuse Field Equalization.

The first method involves using just one on-axis response for each capsule and splitting the FFT into 1/3rd octaves[12]. The average magnitude is measured for all these sub-bands and then the lowest value is used to attenuate all other microphones to that same level.

The second method involves selecting the capsule with the best response by visual inspection and subsequently trying to match all other capsules using Equation 3.1. In this method, once more, a single IR for each capsule is required. Note that this method means user input is required (e.g. this process cannot be automated without

[11]Digital Room Correction.

[12]In electronics, an octave is a doubling of frequency. For example, between 40 and 80Hz there is exactly one octave.

some modifications). One could alternatively calculate the deviation from a flat frequency response, instead of selecting the target response by visual inspection.

The third method is very similar to the second, however, rather than using an on-axis response from one of the capsules, all the capsules are calibrated using a perfectly flat frequency response as the target. The frequency response of the *Dirac delta function* (δ) function serves as this target. Much like with the second method, frequency-dependent regularization is employed (e.g. the $\epsilon(k)$ parameter changes over the frequency range).

The last method, called Diffuse Field Equalization, is the most time-consuming of all, as it requires the full range of IRs. Much like with the other methods, each capsule gets a customized/individual calibration filter. However, to find our target filter, we take all measurements (e.g. all directions), for each capsule, and average them. In this case, only horizontal measurements were used. Middlicott et al. provide the following equation to generate the *Diffuse Field Response* (DFR).

$$\mathrm{DFR(c)} = \sqrt{\frac{1}{D}\sum_{d=1}^{D}\left|FFT_{(c,d,k)}\right|^2} \tag{3.2}$$

where:

DFR(c) is the Diffuse Field Response for capsule c,

d is the measurement position index (1 to D), and,

k is the frequency bin.

As we can see, this equation is closely related to the RMS[13], although, in this case, we use these values to derive a filter over specific frequency bins. With these filters acquired, the authors use the frequency-dependent regularized inverse filter technique, much like with methods 2 and 3, with the $DFR(c)$ filters as targets. We

[13]Root Mean Square.

should note that the author does not make any mention of the phase of the inverse filters, only the magnitude is considered[14].

After the four methods were implemented, the author undertook an evaluation of these different A-format calibration approaches by examining the spherical harmonic polar plots. Before doing so, however, they also generated simulated B-format polar plots using perfect cardioid responses. Even then, the simulated harmonics differed greatly from the ideal harmonic due to the distance between transducers. The authors' particular design featured a 25mm radius and 72° capsule spacing. The calculated spatial aliasing frequency was found to be 4.3kHz using the following equation [15]:

$$\text{Aliasing Frequency } = \frac{Nc}{2\pi r} \tag{3.3}$$

where N is the ambisonic order (2OA in this case)[16], c is the speed of sound, and r is the array radius. The authors also plot the uncalibrated B-format polar responses for comparison with simulated, ideal, and calibrated responses. The ideal responses correspond to perfect coincidence and response, which is impossible in real-world conditions.

In order to encode the raw A-format signals captured directly from the mic into the SH domain the, real-valued SHs are evaluated at the capsule angles according to the array geometry. Since this particular array is uniform, the frequency-independent SH encoding matrix can be applied as is to generate the B-format signals.

Figure 3.2 shows an example using FOA SHs. The matrix of SHs contains four rows corresponding to four capsule positions, and four columns corresponding to a particular ambisonic order and degree. For example, column 0 is the W channel

[14] In [LL18] filters are modified for minimum-phase.

[15] This theoretical frequency has been found to be imperfect in other publications. Visual inspection can be used to determine transition frequency [LL19].

[16] For 3D sound capture of 2OA we'd generally need 9 capsules. This design is for horizontal 2OA (or cylindrical harmonics only). The general formula for this is $(2N + 1)$ capsules are required, rather than $(N + 1)^2$.

which equals 1 for any ϕ or θ. So W will simply be $c1 + c2 + c3 + c4$. Column 1 in ACN is Y, which corresponds to the equation $sin(\phi)cos(\theta)$. We evaluate this SH using all four capsules' polar coordinates in radians, etc. The capsule's signals are multiplied by each element in the corresponding row, and summed to generate the desired signals.

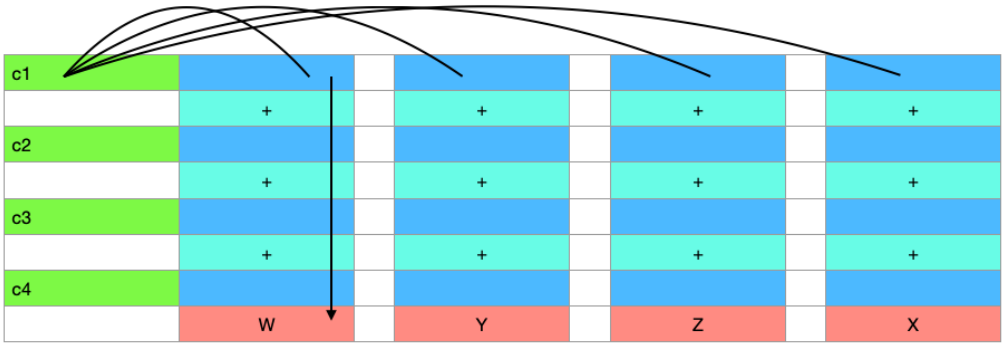


Figure 3.2: FOA Mic Encoding Diagram

Middlicott et al. concluded that the DFR method (method 4) and the average gain matching (method 1) yielded the best results. Considering that method 1 requires far fewer measurements and yields similar results, it would be interesting to evaluate the time savings this could result in - especially in an automated system, which is ultimately desired.

Unfortunately, in this publication, all the measurements were taken only on the horizontal plane, and there was no subjective evaluation of the systems proposed; a subjective evaluation of methods 1 and 4 could help decide which method is superior, if any. Finally, the authors noted the need for post-filtering after encoding to further improve the integrity of the spherical harmonics at all frequencies. The figures in the publication also revealed that despite the absence of radial filters, such as those described in [BPSW11], proper SH responses were achievable at multiple frequencies.

This paper offers great insight into the calibration of capsules for SMAs. Unfortunately, the design of the microphone could not be found. The capsules used were

the JLI-140A-T, which only cost about $5 per capsule. Although these capsules have many attractive features, for HOA arrays, we would likely need a costly ADC to record the analog output of these cardioid mics. In addition, the current design is only capable of sampling SH in 2D, and the parts are not optimized for low-cost 3d printing.

Modular Spherical Microphone Array (MSMA) Gonzáles et al. [GPL18] described in 2018 a design the authors defined as modular, given the system's ability to change radius on the fly. Their design can be considered both multi-tiered or rigid, since in certain configurations it can have more than one operating radius. The design has a rigid body, but capsules protrude from the main sphere in varying amounts, depending on the attachment selected. The project site can be found here[17]. In the cited publication, the authors compared the performance of their design in three configurations: small radius, large radius, and mixed radius.

Some of the key features of this design, in addition to the modularity aspect of the microphone, are the use of FOSS for the *Computer Assisted Design* (CAD)[18] and the use of MEMS microphones, in lieu of ECMs. Unfortunately, the code used to generate the CAD models does not seem to be accessible, however, the files for manufacturing the *Printed Circuit Boards* (PCBs), and the rendered 3D designs can be downloaded and modified.

To validate their designs, the authors used a simulation approach in which they compared a simulation of their system to the ideal SH representation. The "Spherical Array Processing" library by Archontis Politis, publicly downloadable from GitHub[19], and explained in his doctoral dissertation [P+16], was used for this assessment. The spatial correlation measured was given by the following equation:

[17] Accessed: Feb. 2021.

[18] The authors used OpenSCAD, which is distinct from other systems in that the user programs the design using code, rather than manual manipulations.

[19] Link to repository.

$$R_{|S^{sim}|,|S^{SH}|} = \frac{\left\langle S^{sim}, S^{SH} \right\rangle}{\|S^{sim}\| \cdot \|S^{SH}\|}. \tag{3.4}$$

The top part of the fraction involves taking the dot product of the simulated and ideal spherical harmonics. The bottom part involves taking the norm of each vector and then calculating its product - these are both classic operations in linear algebra. The spatial correlation is calculated over the range of audible frequencies for each ambisonic order. For the three given arrangements (e.g. small radius, big radius, and mixed radius), it was found that a smaller radius provided the best spatial correlation. Unfortunately, these simulations can only give us a prediction of the performance of this system. Ideally, the authors would have also measured the response of the microphone in anechoic conditions and performed some listening experiments.

Notwithstanding, the design remains promising given that the parts used for its manufacturing can be printed with low-cost 3D printers. This *Modular Spherical Microphone Array* (MSMA) also features MEMS capsules, making it more affordable to reproduce than other designs. Unfortunately, the 19 analog capsules require a rather costly interface in order to record the signals. MEMS microphones with internal ADCs might provide a cheaper alternative than the proposed analog MEMS devices. In addition, the selected capsules in the MSMA design do not offer the best SNR available for analog MEMS capsules.

Higher Order Spherical Mic Array (HOSMA) TH Köln[20] has published several papers describing the development of a rigid SMA. In [MDLP20] Moschner et al. described their latest efforts to create a 7th Order Ambisonic (7OA) microphone with a total of 64 capsules. As the authors noted, spherical harmonic processing in VST plug-ins is limited to 64 channels (e.g., in SPARTA or IEM suite). There are

[20]TH Köln - University of Applied Sciences, is an institute of higher education located in Cologne, Germany.

currently no 7OA microphones on the market, the highest-order commercial system is the Eigenmike by *mh acoustics* which is of 4th order.

Similar to Gonzáles et al. [GPL18], this microphone system has an inner sphere to which various tubes are attached, however, in the case of the design by Moschner et al., the 64 channels are all comprised of ECM capsules instead of MEMS microphones, and there is an additional rigid baffle which changes the acoustics of the array[21].

HOSMA uses the same capsules as the commercial FOA microphone by Sennheiser called Ambeo VR Mic[22]. The CAD modeling is done in Blender and the analysis is done using the SOFiA toolbox[23].

Another similarity with the Gonzalez et al. publication is that partially automated modeling is possible using Blender and Python scripts. Automated modeling is done in [GPL18] and [LL19] using the OpenSCAD modeling software, which uses a strictly scripting-only approach to 3D modeling. This means other researchers can easily re-purpose the design for systems with fewer microphones, smaller radii, or different capsules[24].

The HOSMA design is based on a Fliege Grid [FM99], created by Jörg Fliege from the University of Southampton's School of Mathematics. The integration formulae for this grid can be found here. There are coordinates and weights for spheres with 4, 9, 16, 25,..., 900 nodes. According to Moschner et al. [MDLP20], this sampling grid allows an "optimal spatial resolution can be archived with a minimum number of microphones." The Fliege grid is particularly useful for ambisonics because, as we can see, the number of points on the sphere follows the $(N+1)^2$ requirements for ambisonic spherical sampling.

According to [Arc]: "the Fliege-Maier nodes are another example of nearly-

[21]Rather than a rigid array with protruding capsules, it becomes a rigid array with flush mounted capsules.

[22]It is unclear from our search if these capsules are commercially available, this makes the experiment difficult to replicate.

[23]Link to HOSMA repository (Accessed: Feb 26, 2021)

[24]The authors note the process is not fully automated in HOSMA.

uniform arrangements that along with their respective integration weights can be used for direct integration through summation." This is one of four "special" sampling arrangements[25], which allow for direct integration. When the function is sampled using non-uniform nodes, then a least squares solution is necessary to encode the array - such as in [LL19].

The parts for this project were 3D printed using *Fused Filament Fabrication* (FFF)[26] via low-cost 3D printers. This is one of the considerations also addressed by [LL19] and [GPL18]. Two other, more robust, 3D printing technologies are SLS and SLA. SLA is used for prototyping; these printers cost around $500. SLS is used for final products and uses stronger materials. SLS printers cost around 5,000 dollars, which not all institutions might have access to. FFF printers vary in price depending on the quality of the print - a cheap one can cost $300. FFF, therefore, is the most accessible type of 3D printing. Unfortunately, it also means designs must be geometrically simpler, and will not be as strong[27].

The HOSMA design uses the Sennheiser KE14 ECMs and PCBs with a dedicated voltage divider and converter, which outputs a balanced signal to a standard XLR-pin connector. Additionally, for improved EMI-shielding[28], copper tape and graphite spray are applied to the tubes enclosing the PCBs. In [DLAP19] by Dziwis et al. (e.g. the first publication associated with this project), the EMI shielding was measured to quantify any improvements. It was shown that the 50Hz hum produced by a voltage outlet was mitigated by using these methods. The authors also used a conductive PLA[29] 3D printing material to further improve EMI shielding.

The most expensive part of this build is unfortunately the recording interface. HOSMA requires 8 microphone amplifiers, each converting eight A-format signals to

[25]Including Gauss-Legendre quadratures, Lebedev grids, and t-designs.

[26]Also known as Fused Deposition Modeling.

[27]After prototyping is complete, one can also send their designs to a 3D printing service, which can return stronger prints at relatively low prices.

[28]Electro-Magnetic Interference.

[29]Polylactic Acid.

the ADAT format. Another interface converts eight ADAT streams to the MADI format for recording. This makes the ultimate design, unfortunately, well outside the price range of independent musicians.

The final part of the paper describes the evaluation of the system. The array was measured via impulse response methods in an anechoic chamber at TH Köln. For comparison, Moschner et al. also simulated a free-field plane wave impinging upon a virtual array with the same Fliege grid. For this simulation, the SOFiA (Sound Field Analysis) Matlab toolbox was employed [BPSW11]. The simulation and real measurement were well-matched for the frequency of 3kHz. This design is expected to show spatial aliasing artifacts above 3.2kHz. In this paper, the simulation and recorded IR are of a single direction.

The authors employ a *Plane Wave Decomposition* (PWD) method to analyze the results. According to [PPQ16] in spherical beam-forming literature, PWD refers to higher-order hyper-cardioids used to "maximize the directivity factor for a given order"[30].

Radial filters in this publication were limited to 0 dB. These are used to compensate for any scattering occurring from the surface of the sphere and also account for the directionality of the sensors. These radial filters are implemented in the SOFiA toolbox.

As noted in Lösler and Zotter [LZ15], simply using a frequency-independent matrix to encode A-format signals into the SH domain does not yield proper ambisonic recordings. This is because "low-frequency signals predominantly map to a non-directional pattern on the surface of the microphone array". This, in turn, causes higher-order signals to require amplification. Holographic/radial filters are therefore required to equalize the B-format signals.

Radial filters are designed based on the nature of the microphone array (open/-rigid), and the individual capsules' responses (omni/cardioid). The formulas them-

[30]This means essentially that the analysis is that of a virtual microphone, not the full sound-field.

selves are based on spherical Bessel and Hankel functions, both of which describe sound propagation in a frequency-dependent manner. According to McCormack et al. [MDMF+18]:

> "The modal coefficients[31] take into account whether the array construction is open or rigid and the directivity of the sensors (cardioid, dipole, or omnidirectional). In the case of a rigid baffle construction, these modal coefficients can also take into consideration the frequency-dependent acoustical admittance of the surface of the array and also the distance between the surface of the array and the sensors; in order to cater for cases in which the sensors protrude from the surface of the array."

To derive these filters, we first need to define and calculate the spherical Bessel functions of the first and second kinds. All of these are related to Bessel functions and are all solutions to the same differential equation [Teu07]. They can also all be defined in terms of Bessel functions of the first kind, which are often called cylinder functions. These are described in Appendix A.

Using these spherical Bessel functions various radial filters can be implemented, based on the configuration of the SMA. Bernschutz et al. [BPSW11][32] provide good formulations for some of the filters for various configurations. McCormack et al. [MDMF+18] provide more information regarding the inversion methods examined in the literature, as the actual filters we will need to use are the inverse of these functions - which require regularization to limit sensor noise amplification.

The HOSMA analysis revealed that the array performs similarly to the simulation at some frequencies but deviates substantially at others. Notably, the array does not perform well at 500Hz, well below the aliasing frequency of 3.2kHz. This could perhaps be improved with some equalization of the B-format signals. The other two reported measurements at 3kHz and 7kHz appear to match the simulation.

While this design is promising given the very high order of the array, and high quality of the components used, it is not suitable for independent musicians or sound

[31] Of the radial filters.

[32] See Section 3.2 of [BPSW11].

engineers working on a budget. Additionally, certain parts do not seem to be available without a partnership with the manufacturer, which makes this project unreproducible for most people.

Nonetheless, the Fliege grid design does offer a great HOA design that could be modified and used with other capsules; one could swap these out for MEMS to create a cheap version of this array. The license of the project, and the use of open-source tools, such as Blender and Python, make this possible. The SOFiA toolbox has recently been partially ported to Python [HA17], which means one does not need a MATLAB license to perform these simulations.

Spherical Harmonic EAR (SpHEAR) Fernando Lopez-Lezcano, from CCRMA, has also written about Free and Open Source SMA designs in [LL16] [LL18] [LL19]. To this date, two designs have been proposed by the author: the TinySpHEAR, and the Octathingy. The TinySpHEAR is a FOA microphone featuring four capsules in a tetrahedral configuration, while the Octathingy, inspired by [Ben12], is a 2OA microphone with 8 capsules arranged in octahedral geometry - capable of sampling 8 out of the 9 SHs required for 2OA.

One of the key features of the SpHEAR project is the OpenSCAD routines written for the generation of 3D models. The OpenSCAD scripting language, as already noted, allows researchers to modify existing models to fit the dimensions of different capsules, and also adjust the array radius without hundreds of manual modifications. The code for the OpenSCAD modeling of SpHEAR is public and can fully automate changes to the design.

The second important part of this work is the Octave GNU calibration routines which allow one to encode microphone signals into the SH domain, and also perform equalization of both A and B format responses. Finally, the author also provides PCB files and diagrams describing how to create these designs, and the accompanying publications explain how to calibrate the systems.

In [LL18], Lopez-Lezcano describes the use of a 5-degree-of-freedom robotic arm for the calibration of the Octathingy[33]. The arm allows automatic measurement of the system with 4096 different points of resolution. With 5 degrees of freedom, the author can calibrate the microphone using either horizontal-only IRs or a combination of horizontal and vertical IRs. Certain spherical harmonics are not well suited for calibration using horizontal measurements only, as the response is completely null at $\theta = 0°$[34].

Mechanical Design and 3D Models One of the key features of this microphone array, similar to [GPL18], is the ability to 3D print these designs using fused filament fabrication (FFF)[35] - instead of more expensive methods such as SLA or SLS. This allows institutions with cheaper equipment to reproduce these designs without having to outsource the printing. The designs are created using FOSS, and the entire project is licensed under GPL/CC licenses.

In [LL19] Lopez-Lezcano shared an update on the project. In this publication, the author described various mechanical designs which were measured and acoustically validated. A second version of the capsule holder was designed to improve the cardioid response of the capsules - which was perturbed by the enclosure. The author noted that the space inside the initial array (shown in Figure **??**) resulted in the offending resonance. This affected the response of the sensors at the frequency corresponding to the size of the cavity.

As a result, in the second design, a new geometry was designed, which no longer featured an open array configuration, but instead positions each sensor at the base of a cone attached to a rigid octahedron. Figure 3.3 shows Octathingy Version 2, featuring this new design. Each cone has apertures for incident sounds arriving from

[33]The robotic arm costs $2,500.
[34]Consider for example the Z harmonic. θ is our elevation variable.
[35]Also known as extrusion method.

behind the capsules[36]. Unfortunately, although the individual capsule response of this new design is better, it also requires a slightly larger overall radius - resulting in a lower spatial aliasing frequency. The TinySpHEAR is based on the capsule holders in Version 1.

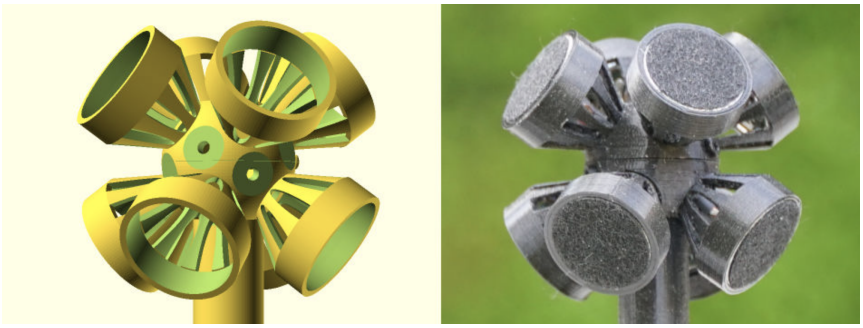


Figure 3.3: Octathingy Version 2 [LL19]

Electronics Both the TinySpHEAR and Octathingy rely on active ECMs requiring phantom power for operation. The operating voltage of sensors varies depending on the package and must be suitably powered for proper performance. Both the 10mm Primo EM182 and 14mm Primo EM200, used in these designs, have an operating voltage of 5V, which means the standard 48V output from phantom power needs to be stepped down[37].

In both the FOA and 2OA designs, this is accomplished using a Zapnspark circuit. The larger 14mm capsules were reported to provide better low-end performance, highlighting a trade-off between capsule diameter, noise performance, and array radius. Larger diameter capsules have been shown to have better noise performance, which becomes important during the SH equalization step.

The Octathingy[38] uses a standard DB25 connector, as opposed to the 12-pin Amphenol DIN connector from the TinySpHEAR, which is not common in audio

[36]The desired directivity is that of a cardioid microphone.
[37]The operating voltage can be found in their respective specification sheets.
[38]Also called OctaSpHEAR.

applications. The ECMs used in this project can be found online and are of exceptional quality, however, they are also quite expensive - reducing the availability of this system.

Calibration By far the most laborious aspect of designing a proper HOA microphone is the calibration process. There are several steps involved for complete HOA array calibration. Below is a reduced list, describing the minimum requirements:

1. Determine truncation size based on the shortest reflection path (unless an anechoic chamber is available).
2. Create a compensation filter for speaker response, taking into account the measurement from step 1.
3. Find suitable systems for IR generation and acquisition. The system should ideally be automated.
4. Determine how many poses are needed for A-format equalization and generate A-format correction filters.
5. Encode corrected A-format signals into the SH domain after they have been equalized.
6. Determine how many poses are needed for B-format measurements, and analyze B-format responses to create SH equalization filters[39].
7. Ideally, download all filters and apply them via a VST or other easy-to-use tool.

[39] Alternatively, use a modeling approach to equalization.

Determine Truncation Size

Much like with other authors (i.e. [MW19]), the first step towards calibration is to simulate a free-field environment by compensating for the effects of the speaker and room. The shortest reflection path will determine the truncation size of our IRs[40], as well as the minimum frequency for reliable measurements. In [LL18], the 4.5 msec direct path translates to a 220Hz limit[41]. This shortest reflection path will also determine the quality of the compensation filter we create for the speaker response.

Create a Compensation Filter for Speaker Response

One may either choose to create this filter themselves[42], or use a software library, such as Digital Room Correction (DRC) to create this filter. Providing as much detail as possible regarding room dimensions, reverb time, and noise floor will help analyze the final design. If an anechoic chamber is not available, then, in order to achieve accurate results, a large room with little to no reflections, and low noise floor, is suitable.

Using the truncation time, re-calculated in samples, all measurements will need to be trimmed after the impulse response has been extracted from an *Exponential Sine Sweep* (ESS)[43]. Canfield-Dafilou et al. [CDCJ18] note that the ESS should ideally be re-recorded using a direct line to the audio interface, to compensate for any effects the interface might have on the signal[44]. This, in practice, is often not performed because of IO[45] limitations. Instead, we consider the interface's effects to be negligible, especially if the bit rate is high. Thus, in practice, the sine-sweep is commonly saved in memory until it is needed for deconvolution. Equation 3.5 shows

[40]Shortest reflection path corresponds to the closest surface relative to the sensor array.

[41]$1/f = T$ corresponds to the period of the longest wave. We can use $f/c = 1/\lambda$, where λ is the wavelength, to calculate f based on the measured distance. For Lopez-Lezcano it was 1.5 meters roughly or 5 feet.

[42]Using, for example, JOS's imp_meas.

[43]The ESS technique is the most common method today for capturing IRs.

[44]The interface belongs to the transfer function path.

[45]Input/output.

the deconvolution process.

$$h(t) = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(r(t))}{\mathcal{F}(c(t))}\right) \tag{3.5}$$

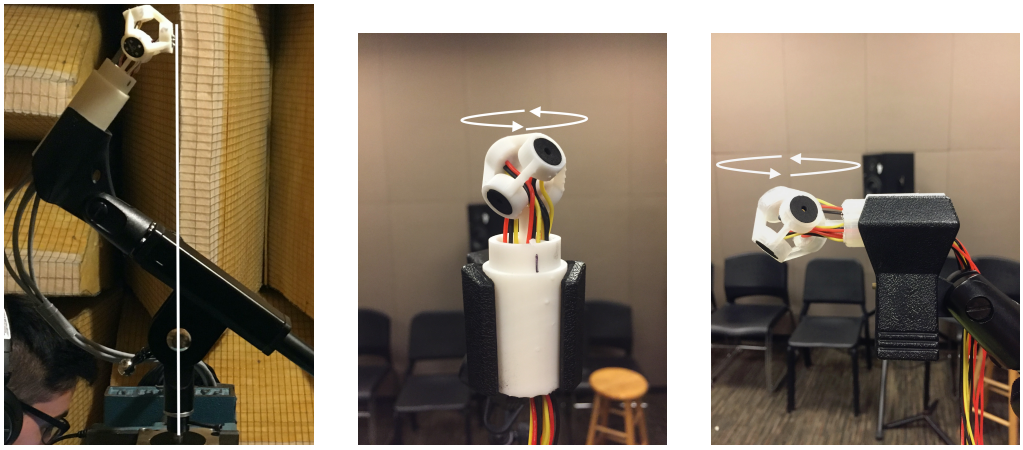
In this equation $h(t)$ corresponds to the IR in the time domain, $\mathcal{F}^{-1}$ is the iFFT operation, $\mathcal{F}(r(t))$ corresponds to the FFT of the recorded signal, and $\mathcal{F}(c(t))$ corresponds to the FFT of the original ESS.

Since convolution in the frequency domain corresponds to multiplication, deconvolution corresponds to division. Once the response of the speaker has been found using a flat microphone, an inverse filter can be generated, such as in [MW19], using the Nelson/Kirkeby method. Lopez-Lezcano does not use this method, but instead inverts the response directly, and subsequently uses the*Minimum-Phase Spectrum* (MPS) method - by Julius O. Smith [Con]. Alternatively, Lopez-Lezcano uses the Digital Room Correction software package.

After the IR is found, it is truncated - and a windowing centered at the impulse is applied to avoid discontinuities in future FFT analyses. The same will be done for all future IRs captured by the array. Lopez-Lezcano recommends using a Blackman window as he states it provides: "empirically better ripple in the passband, which translates to flatter response at low frequencies." The speaker compensation filter will be convolved with all subsequent measurements, negating the effects of the speaker.

Build an Automated IR Acquisition System

Once the compensation filter for the speaker response has been generated, we must acquire a system capable of capturing a dense grid of IRs to analyze and ultimately equalize our SMA. Lopez-Lezcano [LL18] uses SuperCollider (SC) to send serial data to the aforementioned robotic arm and then transfers the recordings to Octave GNU for analysis and filter design. The robotic arm uses an "Arduino-based built-in control processor". Unfortunately, the SC code was not found, however, recreating it would likely not be too hard since all the synthesis involves is an ESS.



(a) Pose 1: Single Capsule - Horizontal (b) Pose 2: Full Array - Horizontal (c) Pose 3: Full Array - Vertical

Figure 3.4: Three Poses for SMA Calibration

The other option is to use Aliki, which is a Free and Open Source IR generator, however, there is no direct way we know to interface Aliki with Arduino; This means the person has to manually move the array, possibly hundreds of times, which could lead to substantial errors. This process would also take a very long time, especially if the sampling grid is very dense. In Section 3.3, we will talk about the solution the present author has implemented in the past, and what possible changes we adopt to make it more accessible.

Calibrate A-format IRs

Once a proper IR measurement system has been acquired, and the IRs of the array have been captured, the next step is to create the equalization filters for the A-format signals. There are two purposes to this: to make the capsule responses flatter and to make the capsule responses match each other. In contrast to cylindrical arrays, such as the one in [MW19], for spherical arrays, we need to use both horizontal and vertical responses.

Let's consider the *pose* of the array - with respect to the speaker - and the

axis of rotation (e.g. the center of gravity), in order to demonstrate this process. Figure 3.4 shows the three poses the present author has previously implemented in the field. Image 3.4a shows an FOA microphone array centered for single capsule measurements, to analyze the directivity of a single capsule. Figure 3.4b shows a different FOA array centered for horizontal measurements of the entire SMA, the axis of rotation is directly below the center of the tetrahedron[46]. Figure 3.4c shows a third FOA microphone, rotated 90° in preparation for vertical measurements of the Z harmonic.

As noted, the purpose of these filters is to fix any irregularities in the frequency response of each capsule. Ideally, A-format calibration filters should be measured using Pose 1, repeating the process for each sensor in the array. Unfortunately, this substantially increases the amount of time required for calibration, since each capsule requires a whole new set of measurements.

In the A-format Octathingy calibration, horizontal measurements of the entire array are used to calibrate the A-format signals (e.g. using Pose 2). Lopez-Lezcano, unfortunately, does not provide much information about the generation of these filters in the associated publications. From inspecting the open source code, we discover that the inverse filter is generated using three measurements from a direction proximal to the angle of the capsule in question - in the azimuthal sense. The measurements are also weighted according to a cardioid response and averaged. The resulting response is then inverted directly without regularization and converted to minimum-phase using the technique described by JOS:[47].

> "Any spectrum can be converted to minimum-phase form (without affecting the spectral magnitude) by computing its cepstrum and replacing any anticausal components with corresponding causal components. In other words, the anticausal part of the cepstrum, if any, is "flipped" about time zero so that it adds to the causal part. [Con]"

[46]This pose was used for X and Y harmonic sampling.
[47]Using the technique in [Con], makes the filter causal.

Finding suitable methods for A-format calibration in SMAs is an open area of research. The method chosen by Lopez-Lezcano works reasonably well using only horizontal measurements, however, the 8-channel system is relatively small[48] compared to other arrays. Additionally, the elevation difference, relative to the horizontal plane, is the same for all capsules; for more complex arrays, where multiple elevation angles are featured, it is not clear that horizontal-only measurements would be the best choice. In the Octathingy design, the calibration of A-format is designed to optimize the reproduction of horizontal signals; this means the optimal recording position of the microphone, would be perpendicular to any musicians.

Encode Calibrated A-format IRs

With the A-format equalization filters derived, the original IRs[49] is filtered and subsequently encoded using a matrix of coefficients derived from the SH representation of the capsule positions. Once the matrix is initialized, it is inverted to produce the encoding matrix [MDMF+18].

The inversion matrix for non-uniform distributions can either be calculated using the Moore-Penrose technique or SVD[50]. Equation 3.6 shows the encoding matrix calculation given uniform and non-uniform distributions. The non-uniform solution in Equation 3.6 describes the Moore-Penrose approach.

$$\mathbf{Y}_e = \begin{cases} \frac{1}{Q}\mathbf{Y} & \text{uniform} \\ \mathbf{Y}^\dagger = \left(\mathbf{Y}^T\mathbf{Y}\right)^{-1}\mathbf{Y}^T & \text{non-uniform} \end{cases} \tag{3.6}$$

If using the SVD method (e.g. $Y = U\Sigma V^T$), the pseudo-inverse will be calculated by:

$$Y^\dagger = V\Sigma^\dagger U^* \tag{3.7}$$

[48] In terms of the number of sensors.
[49] This might include multiple elevation angles.
[50] Singular Value Decomposition.

The star superscript indicates conjugate transpose. If a matrix has all real components then the conjugate transpose is just the transpose. If the matrix has complex entries, you take the conjugate and transpose the matrix [Sin]. $\Sigma^{\dagger}$ is formed by calculating the reciprocal of each eigenvalue[51] in Σ. Both methods should yield the same matrix.

The encoding matrix is only valid below the aliasing frequency where the wavelength of the impinging sound is large enough to be accurately sampled in the SH domain by the mic array[52]. For this reason, additional equalization must be applied, especially above the critical spatial aliasing frequency where the responses begin to deteriorate.

In SpHEAR, the author uses a slightly different approach; Lopez-Lezcano derives the encoding matrix as a function of the measured IRs - much like in [MDB06] or [JEP13]. The matrix of encoding coefficients is not just the inverse of the projection matrix, rather, it is the optimal matrix solving the system of linear equations involving the specific IR powers[53] and the ideal SH representations. Since this array is irregular, the solution is also inverted. The final matrix is frequency-independent; the powers in the critical bandwidth is integrated and averaged to find these values. Unfortunately, the reference given for this process is unpublished. We discovered this by inspection of the code.

In Moreau et al. [MDB06], the authors compare the frequency-independent matrix approach[54] versus the frequency-dependent inversion approach - which yields a matrix of encoding filters. One can use simulated responses, or idealized responses, in order to find the desired filter matrix. Lopez-Lezcano uses idealized responses and real measurements[55] to find a matrix of scalar coefficients, and subsequently derives

[51] One over the original value. The original σ entries are called singular values.

[52] One should also consider the source's relationship to the receiver. Various authors opt for 2m and ESS times of 10 seconds. 20Hz is roughly 17 meters.

[53] At a specified frequency range, where the capsules are considered co-located.

[54] Using a matrix of scalars plus radial filters.

[55] The RMS power is calculated over a specific frequency range determined by the size of the

custom equalization filters based on measurements, rather than applying radial filter theory.

Find Equalization Filters for B-format Signals

As aforementioned, no radial filters or a matrix of filters are derived in this project. Instead, the author analyzes a set of initial B-format signals, resulting from the A-format equalization, plus the transformation from the encoding matrix, in order to create custom B-format equalization filters.

According to the author:

> "Our software defines the shape of the B-format equalization filters by measuring the power in logarithmically spaced bands for measurements near the peak of each recovered lobe. The inverse of this power profile is used to create the FIR filters.[LL19]"

The measurements corresponding to the points where the SH features a maximal response are strictly used to define this equalization profile. The B-format equalization filters are then generated much the same way as the A-format ones via the MPS method inverting these maximal responses.

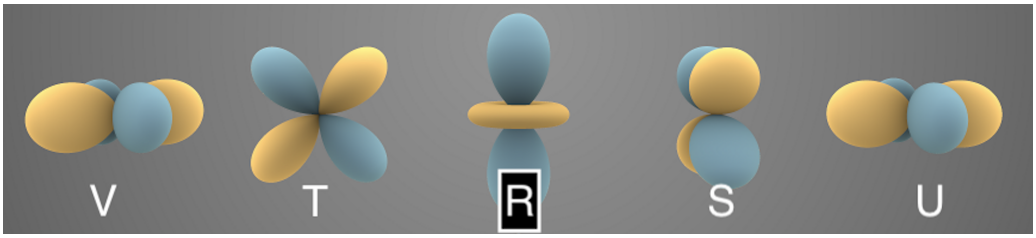


Figure 3.5: SH of Ambisonic Order 2

Unfortunately, the horizontal-only measurements do not provide any information suitable for B-format equalization at certain harmonics. Consider Figure 3.5, here we see clearly that certain harmonics (e.g. T and S) contain nulls at elevation $\theta = 0$. This is for example why Pose 3 was implemented in [Zal19], which allowed us to determine if our own array could sample the Z harmonic[56].

array.

[56]Since Lopez-Lezcano has access to many elevation angles, this is not a problem. However, if

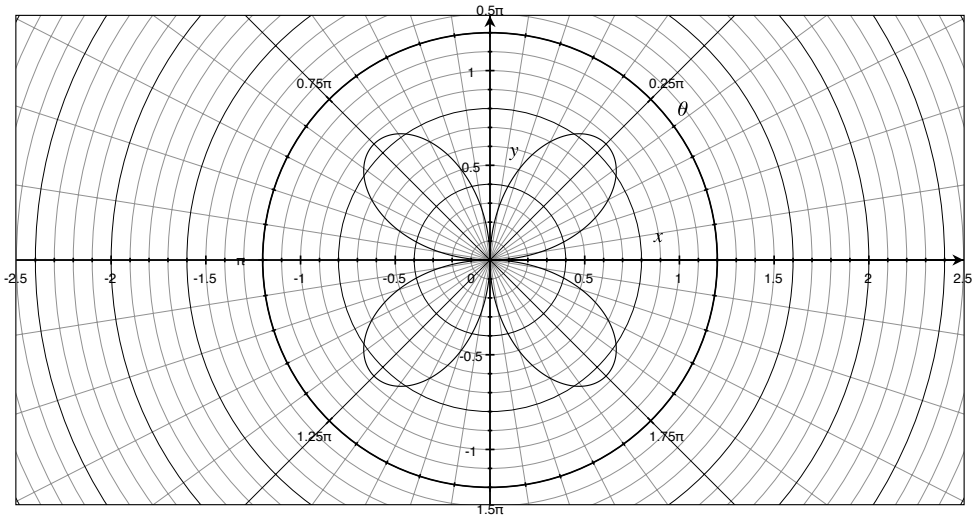


Figure 3.6: V Harmonic - Horizontal Cross Section - Polar Plot

While harmonics Z, T, and R[57] can suitably be equalized with a set of measurements using Pose 3, due to the orientation of harmonic S, it cannot be equalized using either of these measurements[58]. Figure 3.6 shows the horizontal cross-section of harmonic V, using a 2D polar plot, illustrating the result of sampling this harmonic using horizontal only measurements (e.g. Pose 2).

Assuming symmetry between harmonics T and S, we could apply the same equalization to both harmonics. This theory can be further expanded to determine the fewest number of filters required for calibration of all harmonics, however, in the real world, these arrays are never perfectly symmetrical, so the compromise comes at a cost.

Consider now Figure 3.7, which shows SHs of ambisonic order 3 (note, in particular, harmonic O). At higher orders, certain harmonics have nulls using both Poses 2 and 3, which means a different angle needs to be accessed for the calibration filters to

we don't have access to a robotic arm, we will have to figure out some other method.

[57]2, 5 and 6 in ACN ordering.

[58]It is rotated 90°compared to harmonic T.

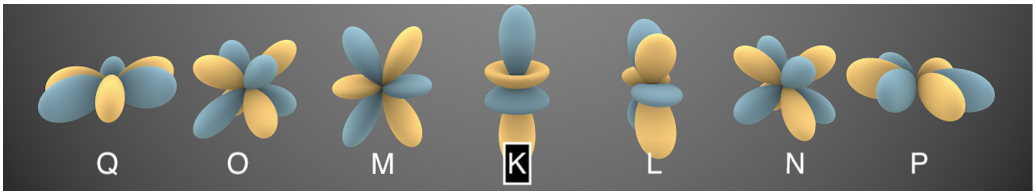


Figure 3.7: SH of Ambisonic Order 3

be derived. For 3OA, we would need at least Pose 2 and 3, and a third measurement where either the speaker or microphone had a 45° tilt. This demonstrates that the calibration processing time, for this method, has a direct relationship with the order of the array.

Lopez-Lezcano also notes that second-order components (e.g. V, T, S, U)[59] drop in volume at 6dB per octave with decreasing frequency. This means that the B-format correction filters provide intense amplification at those frequencies, which results in significant self-noise. As a result, it is necessary to add a high-pass filter to the design. Alternative equalization techniques such as those discussed by Middlicott et al. [MW19] might be worthwhile exploring as an alternate means of B-format correction.

Radial filters also tend to boost the low end of these arrays, which can further exacerbate the noise problem. Care should be taken if one wishes to use both radial filters and equalization filters. Due to the aggressive nature of radial filters, other authors have also suggested high-passing harmonics in the encoding process (i.e. [Ven14]). The high-pass filters should be commensurate with the ambisonic order, which affects the magnitude response of the radial filters (which are order dependent).

Export Filters and Encoding Matrix

In order for musicians and recording engineers to use this system, an easy-to-use interface is required. Having to import recordings into Octave GNU for processing every time would be cumbersome. As a result, the author suggests exporting all the

[59]We don't actually sample R, because there aren't enough sensors.

equalization filters and the encoding matrix for use in a plug-in. In the SpHEAR project, the Faust library is used to create an Octathingy encoder that can run on a traditional DAW.

Lopez-Lezcano notes the inclusion of an expander in the final encoder, to compensate for the SNR of 2nd order SHs. The expander works inversely to a dynamic compressor - instead of making loud signals quieter, and weak signals louder, it expands the dynamic range by doing the opposite. The expander is only applied to the 2nd-order harmonics, which distorts the spatial image - as conceded by the author. Other dynamic compression systems, such as those discussed in Chapter **??**, apply the same attenuation to all harmonics in order to preserve spatial acuity. This time-variant system is implemented in the final real-time encoder[60].

Summary As we can see, while SpHEAR offers a great deal of insight into the design and calibration of SMAs for FOA and HOA recording, the calibration process, in particular, makes these systems difficult to implement. The use of high-quality ECMs also makes the design expensive, especially if considering 3OA or 4OA. The analog nature of these capsules implies the need for multi-channel audio interfaces, which further increase the cost, especially as we increase the ambisonic order.

Despite these limitations, SpHEAR is undoubtedly the most comprehensive, accessible, and high-quality design we have found. Every piece of software used for the design is FOSS. This makes it possible for anyone with enough patience and perseverance to develop one of these microphones at home. The models are designed for low-cost 3D printers and the electronics are distributed publicly[61].

One of the main problems with the 2OA mic in particular, is that the missing harmonic R might make it incompatible with other VSTs and effects. The calibration process is not only laborious, unfortunately, it is also necessary, since these capsules

[60]The code for this process was unfortunately not located by the present author, so not much more can be said about it.

[61]The authors even created additional models for a windscreen and microphone holder.

have a ±3dB manufacturing tolerance. In contrast, the MEMS capsule we will talk about in the next section, have a ± 1dB tolerance, arguably making individual microphone calibration less critical.

3.3 FOA Microphone

Ambisonics Z Array Over the last few years, the present author has been investigating the viability of MEMS sensors for the development of a free and open-source ambisonic array. In our latest publication [Zal19], we exploited the small size of the MEMS to increase the coincidence of capsules, thus increasing the spatial aliasing frequency to good effect.

The polar plots resulting from acoustic measurements of the constructed device, in this case, a tetrahedral FOA MEMS array, showed improvement in the polar response of dipole elements as a result of increased coincidence - as predicted by the theory. No phase compensation was needed in order to generate the velocity elements in the system. An error in the normalization process led to distorted W responses (described in the paper), however, newer analysis of these systems using corrected software shows the arrays adequately sample the omnidirectional pressure signal. A subjective evaluation of the system was mentioned in the associated publication, however, this experiment has not yet been conducted. This experiment is the subject of future work.

MEMS MEMS sensors offer many benefits compared to ECMs:

1. They are very cheap.
2. The manufacturing tolerance is very small, so the frequency responses are already fairly well-matched.

3. They are small, which means they can be placed closer together to increase the aliasing frequency.

4. Digital MEMS can be directly recorded without the need for an external ADC.

Unfortunately, they also have many compromises:

1. Their SNR is worse than ECMs (64dB SNR v. 78dB)[62].

2. As a result of the poor SNR the dynamic range is limited.

3. They need to be surface mounted (e.g. they are soldered using a reflow oven or something similar).

4. They have a Helmholtz resonance above 10kHz.

5. They are omnidirectional.

6. They are very fragile, which means the sound port needs to be protected. Acoustically transparent foam might solve this.

The commercial Zylia microphone uses MEMS capsules internally. This is promising as it indicates that MEMS capsules can be used for HOA in professional recording conditions.

Automatic Rotating Microphone Mount (ARMM) In our experiments, we have been using the ScanIR [VGR19] MATLAB toolbox to generate and capture our impulse responses. Our automatic rotating system is based on a cheap stepper motor with a custom attachment, which is driven by an Arduino. This system only rotates along one axis. As a result, for a vertical sampling of our array, we manually adjust the pose of the microphone. This is trivial for FOA but might be less accurate and cumbersome for HOA.

[62]For the ICS43434 or ICS-41350, two capsules we have chosen, and the Primo Capsules.

One of the recent additions to this project is a 3D printable model which allows the NEMA23 stepper motor to be used as a microphone measuring system. The ARMM is very easy to build as it consists of simple electronics such as an Arduino and a stepper motor shield. The current calibration system is running in MATLAB unfortunately but creating an alternative IR acquisition method in Pd is possible using Arduino packages for Puredata. If Pd Arduino packages prove difficult Arduino also has OSC support so it should be possible to send data back and forth between the Arduino IDE and Pd via [netreceive]. Another alternative we are considering is using Octave GNU to port the existing MATLAB code. Our version of ScanIR can be found in the repository associated with this project on GitHub. The newest official version of ScanIR has not been tested yet.

Open Questions The designs proposed by Lopez-Lezcano provide a good starting point for building and calibrating HOA microphones of various sizes, unfortunately, the calibration software is far too complicated for most musicians to use - even with the example files provided. Beyond using different components (i.e. MEMS capsules and a MicroController), we would also like to investigate how the tolerance of MEMS contributes to the need for individual array calibration.

In other words: *is individualized microphone calibration noticeable by listeners given the improved tolerance of the capsules?* If it's possible to demonstrate that using MEMS results in imperceptible microphone calibration, we could distribute a single software solution that works for all people wishing to build this design. This would mean that the musician would only have to worry about the mechanical aspects of the design.

An additional open question regards the SNR of the system, which as we have noted is inferior to those with ECM capsules. An easy experiment to conduct in the short run would be to record various stimuli using our existing FOA designs, and professional FOA microphones available to us, and compare their sound reproduction

with music samples featuring wide dynamic ranges. In addition, we should evaluate the amount of noise present in higher-order SHs, to validate the proposition of MEMS for HOA.

Since the directionality of MEMS capsules is omnidirectional, we must also investigate how this affects the array responses. Finally, we would like to look into radial filters to improve the performance of these arrays, which one can combine with equalization methods.

3.4 HOA Microphone

Over the last months, we have been designing and evaluating HOA designs of various types using custom and commercial PCBs, various types of manufacturing processes, and several types of encoding methods. The three types of manufacturing processes proposed include Fused Deposit Modeling (FDM), Selective Laser Sintering (SLS), and Laser Cutting. Various protocols are compatible with the MCH Streamer Board we have selected for this project including I2S, PDM, and TDM.

Early in the design process, we opted for a design that could elegantly complement a video recording via a proprietary camera system. We selected a 360 camera by GoPro and began designing our models around the assumption that this device would be paired with our array. Much of my own artistic practice includes visual elements - I wanted to devise a system that could capture ambisonic audio but combine seamlessly with video elements. The design which I will focus on in this text is the 16-channel PDM array with an Atmos-style configuration. We were inspired by the home-theatre set-ups recently presented by Dolby which feature overhead speakers pointing down towards the listening area. This design also allows us to fit a large number of capsules in a relatively small space.

Inspired by the modular approach of the Octothingy [LL19] and the MSMA array [GPL18] we created an octagonal prism design where each of the side faces is

an independent part. The benefit of this is that if there are errors in the printing process one does not have to reprint the entire array. The top parts are also modular and, like the rest of the design, can be printed using an FDM printer which is more affordable than the SLS counterparts (see Figure 3.8).In order to connect our 16 PCBs with the mounted MEMS mics to our DSP board we use a modest perforated board and carefully solder our different lines to the appropriate pins. The major benefit of using Adafruit components is that they are tested by the manufacturer, so we can be sure they work properly. We have had some success with surface mounting parts at home, but this process is not only dangerous but also time-consuming. The Adafruit components are relatively cheap, around 5 dollars per capsule, and the documentation greatly facilitated the design process.

After all the parts are connected correctly, using the specification sheet for the MCH board and the Adafruit documentation, we connect the board to a computer and load the appropriate firmware for the PDM protocol. The software and support make this device very easy to operate and at 100 dollars it is a suitable cross-compatible solution for developing MEMS-based HOA arrays. After ensuring that all capsules are working we move into a quiet space to measure the array. Figure 3.9 shows the array being measured in the diagonal configuration for harmonics not present in the vertical or horizontal plane.

Our initial round of measurements where done in CPMC365 but we discovered that the measurements had some unwanted artifacts resulting from hard objects in the room reflecting sound. We decided ultimately to move to a larger space with a low-noise floor, specifically, the spat lab at Calit2. We recorded our impulse responses using a medium-sized Genelec speaker placing the array in the middle of the room to avoid reflections as much as possible. The speaker was placed 2m away from the source and our ARMM was employed once more to capture our measurements. The 16-channel array was measured at three angles (or poses) allowing us to sample all 16 harmonics present at 3OA.

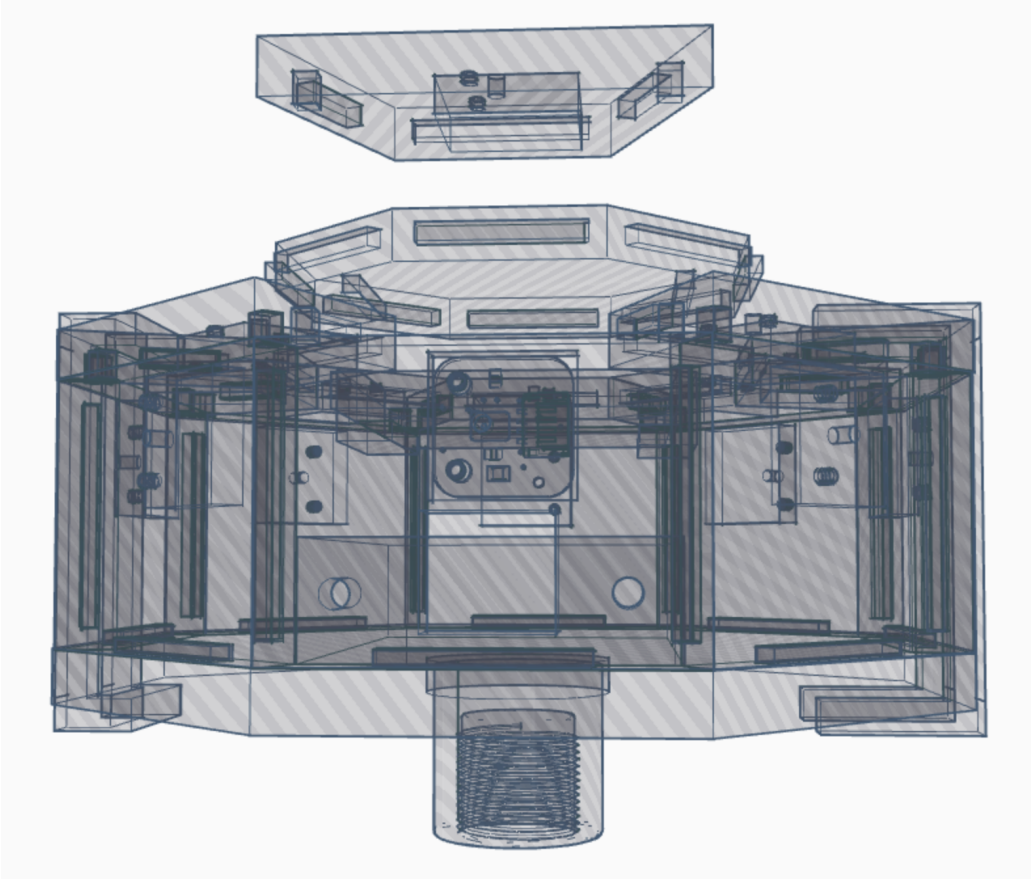


Figure 3.8: 16 Channel HOA Mic (PDM) 3D

With the measurements gathered we then needed to create an encoding matrix, or filter matrix, and, optionally, calibration filters for the design. Our current solution is composed of A-format (AF) calibration filters, a static encoding matrix, and B-format (BF) calibration filters. To derive the AF calibration filters we opted for the Diffuse-Field Response (DFR) method suggested by Middlicott [MW19]. We use all three sets of measurements which include a horizontal, vertical, and diagonal sampling of the sound field. These measurements are then averaged and a single DFR is used as the target in the filter inversion algorithm. We can see the effects of this filter by convolving the inverse "global" filter with the DFR of each capsule.

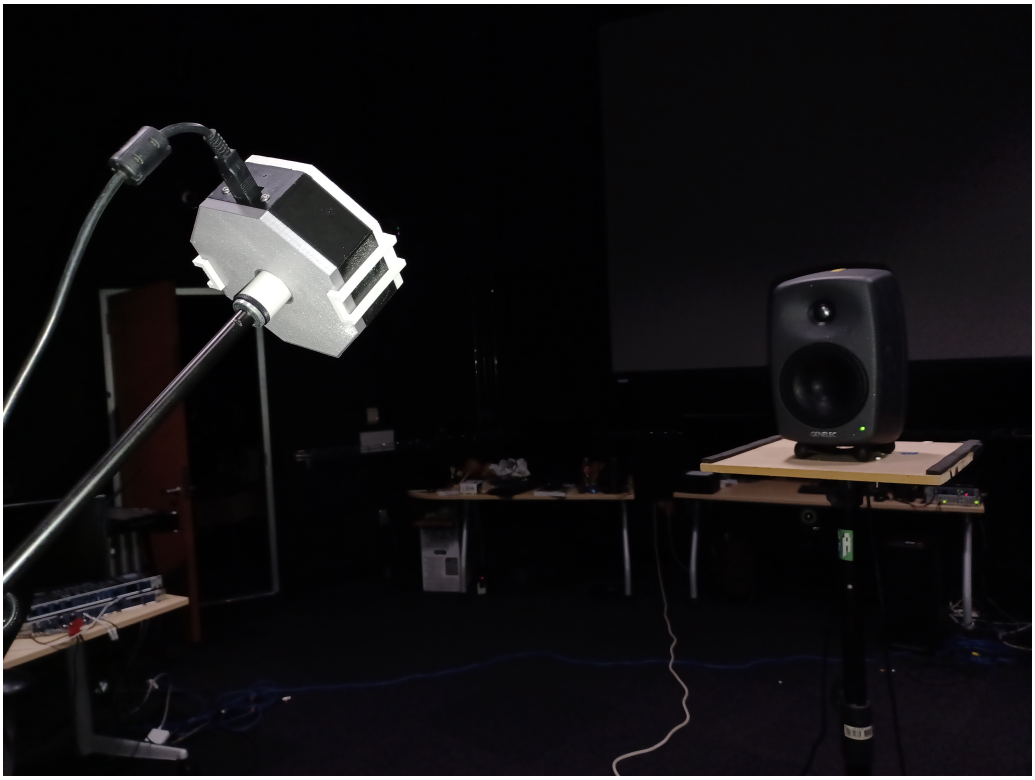


Figure 3.9: The microphone Measured for 3OA SHs (45 degrees)

The rationale behind this approach is that each array will be slightly different based on the response of the capsules, thus "local" DFRs will not be applicable to alternate arrays; this generalized solution will not be as precise, but it allows the hobbyist building this array to calibrate their system using a similar response to that of their own. After the AF filters are employed the raw signals need to be encoded into ambisonics. There are several ways to do this but we opted for the simplest solution for now, which entails describing the positions of the capsules in the spherical domain, similar to the idea of synthesizing a sound field. For example, in an FOA system, we would use the azimuth and elevation of each capsule to derive a 4-by-4 matrix with four coefficients for each harmonic. The elements of column 1, corresponding to W, are summed together - this special case is simply the sum of all

four signals. The same approach is taken for the other columns which are evaluated using the corresponding spherical harmonic functions.

Our current system relies on two open-source plug-ins that apply filtering and matrix multiplication to our signals. Users building a copy of this system can find the filters and encoding matrices in the project repository[63] and load the associated files. We have also made Reaper templates for different versions of the array which have these plug-ins pre-loaded and all the routing done for the user. Since the plug-ins are open-source, copying the code to a proprietary VST should be mostly trivial. The sparta_multiconv plug-in by Aalto is used for AF and BF calibration, while the MatrixMultiplier VST by IEM is used for encoding. The corresponding WAV and JSON files needed for these two effects are in the repository for users to download.

Once the array's raw signals have been equalized and the array has been encoded, we apply a final EQ step which attempts to make the SH flat at the incident angle where the energy is maximized. To do this we use a peak finding algorithm at a particular bin to determine the direction from which energy is maximal (for example with the X harmonic, we can assume azimuth 0 will have the maximum energy). Using the frequency response from that direction we create an inverse filter using a Dirac delta function as the target response. The goal of this last phase of equalization is to make the response flat at these SH peaks.

Using these filters and encoding matrix a subjective evaluation of the system was undertaken using the HOAST [DMKH+20] platform, Google Forms, and Prolific[64]. Users were instructed to listen to several audio-visual media which were identical visually but differed in the audio component. We labeled these different versions AB, AC, and B, where A corresponds to our ambisonic array, B corresponds to the FOA sound field captured by the camera, and C corresponds to the stereo signal captured by the camera, encoded into ambisonics. The following attributes were

[63] https://github.com/gzalles/ambisonics-z-array/wiki
[64] https://www.prolific.co/

then rated on a 5-point scale (AB means that A and B are mixed together).

1. **Noise**: how much noise is there in the sample? A low amount of noise should represent a higher score. If the noise level is high, the score should be low.

2. **Dynamics**: corresponds to the range between the quietest sounds and loudest sounds. If the microphone feels like it is able to capture very faint sounds, as well as loud ones, this score should be high.

3. **Realism**: how realistic does the microphone recording sound? If it sounds very real the score should be high.

4. **Spectrum**: how balanced does the recording sound in terms of low frequencies versus high frequencies? Can you hear the bass frequencies as well as the treble? If so the score should be high.

5. **Quality**: the overall quality of the recording, the better the quality the better the score.

Twenty online participants took part in the experiment. Surface statistics revealed that the microphone performed slightly better when combined with the FOA sound field or the stereo encoded into ambisonics than just the FOA audio. This suggests that the array can improve the audio quality of the resulting video relative to the baseline audio recording. Deeper statistics involving ANOVA reveal that the three treatments are not statistically different. This demonstrates that at the very least the PDM HOA array can replicate the sound quality of the GoPro camera, which means it is tolerable as a substitute for camera systems that don't include built-in microphones. We believe part of the reason the treatments were statistically identical is the small 5-point scale we used.

The results of this short experiment ought to be taken with a grain of salt, as we were unable to verify the extent to which participants engaged with the media

(e.g. if they listened attentively) or if they even wore headphones (as requested). Nonetheless, this methodology incorporating paid online participants feels promising as a means of establishing baseline results, which can then be used to modify the experiment. It is extremely easy to set up and administer tests, and if enough participants are employed, real trends can be obtained.

3.4.1 Objective Measurements

Below are some of the resultant polar plot measurements resulting from our analysis. Figure 3.10 shows the on-axis response of the eight I2S MEMS in our first design. As you can see there is quite a bit of variability between our measurements - this is most likely because the center of gravity of the stepper is not perfectly aligned with the microphone, and because the room is not perfectly anechoic. One of our future goals is to average multiple sets of measurements to validate our initial data.

After the on-axis measurements have been plotted we calculate the DFR (Diffuse-Field Response) of the system by averaging, per capsule, the IRs - all directions collapse into a single measurement[65]. The DFRs, per capsule, called the "local" A-format (AF) DFRs, can be used to create another inverse filter that attempts to calibrate the response of the array to a target response set by the user. In this case, we decided to use a delta function as the target IR as it exhibits a perfectly transparent profile. Since we can't expect the reader's capsules to perfectly match our capsules we decided instead to generate a single AF calibration filter based on the "global" DFR, which is the average of all "local" DFRs. The result is a single AF calibration filter which is applied to each of our 8 channels prior to encoding the AF data into ambisonics[66]. Initially, we suggested Kronlachner's convolution VST for applying these filters but found that the Aalto suite is better equipped to handle

[65] A real DFR would include height information as well, here we only focus on horizontal measurements.

[66] Our processing order goes: AF calibration ¿ encoding ¿ BF calibration.

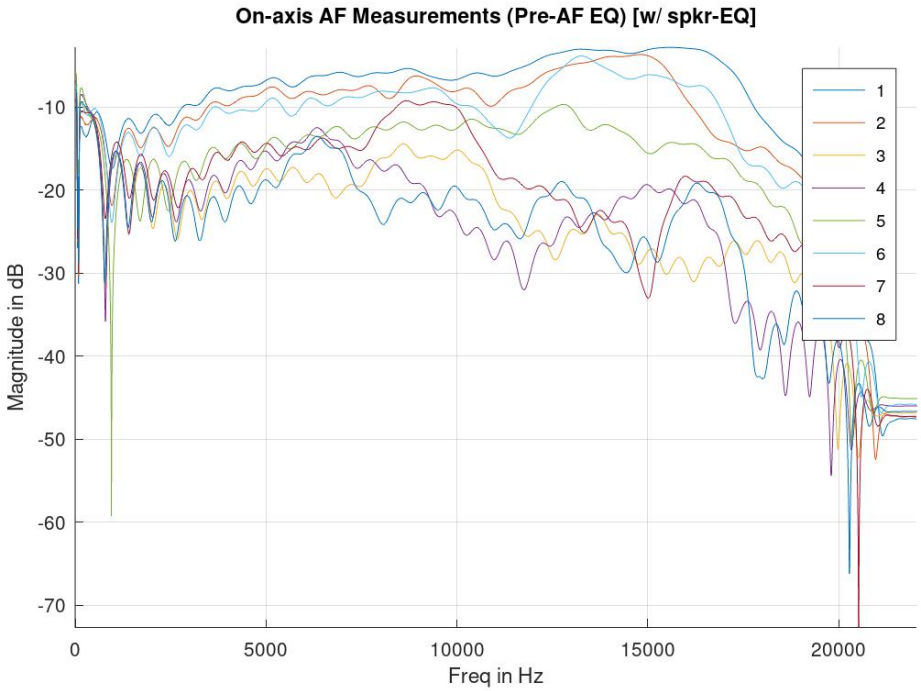


Figure 3.10: On-axis measurements of our data from design 1.

this task. The filters themselves along with Reaper templates can be found in the repo associated with this project.

Figure 3.11 shows the DFR of the AF data from design 1. Figure 3.12 shows the DFR after EQing with the inverse filters (this uses 8 "local" DFR inverse filters). As you can see the inverse filters have compensated the DFR of all Q sensors such that their response is flat under these conditions.; this is the idealized case - with personalized filters.

After the AF IRs have been equalized we must encode them into ambisonics to determine the response of the SHs. The encoding matrix can trivially be calculated using the spherical coordinates of the sensors and the standard ambisonic equations. The responses undergo encoding via matrix multiplication between the encoding matrix and the IRs. For a mathematical description of this please refer to [MDMF$^+$18].

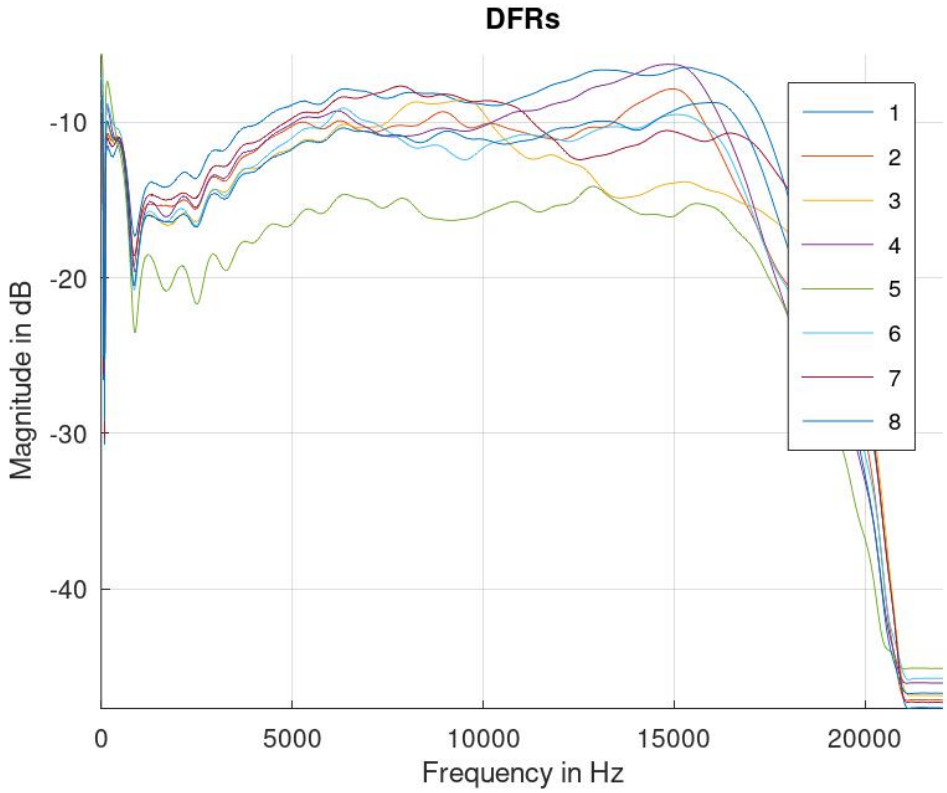


Figure 3.11: DFRs of eight capsules in design 1 ("local DFRs")

Figure 3.13 shows the polar response of the seven SHs in 2D available for this array (at 2kHz).

The final step in our chain is the equalization of the B-format components for which we implemented two solutions. The first solution is based on finding the incident angle of the most sensitive direction of the polar response and calibrating the array such that this response is flat. For example, in ACN the second harmonic (Y) corresponds to the L/R direction, so the peak would be at 90 degrees which corresponds to step 50 in our data. Using this specific measurement, we then create an inverse filter using the delta response as our target once more. Figure 3.14 shows the equalized harmonic on the left, and the raw SH on the right using this method

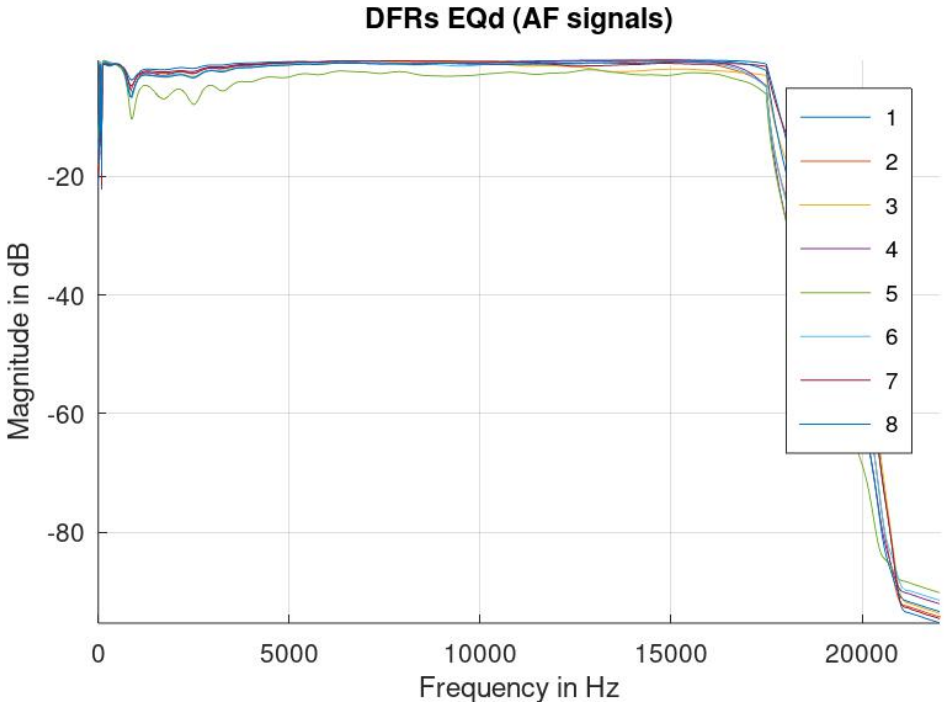


Figure 3.12: DFRs after EQing with inverse filters

at 2, 4, and 8kHz.

The other approach to equalizing the SHs is to find the DFR of the SHs and create an inverse filter which makes it so the DFR of the SH will be flat. For the selected frequencies, these two methods proved to be fairly similar with visual inspection revealing that the "peak" BF method performs better than the DFR approach. Figure 3.15 shows a HOA SH equalized using the DFR method at 2, 3, 4, and 8kHz. In the future we intend to use analysis toolkits to evaluate the differences objectively.

Once we have calculated all our filters and the encoding matrix we must import these into a suitable DAW (Digital Audio Workstation) to encode the raw signals (e.g., AF) into ambisonics. The best solution we have found is Reaper in conjunction with the toolkits: ambix, SPARTA, and IEM_plugin_suite. There are two convolu-

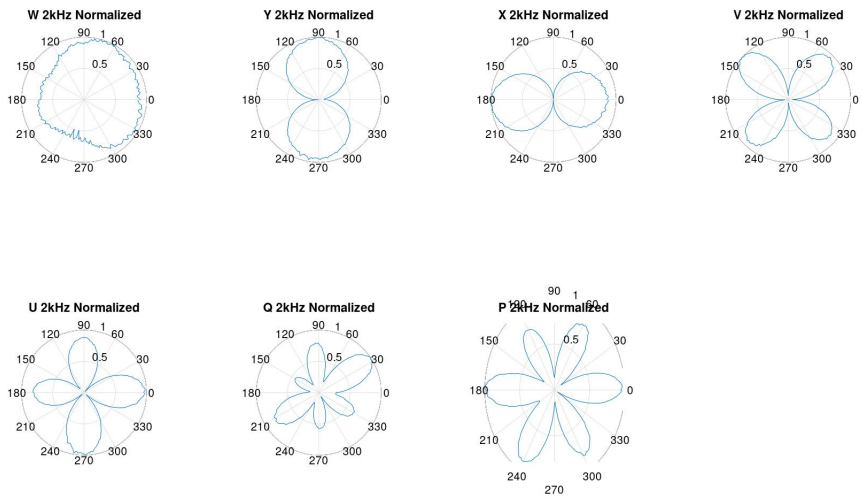


Figure 3.13: SHs of all orders in 2D for 3OA (Design 1)

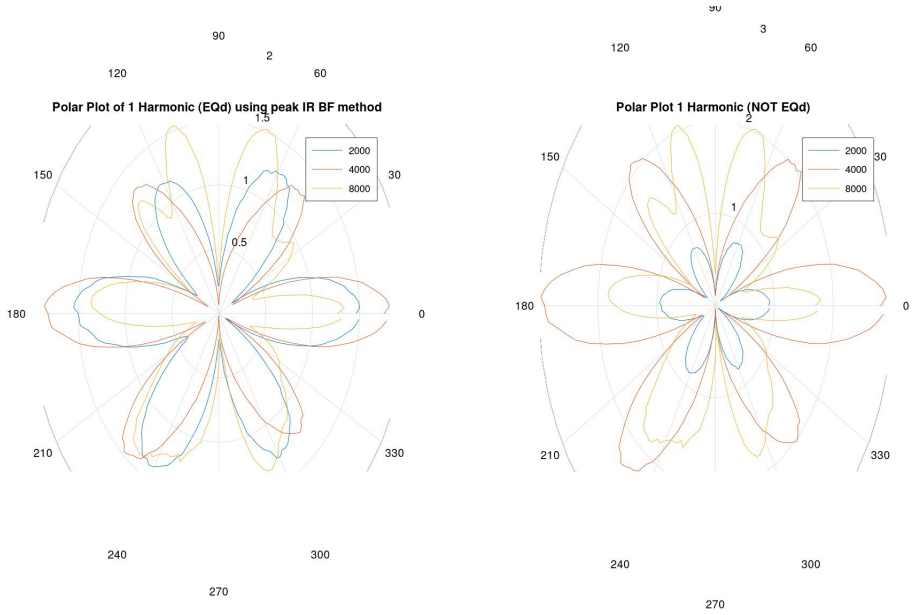


Figure 3.14: Design 1 SH - BF equalized using "peak" method (right side no EQ)

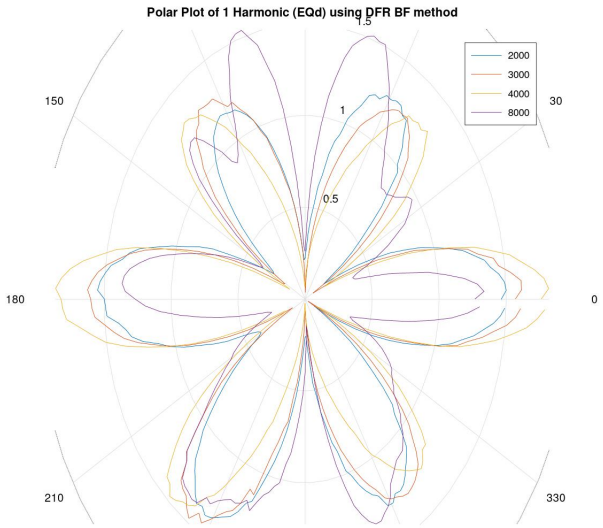


Figure 3.15: Design 1 SH - BF equalized using "DFR" method

tion FX in SPARTA for filtering and thus calibrating our HOA array. The first, sparta_multiconv, is used to apply the AF calibration filter as well as BF calibration filters.

For design 1, the plug-ins are configured with 8 ins and 8 outs, each channel being convolved with the corresponding filter. In order to encode the AF into BF we use the MatrixMultiplier by IEM which allows us to load JSON files. There is an audio recording of this system in the repo's wiki - the audio is binaural. Since the BF filters are all targeting a delta response with full energy at all frequencies it is important to further study how this contributes to the normalization expected (e.g., SN3D); it might be necessary to examine the energy of the signal after filtering, to ensure the balance is correct between orders.

3.5 Design 2

Upon completion of the first prototype, we decided it was important to create a concept with height information encoded in the signal. The second prototype features 8 additional capsules on the top side of the octagonal prism, which we measure using different poses, to sample the vertical components. In addition to horizontal measurements, two additional planes were sampled: the vertical plane and the 45-degree diagonal plane, intersecting the XY plane (we call each of these configurations "poses"). Unfortunately, we do not have access to an automated arm, so these pose changes are done manually.

Much like before the first step is to equalize the AF, raw, audio. Figure 3.16 shows the frequency response of all the capsules from azimuth 0, using pose 1 (e.g., horizontal measurements only).

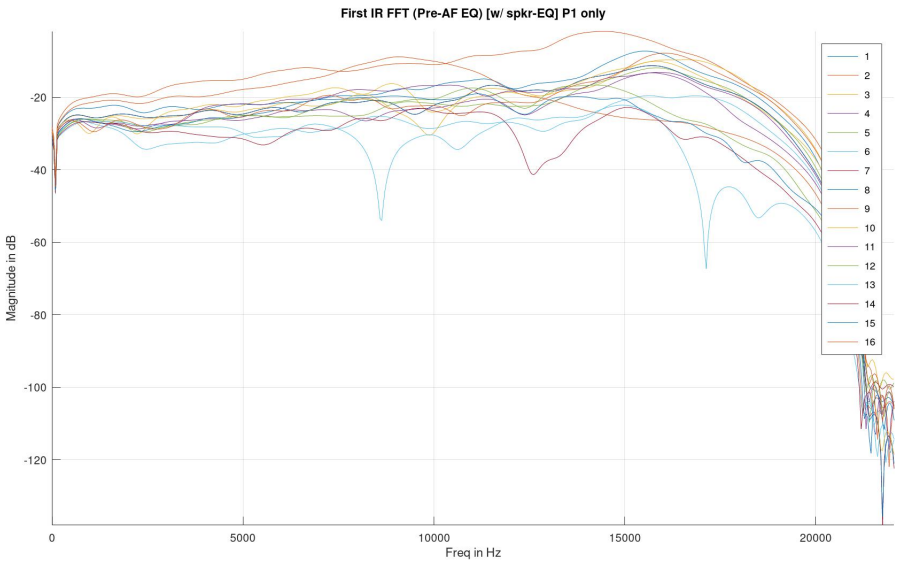


Figure 3.16: Design 2 - AF First IR - Pose 1

We then calculate the DFR, but this time we use three poses and average all these.

Each capsule can then have a custom AF calibration filter designed to compensate for the response and help match all sensors. Like before we use the delta function as a target; the filters are turned to minimum phase using the technique developed by Julius O Smith [**?**][67]. Following this process, we use the same equations as before to calculate our encoding matrix for this new array and then encode all our IRs. Figure 3.17 shows the frequency response of all the capsules, again only step 0, using pose 1 after AF calibration.

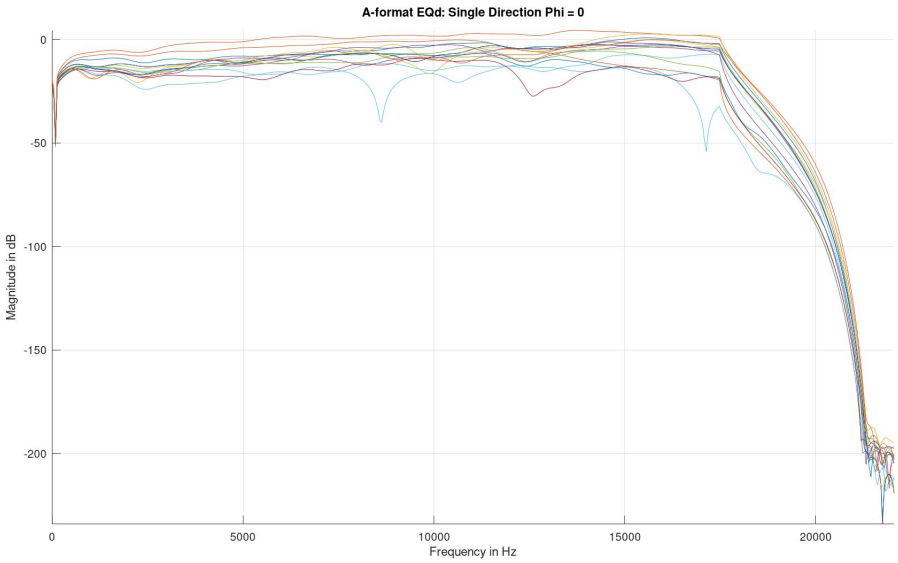


Figure 3.17: Design 2 - AF First IR [EQ] - Pose 1

We have to trim some data; the information in certain poses does not yield anything valuable. For example, pose 1 (horizontal) does not have any Z information. For all SHs with degree = 0, we can use pose 2 (vertical), however, some harmonics have no information in either of these perspectives. Using a combination of all three sets of measurements, however, we can create a data structure will all harmonics

[67] Another option would have been using three DFRs, one per pose, to calculate the AF calibration filters. For example, channels 9-16 use only pose 1.

sampled. Figure 3.18 shows the SH polar responses of all orders available at 3kHz - note this is a combination of multiple poses.

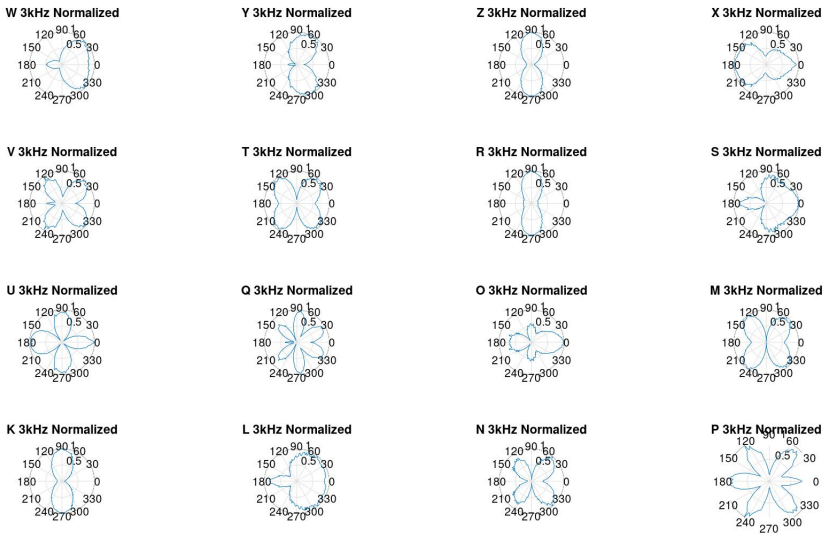


Figure 3.18: SHs of all orders in 3D for 3OA (Design 2)

After we encode our IRs, much like before, we calculate a set of Finite Impulse Response (FIR) filters to compensate for the inaccuracies in the SH response. Subsequently, we find the peaks of our SHs and use those measurements to generate inverse filters. In order to find the peaks we can simply calculate the "perfect" SH computationally and find the index corresponding to the maximum[68]. We can also derive BF calibration filters from the DFR of the harmonics. Figure 3.19 shows the DFR BF calibration effect on harmonic 12 (ACN). We can see some objective improvement in using this approach, and visual inspection suggests it's an improvement over the "peak" method.

[68]For pose three we use the same theta angle as the orientation of the array, 45 degrees.

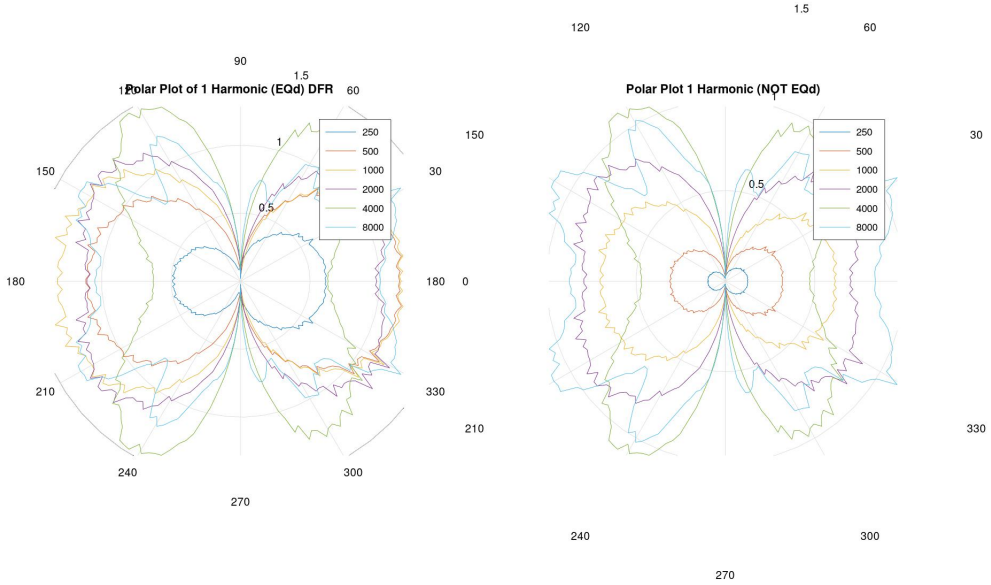


Figure 3.19: Design 2 - BF equalized using "DFR" method

3.5.1 Findings

This section summarizes some of the main findings we have discovered during the last year of developing this project:

1. More FFT bins in our FIR filters mean better accuracy but require more multiplications. This might be important for small computers like RPIs or other MCUs.

2. When encoding Spherical Microphone Arrays (SMAs) into HOA we may choose either a static matrix with one coefficient per element in the matrix or a matrix of filters. The computational load-to-quality ratio will be the subject of a future study.

3. In the past we suggested using a Helmholtz resonance filter before AF calibration, this is unnecessary since the target function in AF calibration is a delta

function, which eliminates the Helmholtz frequency by its design.

4. Platonic solids might be ideal for isotropic sound recording but harder to integrate with camera systems.

5. The smaller the array the better the spatial performance at high frequencies but too small and low frequencies are not sampled accurately. The smaller arrays also require high-quality printers making them more inaccessible.

6. Adafruit components make building these arrays significantly easier and less prone to errors. Given the low cost of the components and the complexity of surface mounting at home, this might be a better overall solution. Stencils make surface mounting at home significantly less error-prone.

7. Calibration is not necessary for each array, but in our case, it showed us defects in the construction which allowed us to diagnose and repair our system.

8. It is impossible to create a set of AF calibration filters that perfectly EQs any array. Therefore, we opted to use the "global" Diffuse Field Response (DFR) of our systems to generate our AF filters, providing a generalized solution.

9. Some arrays might sample vertical harmonics better than others based on sensor distribution. Humans are better at discriminating sources in the horizontal plane, thus perhaps cylindrical arrays are more appealing musically.

10. We can laser-cut parts for HOA arrays with extreme speeds relative to printing.

11. We should treat our arrays for EMI however we can. Two good methods are shielded cables and copper tape.

12. Defects in the print can cause sound to enter the model, resulting in unwanted resonances.

13. We should be mindful during B-format calibration not to destroy the SN3D normalization. The overall energy gain/loss of the filters should be zero (e.g. the filters should be transparent).

14. It should also be possible to B-format equalize using the DFR of the SHs, this has not been evaluated yet.

15. On-axis AF DFRs might differ substantially from what we expect, this is likely in part due to imperfect sensor positioning in the ARMM.

16. In small on-board computing platforms, we could reduce memory load by using one B-format filter for symmetric pairs. For example, X and Y could employ the same filter.

17. When evaluating the sound quality for our array versus the GoPro microphones we noticed the binaural decoder had an enormous effect on the signal. The decoder used in HOAST made these two treatments sound very similar, while the IEM binaural decoder sounded incredibly different for each treatment. When decoding using IEM's binaural decoder our array sounded much better than the MAX360 GoPro FOA recording. It would be worthwhile also comparing these treatments over loudspeaker arrays.

18. It might be worthwhile to compare subjectively B-format calibration using measurements versus radial filters derived analytically.

19. Not all MEMS have resonances, the SPH0645 used in the I2S 8-Channel array we built does not need AF calibration.

20. Wind can become a factor outdoors and is hard to differentiate from internal noise.

3.5.2 Contributions

1. Modified the ICS-43432 boards such that we can use the TDM ICS52000 mics. I noticed the landing pattern layout was the same, so I used EasyEDA to modify the board, adding an output for the WSO clock. The microphone that uses this board is still under development.

2. I also modified another copy of the 43432 board with the select pin tied to V so we can do L/R I2S 8CH mics with the MCH MiniDSP Board. This microphone is also under development.

3. Designed several 3D CAD models that are suitable for HOA arrays and also proposed hybrid designs that incorporate laser cut parts, FDM parts, and SLS parts.

4. Evaluated the efficacy of various calibration methods objectively.

5. Ran a subjective study where I studied the viability of one of these microphones and ran deep statistics on the results.

6. Designed a 3D-printed part that can be used to recreate the ARMM system with conventional tools.

7. Created a library of functions for encoding and EQing the arrays.

8. Created a composition that incorporated the array and created a HOAST instance to share the work online.

9. Made a few different buffer boards for the ICS-43150 PDM mic and the ICS-52000 TDM mic.

3.5.3 Future Work

1. Measure the arrays with the camera mounted to see if/how this affects the response[69].

2. Compare array2SH Sparta plug-in to filter matrix, to static matrix plus AF/BF calibration using ANOVA study.

3. Make a robotic arm with multiple DoFs. There are open-source designs online.

4. Create datasets for ML sound field synthesis. We could create new special datasets which have multiple arrays for 6DoF.

5. Make open source Pd IR recording system or port ScanIR to Octave.

6. Make JUCE plug-in rather than Reaper templates.

7. Compare in an experiment AF with individually calibrated filters (Q filters), versus no filters, versus global DFR.

8. Compare Max360 camera audio only, versus hybrid, versus audio only from z-array.

3.6 Conclusion

This chapter has presented a literature review and recapitulated some recent advancements in our microphone array project. In contrast to ambisonic synthesis SMA naturally records sound fields which might be beneficial in remote environments with uneven terrains. If a musician is moving around a space there is also no need to track their movement with this system, since the recording is spatially accurate. Over the last year, we have resumed our work in this domain and produced several different

[69] This might be difficult for some orientations since the camera adds significant weight.

microphone arrays using MEMS. The designs consider the integration of a camera system paramount to the success of the device. A 16-channel PDM MEMS array was designed and evaluated both objectively and subjectively. Various encoding methods were evaluated and a simple musical oeuvre was created to demonstrate the potential of the system. The designs, code, and instructions for building the system were published online along with examples showing how to calibrate and use the device for artistic creation.

Chapter 4

Conveying Spatial Music

4.1 Introduction

This dissertation chapter discusses WebXR technologies and how these can be leveraged to disseminate spatial audio compositions online. Extended reality (XR) is an umbrella term that includes Augmented Reality (AR), Virtual Reality (VR), and Mixed Reality. The term Mixed Reality has recently faded out of favor and is now synonymous with AR. In the past AR was limited to displays with no user input, and Mixed Reality was reserved to displays which allowed users to interact with the system.

This dissertation will showcase three VR projects created using online Javascript (JS) libraries. The target environment for these pieces is mobile devices rather than standalone Head Mounted Displays (HMDs) since this increases the accessibility of these works dramatically at the time of this writing. The promise of WebXR is a system that works automatically on desktop computers, mobile devices, and HMDs without needing to recompile code for all these platforms or needing to worry about Operating Systems (OSs).

Unfortunately, even with WebXR, there are still barriers caused by manufacturers

who limit availability to sell more products and services to their consumers. Some audio formats may be for example restricted by Apple, or different browsers may not support libraries needed for WebXR to work on some hardware. Despite these barriers, we still believe that WebXR is one of the most open XR platforms allowing us to share spatial music at scale.

4.1.1 Outline

This chapter will begin by talking about the history of VR, which owes its origins to the development of the moving image. Secondly, we will have an overview of binaural audio, which allows us to share these XR works with simulated spatial audio online. Finally, we will conclude by sharing with the reader three WebXR projects created throughout our studies that use three different spatial audio JS libraries.

The first of these projects is Pigments of Imagination, which is a web version of Tim Gmeiner's newly expanded project as part of the IDEAS[1] series, in collaboration with Eito Murakami (2023). This project was developed with both students during their undergraduate studies as part of our collaborative incubator SElectOr[2] remotely while the pandemic[3] prevented us from being physically together.

The second project is called Continuum and it was developed during the fourth year of SElectOr with the support of several other undergraduate students, mostly in the ICAM program. This second project is a conceptual interactive instrumental album featuring various remixes created by the students. The user is allowed to navigate the album virtually, which features seven short compositions all featuring spatial audio. The names of all the student collaborators can be found on the SElectOr website.

[1] Initiative for Digital Exploration of Arts and Sciences
[2] ucsdselector.com
[3] COVID-19 pandemic of 2020

The last project is called Bits and is a solo XR project by the author which was started during his first year as a Ph.D. student during his studies with Natacha Diels. The project is an exploration into digital hoarding and excess featuring simulated surround sound and a virtual multi-panel display. Part of the goal of the project was to explore how inaccessible infrastructure could be simulated in virtual worlds and to showcase a piece with thousands of sounds and images (e.g., large-scale "big data" composition). The piece was inspired in part by the classical computer music work Williams Mix (1951–1953) by John Cage and Poème Électronique.

4.2 What is XR?

Before we dive into the role of WebXR in computer music, we should understand what exactly XR is. More specifically, for own work, we should understand what *Virtual Reality* (VR) is. Every XR experience is a *perceptual illusion*. A programmer, or designer, has the task of creating an environment that attempts to fool another person's senses. This illusion can be aural, visual, or involve multiple senses - including smell, touch, and taste. It should be noted, however, that this does not always mean the end goal is to make XR experiences as realistic as possible. Animations and *lo-fi*[4] graphic scenes are not only popular but sometimes preferable to highly realistic simulations.

LaValle defines VR in the introductory chapter of his free book "Virtual Reality" [LaV16] as a system containing four key elements:

1. **Targeted behavior:** The organism is having an "experience" that was designed by the creator. Examples include flying, walking, exploring, watching a movie, and socializing with other organisms.

2. **Organism:** This could be you, someone else, or even another life form such as a

[4]Low-fidelity.

fruit fly, cockroach, fish, rodent, or monkey (scientists have used VR technology on all of these!).

3. **Artificial sensory stimulation:** Through the power of engineering, one or more senses of the organism become co-opted, at least partly, and their ordinary inputs are replaced or enhanced by artificial stimulation.

4. **Awareness**: While having the experience, the organism seems unaware of the interference, thereby being "fooled" into feeling present in a virtual world. This unawareness leads to a sense of presence in an altered or alternative world. It is accepted as being natural.

In contrast to *Mixed Reality* (MR) and *Augmented Reality* (AR), VR experiences completely block out real-world sensory stimuli[5]. Simple, overlay-style, applications are sometimes labeled as being *AR experiences*. Meanwhile, MR is sometimes reserved for more sophisticated systems, which require *machine vision*. This, however, is not always the case. There are various MR systems that are marketed as being AR in nature[6].

In an MR experience, an *avatar* that is displayed in your *Field of View* (FoV) might, for example, fall off a real table, and respond to this event by holding its knee. This is not possible without a geometric representation of the space, which is obtained using some sort of light *sensor*[7]. Machine vision refers to the ability of a system to use camera data to detect characteristics of the real world. These include, but are not limited to, the positions of objects, or, a person's facial expressions.

In an AR experience, in contrast, we might see a simple display of time, temperature, coordinates, and humidity in the corner of one's glasses. This does not require any camera data to be generated, but instead uses other types of sensors. This information then gets automatically updated as the user moves from one location to

[5]Real-world visual stimuli is totally occluded.
[6]Most Apple products are marketed as AR, but behave much like MR systems.
[7]Such as a camera.

the next. In these AR systems, there is no link between the geometrical space, the objects inside it, and the system.

Schmalstieg and Hollerer [SH16] authored a comprehensive review of AR in their book: "Augmented Reality: Principles and Practice". According to the authors, AR implies the confluence of three elements:

1. Combination of real and virtual worlds.
2. Interaction in real-time.
3. Registration in 3D space.

The first and second parts of this definition are quite clear: the system must bring together elements of the real and synthetically generated world, and, the system must respond to user interaction. The last element is defined by the authors as:

> "...precise real-time alignment of corresponding virtual and real information. This mandate implies that the user of an AR display can at least exercise some sort of interactive viewpoint control, and the computer-generated augmentations in the display will remain registered to the referenced objects in the environment."

In our previous example, the data from the real world is limited to time, temperature, coordinates, and humidity. Therefore, according to Schmastieg and Hollerer, this does not qualify as AR. The MR example we provided earlier, however, would be called AR. This underlines the fact that companies and academics have different definitions of AR and MR, likely due to the infancy of the technology.

Milgram and Kishino described in 1994 a continuum that can help categorize different XR experiences [SH16]. Figure 4.1 shows this spectrum. As you can see, in their definition, MR is a spectrum within which AR resides. The term AR here, refers to systems that are more real than virtual. *Augmented virtuality*, refers to

systems that are more virtual than real. For example, we can imagine computer-generated worlds where avatars' faces are real, but the rest of the environment is digital.

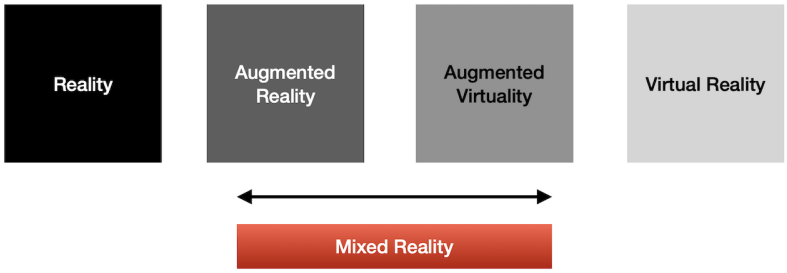


Figure 4.1: Milgram and Kishino Continuum

Much like other authors in the field, if necessary, we will simply use the term AR in this text to avoid confusion. MR will be included in AR; we make no distinction between the two[8]. As new systems emerge, we may see even more types of "reality" emerge both in the academic and commercial sectors - note: lines are not always clear, especially when senses are stimulated by systems or real-world stimuli, they were not "intended" to be stimulated by.

Consider a person seeing a VR experience with a VR display[9] but listening to sounds from speakers, which are inadvertently being blended with real-world sounds. The confluence of synthetic and real stimuli in the auditory domain points toward an AR experience. However, what if the real-world sounds are not intended to be there? Is one of their senses in AR, and another sense in VR? If the real world sounds subside, does that change how we label the experience?

[8]It appears that MR is most prominent in industry products, however, fewer academic authors use this label.

[9]Otherwise known as Head Mounted Displays (HMDs), these are "helmets" one wears which display images via screens mounted inside them. These will be covered in more detail in section 4.3.1.

4.3 History of XR

Much, much before the first ever "real" VR experience was ever developed, pioneering painters, from as early as the 15th century, were already working on the concept of depth perception and optical perspective. Today we take for granted the idea that 2D representations of 3D space contain any dimensionality, but before the establishment of cameras, this was a feature that had to be creatively articulated by artists. The idea of a *vanishing point*, the point at which receding parallel lines viewed in perspective appear to converge, shaped an entire generation of painters. Figure 4.2 shows Pietro Perugino's use of perspective in the "Entrega de las Llaves a San Pedro" fresco at the Sistine Chapel (1481–82).

Figure 4.2: Vanishing Point Painting [Erz]

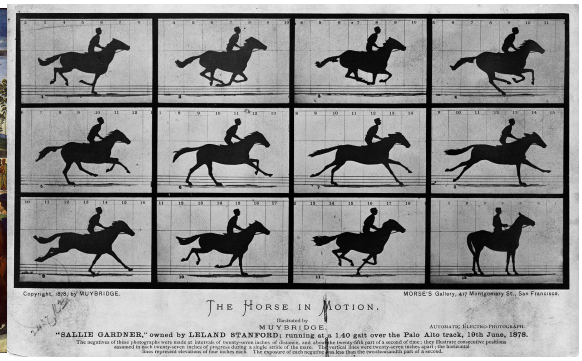


Figure 4.3: The Horse in Motion [Kal]

Much, much later, in the 19th century, the first *stereoscope* was invented by Charles Wheatstone[10] [HF20]. A stereoscope is a device in which a pair of slightly horizontally offset images of the same scene are presented to each eye, to provide additional depth perception. This exploits the fact that our visual system is always integrating two separate images, offset horizontally by the *inter-ocular* distance, in normal viewing conditions. A picture of this device can be seen in Figure 4.4. In the 1930s, a portable version of a stereoscope called the *ViewMaster* became commercially successful. The toy allowed people to switch images and see immersive

[10]English scientist and inventor of many scientific breakthroughs.

photographs of places around the world. Some of the innovations of the stereoscope were the increased FoV and blocking of distracting boundary stimuli resulting in increased immersion. The ViewMaster, and other similar products, can be seen as an early precursor to today's *Head Mounted Displays* (HMDs), commonly used in VR experiences.

Figure 4.4: Brewster-type[11]stereoscope, 1870 [Filb]

In 1878, a few years after Wheatstone's stereoscope, one of the first examples of stroboscopic apparent motion was created by Eadweard Muybridge[12]. This effect refers to the ability to generate apparent motion by flipping through a sequence of images at a fast rate. Figure 4.3 shows Muybridge's famous work: "The Horse in Motion", which depicts several frames of a running horse, as it seemingly moves across space. To capture this sequence, 24 separate cameras were required. These were triggered by the horse as it moved along a track using ropes, which the horse's legs pulled on when running. The cameras were offset equidistantly to capture movements in a synchronized fashion. To reproduce the images, a *zoopraxiscope*, also an invention of Muybridge's, was employed. The zoopraxiscope is a disc containing multiple image frames which can be used to project a moving image in a recirculating fashion [LaV16].

[11]Sir David Brewster (11 December 1781 – 10 February 1868) was a Scottish scientist, inventor, author, and academic administrator.

[12]Muybridge was an English photographer important for his pioneering work in studies of photographic motion (9 April 1830 – 8 May 1904)

Some 40 years later, in 1915, another well-known method in film-making used to increase realism was invented by Edwin S Porter[13]:*3D film-making*. Modern 3D movies are created using special camera lenses and subsequently viewed using *polarized* light filters. These filters, at the subject's eye level, allow certain frequencies of light to pass into one eye, while blocking others. While there is a single image on the screen, two different images are perceived by each eye, much like with the stereoscope. *Stereopsis* is the formal term describing the process of integrating two overlapping visual fields to achieve depth perception. When images are encoded into this format the resulting images are also called *anaglyphs*. 3D movies are still in use today in some VR experiences. They can also be found in commercial movie theatres across the world.

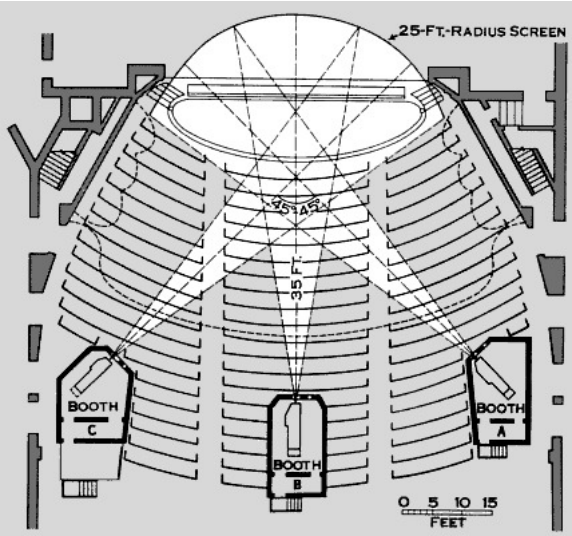


Figure 4.5: Cinerama Diagram [aEW]

30 years later, a different technique was created to improve the sense of realism in movie systems. The method consisted of increasing the FoV by using a wider screen that surrounds the viewer, accounting for viewers' peripheral vision. The *Cinerama*,

[13]Edwin Stanton Porter (April 21, 1870 – April 30, 1941) was an American film pioneer, most famous as a producer, director, studio manager and cinematographer.

from the 1950s, is an example of such a system. It used three projectors to extend the FoV and projected the image upon a concave surface for a richer viewing experience. It was created in the 1950s, and can be seen as a precursor to CAVE (Cave Automatic Virtual Environment) systems used for VR in the 90's[14].

A few years later, in 1957, Morton Heilig introduced the *Sensorama*, which combined: motion pictures, stereo sound, vibration, wind, and even smells, in a personalized stereoscopic *multi-modal* experience. The device resembles an arcade machine with an enclosure surrounding the subject's head, in order to reduce distractions. Unfortunately, the Sensorama had a fixed perspective, which meant the user was limited to a single orientation. It was also very large and expensive, which ultimately led to its downfall. Figures 4.5 and 4.6 depict the Cinerama and Sensorama respectively.

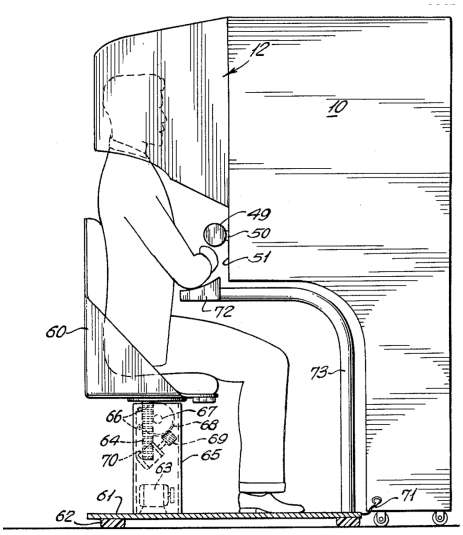


Figure 4.6: Sensorama Patent Figure [Hei61]

Then, in 1965, a major step would be taken toward the realization of VR. That year, Ivan Sutherland[15] introduced the concept of "The Ultimate Display", which

[14]These are described in more detail later.

[15]Ivan Sutherland (born May 16, 1938) is an American computer scientist and Internet pioneer,

he described as: "a room within which the computer can control the existence of matter." [Sut65] A few years later, Sutherland and his team would build *The Sword of Damocles*[16], regarded today as the first VR Head-Mounted Display (HMD). The display was a ceiling-suspended device capable of displaying simple wire-frame shapes according to the users' head movements [HF20]. This demonstrated for the first time in a virtual system the idea of the *perception of stationarity*, which consists of making a static object appear to remain in its position while one moves their head.

In the 1980s a number of advancements were made in the field, mainly by government agencies:

1. In 1982, an advanced flight simulator called the *Visually Coupled Airborne Systems Simulator* (VCASS) was created by the US Air Force Medical Research Laboratory.

2. In 1984, the *Virtual Visual Environment Display* (VIVED) was developed by the NASA Ames research center.

3. In the late 1980s, the term "Virtual Reality" was coined by Jaron Lanier, founder of the Visual Programming Lab (VPL). VPL would go on to develop the *DataGlove* and the *EyePhone* HMD. "Although all normal vision is lost wearing an HMD, the data glove allows the user to hold up their gloved hand in front of their face and see a digital representation through an HMD." [Dix06]

Lanier went on to expand the DataGlove into a full-body *DataSuit*, capable of tracking users' different body joints. Using these suits, multiple users were able to interact inside a virtual environment, and even change their physical appearance. In gaming, one's virtual representation is often called his or her *avatar*.

A different approach, developed in the '90s, was the *CAVE* (Cave Automatic Virtual Environment) [Muh15]. This environment, developed in 1992 at the University

widely regarded as a pioneer of computer graphics.

[16]You may find information regarding the myth here.

Figure 4.7: EyePhone HMD and DataGlove by VPL [Pap]

Figure 4.8: HMD and Wired Gloves - Ames Research Center [NAS06]

of Illinois, uses a large number of projectors to, ideally, cover the entire surface of the room. Some CAVE systems also use 3D video in order to improve the depth perception of the image. The benefit of this approach is that the user does not need to wear any heavy equipment on their face which might restrict their ability to move. The downside is that such environments are often very costly to set up given the large number of projectors needed to create the visual illusion. There are similar environments that use arrays of display panels in lieu of projectors to achieve higher visual fidelity.

During this time there were also a small yet important number of artists who experimented with VR. Kazuhiko Hachiya created a two-person VR experience called *Inter Discommunication Machine* in 1993, which switched one player's sight and sound for the others. The idea was to blur the lines between "you" and "me" [Dix06] by switching sensory stimulation. At the Banff Centre, in Alberta, Canada, several VR projects were developed by a range of artists, all documented in Moser

Figure 4.9: CAVE [Dav01]

and McLeod's 1996 book "Immersed in Technology". Dixon [Dix06] describes some of these projects in more detail.

4.3.1 Hardware

Degrees of Freedom (DoF)

Chapter 2 of LaValle's book [LaV16] provides us with an overview of hardware practices important to XR. An important distinction LaValle makes in regards to hardware is control in either *three degrees of freedom* (3DoF) or *six degrees of freedom* (6DoF). An ordinary object moving and turning in 3D space has 6DoF. Three of its degrees of freedom correspond to its changing position in space. These *translations* include:

1. Horizontal (side-to-side) or $y - axis$ motion.
2. Vertical (up-down) or $z - axis$ motion.
3. Distal (front-back) or $x - axis$ motion.

The other three degrees of freedom include roll, pitch, and yaw, which correspond to rotations along the x, y, and z axes[17]. Sometimes the *user* will be given only 3DoF, in which case this is normally in the form of rotation changes. This is generally the case when viewing 360° videos, for example, videos in which a scene is captured in every direction using spherical lenses or arrays of cameras. The person can experience all directions of the video, by moving their head, but, generally speaking, cannot change the point of capture. In other words, translations in space are restricted. In this case, the *tracking system* (ie. the sensors in the display) only provide 3DoF.

Some controllers, such as the Sony Playstation *DualSense* controller (Figure 4.10), have only 3DoF, which means they can send rotational information but are not designed for translation tracking. Sony's VR ecosystem uses the *Move Controller* (Figure 4.11), which has an LED that is tracked by a camera system, which, when paired with its internal sensors, allows 6DoF. Some HMDs offer 3DoF only, while others are paired with external sensors, which report the position of the HMD in space to the VR system. This is generally accomplished using external cameras which track lights on the HMD. Other, newer systems[18], use internal cameras to perform this task.

World-fixed v. User-fixed

Another important distinction we should make is between *world-fixed* and *user-fixed* systems. As we will see, user-fixed aural and visual displays offer a far more cost-effective and portable way of reproducing spatial audio and virtual scenes. World-fixed systems are less burdensome to the user, but often cost more to implement.

World-fixed systems refer to surround-sound systems, in the aural domain[19], and CAVE-like environments, in the visual domain. User-fixed systems refer contrastingly

[17]This is our preferred coordinate system, it should be noted that others exist.
[18]For example the Oculus Quest.
[19]Not exclusively, this could also be a *cross-talk cancellation* stereo speaker system.

Figure 4.10: Sony DualSense Controller [Coc]

Figure 4.11: Sony Move Controller [EA]

to binaural headphone systems and binocular HMDs, in the two respective domains. Both of these approaches to presenting XR experiences can be mixed. Each of them also has its own benefits and drawbacks. For sound systems, some of the key trade-offs between user and world fixed systems are:

1. World-fixed systems require much more power (energy). Consider the power savings from binaural (headphone) reproduction versus surround-sound systems.

2. Headphones allow a sense of privacy, whereas surround-sound systems might disturb other people in your environment or your neighbors.

3. User-fixed systems can be uncomfortable if used for long periods since the user is required to wear heavy electronics on their head.

4. World-fixed systems might allow for multi-user experiences more seamlessly. However, there might be a noticeable drop-off in quality if users are outside the sweet spot[20].

5. Multi-user experiences are possible over headphones but require different sound processing for each user, which might prove computationally more expensive.

6. Headphones are often much cheaper than surround-sound systems.

The trade-offs are quite similar in the visual domain. In world-fixed environments, it will only be necessary to modify the sound and visual scene according to users' translational movements[21]. Since they are already surrounded by audio-visual stimuli, natural rotational movements do not need to change anything in the environment. In contrast, user-fixed environments will need to adapt to user head rotations to create the illusion that they are inside an encompassing environment, which requires sophisticated graphics systems and a powerful computer.

VR Sickness

An important problem to note here is the sometimes poor updating of images in VR within user-fixed experiences. The slow frame rate, or update speed, of these systems, can cause what is known as *VR sickness*. This is caused by a mismatch in sensory information coming into the brain. This happens when one's *vestibular* organs, used to maintain balance, are not in sync with one's visual organs. This is similar to the nauseating feeling we get when reading a book in a moving car. This

[20]Optimal viewing or listening position. HMDs and headphones do not have sweet spot problems.
[21]If 6DoF is desired.

conflict of sensory information is also called *vection*. Over the years, GPUs[22] have become increasingly powerful, and, as a result, it is possible now to update images on HMDs quickly enough that VR sickness is eliminated[23]. CAVE-like environments do not suffer from this problem, however, much like in surround-sound systems, there is an optimal viewing area beyond which the image will not be of good enough quality. Ideally, in user-fixed systems, the frame rate is at least 90Hz, meaning there are 90 *frames* per second. This number has become standard in the VR industry since it seems to be the smallest frame rate at which VR sickness is reduced to tolerable levels.

4.4 Binaural Audio

4.4.1 Introduction

This dissertation focuses on binaural playback as a primary method of spatial audio playback given its smaller economic footprint. There are two types of binaural audio: fixed-perspective and head-tracked. Fixed-Perspective Binaural Audio (FPBA) is captured using in-ear microphones or a dummy head (e.g., a mannequin head with microphones inside its ears). FPBA can be effective under certain conditions, one of the most popular examples of it is the virtual barbershop video where we hear scissors and clippers moving around our head providing a visceral experience. In contrast, Head-Tracked Binaural Audio (HTBA) relies on a gyroscope system that can track the user's head movements and modify the sound field accordingly to simulate acoustic object permanence. In our work, we opt generally for HTBA as we find it more engaging and effective at conveying spatial audio than FPBA. The three

[22] A Graphical Processing Units (GPU) is the part of the computer responsible for calculating and displaying images.

[23] The amount of sensitivity to VR sickness is different from person to person, but, generally speaking, these systems work well enough now that they are mass produced and sold worldwide.

XR projects featured in this chapter feature 3DoF spatial audio using the built-in gyroscope from mobile devices as the system for rotating the sound field. There are currently no open-source HMDs that could be adopted for 6DoF spatial audio in virtual environments. The other motivation for binaural audio is the cost. One does not need to have a large number of speakers to experience virtual surround sound, a cheap pair of earbuds and a smartphone or desktop computer are all that is needed. These HTBA systems require convolution with HRTFs, the three JS libraries which we make use of for our three XR projects use this method to simulate spatial audio over headphones.

4.4.2 Resonance

As noted, there are several frameworks available for creating binaural spatial audio online today. Among them, the most popular open-source solution is likely Resonance[24]. Gorzel et al. [GAK+19] describe some of the techniques used to optimize the HOA encoding and playback implemented in Resonance, which allow it to run on cheap hardware and older devices.

Encoding Optimization

In their publication, the authors describe the implementation of a *Look Up Table* (LUT), instead of using *cos* or *sin* functions in JS. The *spherical harmonic* (SH) coefficients, needed to encode sources into ambisonics, are pre-computed and retrieved based on the current angle of the sound source. Encoding in ambisonics is the process of taking a raw audio signal and creating a sound field from it, with the audio at a particular position[25]. Since the angle of the source is not known a priori, an entire quadrant of the SH is computed. In order to provide smooth transitions between coefficients interpolation is applied. The use of a LUT is a common technique

[24]From Google.

[25]Encoding ambisonics should not be confused with parametric compression encoding.

in audio software applications which improves the speed of the algorithm on slower processors.

Exploiting Symmetry

This paper points out the possibility of exploiting SH symmetries to reduce the memory requirements of the system. Gorzel et al. note that SHs have well-known symmetries, making it possible to derive all coefficients from a single quadrant, determined by the azimuth and elevation angle of the sound source. By simply calculating the front-left-top quadrant of the sphere, all the other values can be calculated simply by applying sign changes. This memory reduction means faster loading times for the user and consequently allows people with less powerful devices to experience these works.

Assuming Head Symmetry

Another technique employed by the authors involves reducing the number of convolutions required for binaural synthesis by assuming left/right hemispherical symmetry. Consider a binaural sound source at 45° to the left of the subject, with an elevation of 0°. If we were to simply swap the left and right channels then, in essence, it would be like presenting the sound source at 45° to the right of the subject[26]. By extension, a single HRIR is used for sources with azimuth 0°. By assuming symmetry, the number of convolutions is reduced by half, which allows for a smoother experience on cheaper hardware. This method is also adopted by other authors (e.g., [PPQ16]).

[26]As noted, this assumes that the head is exactly symmetrical; further studies are required to see if this feature degrades the sound quality of the renderer substantively, especially when using personalized HRTFs.

HRTF Expansion

Gorzel et al. use spherical domain representations of the SADIE HRTF data set [KD15][27] in order to further reduce the number of convolutions required. In this approach, a *decoding matrix* is first calculated based on the positions of the desired IRs. A matrix of impulse responses, corresponding to the same directions, is then multiplied by the transpose of this matrix, which effectively encodes the IRs into the spherical domain. This method, ubiquitous today in binaural ambisonic reproduction, is called the *Least Squares* (LS) solution to the system of linear equations.

The decoding matrix is found via a *Moore-Penrose pseudo-inverse* of the matrix $\mathbf{L}$, shown in Equation 4.1[28]. As we can see, each column of our matrix corresponds to a direction, and each row corresponds to a harmonic. The values are calculated using our standard real-valued ambisonic equation[29]:

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

$N_n^{|m|}$ corresponds to the normalization[30], and $P_n^{|m|}$ corresponds to the associated Legendre polynomials, which are solutions to the wave equation, generally computed using numerical libraries[31].

$$\mathbf{L} = \begin{bmatrix} Y_0^0(\Phi_1,\Theta_1) & Y_0^0(\Phi_i,\Theta_i) & \dots & Y_0^0(\Phi_N,\Theta_N) \\ Y_1^{-1}(\Phi_1,\Theta_1) & Y_1^{-1}(\Phi_i,\Theta_i) & \dots & Y_1^{-1}(\Phi_N,\Theta_N) \\ \vdots & \vdots & \vdots & \vdots \\ Y_n^m(\Phi_1,\Theta_1) & Y_n^m(\Phi_i,\Theta_i) & \dots & Y_n^m(\Phi_N,\Theta_N) \end{bmatrix} \tag{4.1}$$

Computing the Moore-Penrose pseudo-inverse is given by [GAK+19] as:

[27] In lieu of the traditional virtual speaker approach.
[28] Matrix $\mathbf{L}$ is the SH representation of our HRTF directions.
[29] The equation is unnumbered since it is discussed in detail in Chapter 2.
[30] SN3D in this case.
[31] Refer to Chapter 2 for more information on both normalization and Legendre polynomials.

$$\mathbf{D} = \mathbf{L}^{\dagger} = \mathbf{L}^{\mathbf{T}} \left(\mathbf{L}\mathbf{L}^{\mathbf{T}} \right)^{-\mathbf{1}}. \tag{4.2}$$

Any matrix, regardless of its dimensions, can be multiplied by its transpose to yield a square matrix - transposing is as trivial as swapping rows and columns. The inverted matrix[32] is subsequently multiplied with a transposed version of the original matrix $\mathbf{L}$. The transposed matrix $\mathbf{D}$, the decoding matrix, is used to expand the HRTFs into the spherical harmonic domain. Equation 4.3 shows the HRFT spherical harmonic expansion.

$$\hat{\mathbf{H}} = \mathbf{H}\mathbf{D}^{\mathbf{T}} \tag{4.3}$$

The result is a matrix with $(N+1)^2$ columns, and as many rows as there are filter coefficients. This matrix can directly be convolved with the B-Format ambisonics spherical harmonics (e.g. these two matrices have the same number of harmonics). The binaural signals are then the result of the inner product (e.g. convolution) between the ambisonic coefficients and the SH coefficients of the HRTFs [PPQ16]. In the case of a 3OA[33] signal and a 26-point Lebedev grid[34] the number of convolutions is reduced from 26 to 16 resulting in a $\sim$62% reduction in the number of convolutions.

An added benefit of this method is that any *shelf-filtering*, such as a *MaxRe* shelf-filtering, can be done offline, rather than in real-time during decoding. This reduces the required amount of CPU and as a result improves speed on low-end mobile devices [GAK+19] (e.g. since the decoder is known ahead of time, we can filter the $\hat{\mathbf{H}}$ matrix in advance). The same process is implemented in Politis et al. [PPQ16], however, instead of using *MaxRe* decoding filters, the authors describe

[32]Inverting the square matrix involves Gaussian elimination. This can be done by hand but for large matrices we usually use computing libraries to perform the task.

[33]Shorthand notation for third order ambisonics.

[34]Lebedev quadrature is a special sampling scheme for spherical harmonics in which the number and location of points, along with integration weights, are determined by enforcing exact integration of spherical harmonics up up to a given order. T-designs and Fliege grids are other examples of sampling schemes that enforce exact integration.

using the AllRAD [ZF12] method for creating the decoding matrix, which is then binauralized using a VLS approach.

It should be noted that Gorzel et al. make no mention of the well-known timbral issues with the LS method (see [SZH18]). They also make no mention of alternate decoding strategies such as *mode-matching*, *AllRAD*, or EPAD [ZPN12][35]. A possible improvement to Resonance, therefore, could be to incorporate better binaural filters, based on the *Magnitude LS* (MagLS)[36] described in [DMKH+20], and test the effect of other decoding methods described in [PPQ16].

4.4.3 Virtual Loudspeaker Method (VLS)

In the virtual loudspeaker approach, we imagine that we are decoding to an array of real loudspeakers, and find a suitable decoding matrix according to this criteria. Any of the three aforementioned decoding methods is valid here.

Subsequently, we derive our matrix of decoded signals and match each speaker direction with its corresponding HRTF from our data set. Naturally, this means that the decoder will be designed according to the HRTFs present in the set. The main constraint in this method is that the number of decoding directions must be larger than the number of harmonics, in order to accurately represent the sound-field[37]. With the SH HRTF expansion, the spherical sampling scheme must contain more points than the number of harmonics in that order, or at least be equal.

Politis et al.[PPQ16] provide Equation 4.4 describing the *virtual loudspeaker* (VLS) approach to binaural decoding in the frequency domain. E.g.

$$\mathbf{x}_{\text{bin}}(f) = \mathbf{H}_{\text{LR}}(f)\mathbf{D}_{\text{vls}}\mathbf{b}(f) = \mathbf{D}_{\text{bin}}(f)\mathbf{b}(f) \tag{4.4}$$

[35] Although for regular designs these might be equivalent.

[36] [SZH18] also describes the *Time Alignment* (TA) and *Spatial Re-sampling* (SPR) methods, which Resonance does not yet offer.

[37] If the decoder does not meet this criteria, multiple orders of SHs can be discarded at the expense of spatial resolution.

where,

$\mathbf{x}_{\text{bin}}(f)$ corresponds to the 2 by 1 binaural bins at frequency f[38],

$\mathbf{H}_{\text{LR}}(f)$ is the 2 by G[39] matrix of L/R HRTF filters suitably arranged,

G is the number of decoding directions, and

$\mathbf{D}_{\text{vls}}$ is the G by $(N+1)^2$ virtual loudspeaker matrix (*not in the frequency domain*).

Additionally, $\mathbf{b}(f)$ is the $(N+1)^2$ by 1 encoded audio signal in the frequency domain, where each value is a single bin of a single harmonics[40], and $\mathbf{D}_{\text{bin}}(f)$ is the binaural decoding matrix for the virtual loudspeaker approach. Much like in Gorzel et al. [GAK+19], the number of HRTFs is reduced by 50% in this implementation by assuming L/R symmetry. For a web-based system, binaural signals filtered with the proper HRTFs are then delivered to the client using the *Web Audio API* (WAA) browser specification. This is the approach used in the JSAmbisonics library embedded in HOAST [DMKH+20] which we used to showcase material from Chapter 2.

As already discussed, SH expansion of HRTF data has timbral problems which have since then been addressed in other publications [SZH18], but not implemented in the JSAmbisonic kit. The VLS method provides us with the flexibility to decode HOA using alternate strategies, which might provide better sound quality. This library, therefore, provides us with a way to test out and compare various binaural decoding methods, in order to determine if optimized SH methods, for example, can achieve improved sound quality compared to VLS binaural decoding. It remains to be seen if the additional computational load of VLS can take precedence over the efficiency of the direct approach (e.g. SH HRTF expansion).

[38] One for each ear.
[39] Where each row corresponds to a direction, and each column to an ear.
[40] The paper uses **a** but we prefer **b** to denote B-format signals.

4.5 Selected Works

4.5.1 POI

Pigment of Imagination is a popular music track touching on themes of imagination and space travel. The song is composed and produced by Timothy "Ill Poetic" Gmeiner and features vocals by both Ill Poetic and Nick Tolford, as well as flugelhorn by renowned jazz artist Stephanie Richards, and additional synthesizer parts by DJ/Producer King Britt. The stems from the song were used to create a special HOA (Higher Order Ambisonic) 3D spatial audio mix, which you can hear online via the A-Frame project we created. In addition to the HOA, the A-frame project features a custom movie that provides dynamism to the experience. During our experiments, we found that our HOA files did not load on most mobile devices, thus, alternate mixes in FOA (First Order Ambisonic) and static binaural are also made available. POI (WebXR) was submitted to the DAVAMOT AudioVisual Symposium and was created in collaboration with Eito Murakami and Tim Gmeiner. Currently hosted on: https://gabrielzalles.neocities.org.

Choice of Medium

A-frame and WebVR were selected due to their interoperability between hardware devices and because it is free and open source. Anybody with a desktop computer or phone can view our work, it is not necessary to have a VR (virtual reality) headset. We are interested in the problem of socio-economic access in spatial audio compositions, a dynamic that has led to a technocratic society that values channel count over artistry. Instead, here we focus, both in process as well as product, on the accessibility of our components, from the economic perspective.

Ambisonic has an isotropic characteristic which means it considers all directions as equally important in the reproduction and capturing of audio. This method of

spatial audio synthesis is characterized by its use of basis functions in lieu of channel-based information. In surround sound, we usually have 6 channels corresponding to left, right, center, left surround, right surround, and subwoofer. Ambisonics is not "channel-based" rather, the four audio channels in FOA correspond to information in the X, Y, and Z axes, as well as a W channel which corresponds to an omnidirectional signal (e.g., sound arriving from all directions equally).

Ambisonics is very flexible because, with an HOA or FOA mix, one can reproduce the material over any number of speakers from 1 to infinitum. Furthermore, ambisonics counts with linear transformations such as rotation, which allow us to seamlessly reproduce audio binaurally. In binaural reproduction, we defined a number of virtual loudspeakers and played the audio through these consequently filtering the sound with HRTFs(head-related transfer functions). These HRTFs contain the time/gain/frequency differences for sounds played from various positions, for both ears. In other words, they characterize the effect our heads and ears have on sound.

This project is part of a larger initiative called SElectOr which seeks to bring undergrads, graduate students, and faculty together in collaborative music-making projects. In addition to these three main groups, we also invite alumni, staff, and external members to contribute. The aim is to broaden access to the tools and skills required to create modern multi-media works and forge community in the process. The group has been operating since 2018.

Technical Elements

This piece does not require a traditional multi-channel audio system or projection equipment. Rather, for the presentation, we suggest a Google Cardboard device (and an Android phone) with the website pre-loaded. Guests would be able to individually come up to the table and watch/listen to the work. Naturally, given the binaural nature of one would also need to provide a pair of headphones. Another option is to set it up on a computer, although we find the Cardboard experience to be slightly

more immersive. The project is currently not operational on HMDs, such as Oculus or Vive. Unfortunately, none of the creators has access to one of these devices, so we were only able to test on Cardboard and Desktop.

Figure 4.12: Google Cardboard

The main element allowing this work to be possible is the A-Frame Ambisonic component, which is an A-Frame component providing a high-level abstraction to the Omnitone API. The Omnitone project is responsible for the binaural decoding of HOA/FOA sound files which were created in Reaper, a cross-platform DAW, using the IEM Plug-In Suite (Figure 4.13 shows the GUI for the IEM stereo encoder).

Following the creation of our raw HOA sound file, with 16 channels, in uncom-

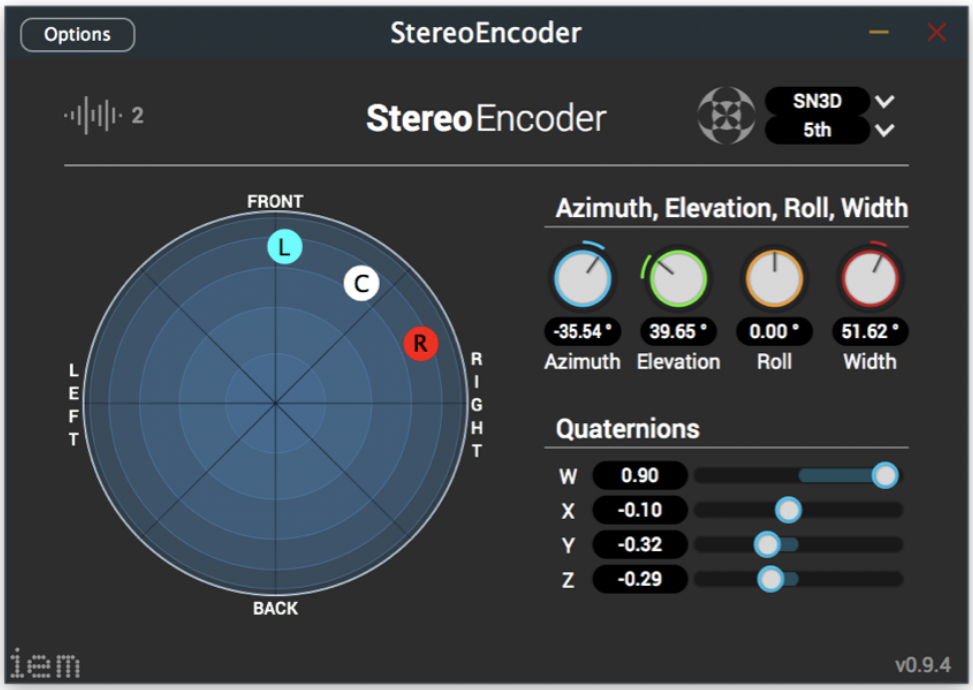


Figure 4.13: IEM Stereo Encoder GUI

pressed format, the sound files were imported into Audacity for compressing and splitting into two 8-channel files, as per the provided example. The OGG codec was chosen, also as per the example, and different quality settings were tested to determine a compromise between load time and over-compression artifacts. Due to the length of the song, and the number of channels, we used a timeout feature in A-Frame which gives the listener's connection 10 seconds to download the 67MB of audio required for the project (as well as other assets). We ultimately selected a quality factor of 5 in the OGG compression, out of the possible ten quality levels.

As per the recommendation of the A-Frame documentation, we opted for the server solution provided by neocities.org (which is open-source). The animation-timeline component was used to create the sequences involving A-Frame primitives. A particle system component was adopted to give the piece some additional move-

ment. Several spheres were textured and set to rotate around the listener serving the theme of outer space. A 360-degree photograph from the European Southern Observatory was used to texture our sky element after it had been resized to a lower resolution (4K) in order to reduce its size. Some GLTFs (3D models with low-polygon count) were also imported into the scene which were collected from Google Poly.

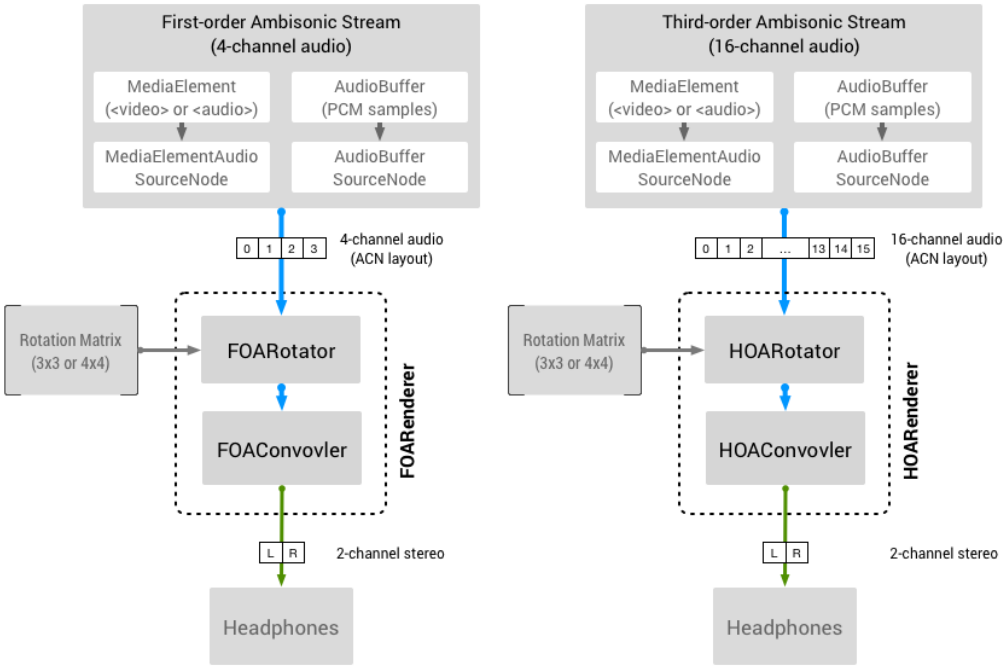


Figure 4.14: Omnitone Diagram

The project was tested using several browsers and mobile phones with various internet connections, all in the San Diego area. While the assets loaded properly with some connections, a longer timeout was needed on other networks. The preload feature of A-Frame has not yet been implemented into our project for HOA. On mobile devices, we were unable to get Omnitone to playback HOA, so alternate versions in FOA and binaural were created to satisfy the needs of each user, based on their connection.

The HOA version on Desktop browsers such as Chrome and Firefox sounds really

distinct from the fixed binaural versions as the virtualization of moving sources is clear and distinct. The vocalists' stems, which are fixed at the front sonically, clearly change in timbre as one changes perspective. In future projects, we are interested synchronizing the bimodal elements of the audio-visual experience to further facilitate the localization of sound sources – animations of flying birds for example could be attached to sound sources representing these objects. Future work might involve comparing these three reproduction methods via a survey.

4.5.2 Continuum

Continuum is a WebVR experience inspired by the concept of the multiverse: the idea that there are an infinite number of possible universes all occurring simultaneously in other timelines. In this project, the manner in which this metaphor is conveyed is by the use of remixes, which are essentially permutations of an original musical idea. Continuum is as much a WebVR experience as it is a conceptual album, the experience adopts the formal structure of a binary tree forcing the listener to choose between two possible scenes at each layer. The project is part of an exploration into open-source tools that can be exploited for the dissemination of spatial audio.

Introduction

WebXR is a powerful medium for the dissemination of spatial music as the experiences can be accessed not only on mobile devices but also on desktop computers or head-mounted displays (HMDs). In contrast to native mobile applications, which can be developed using an operating system (OS) specific software development kit (SDKs), WebXR experiences generally do not require one to own a smartphone device or HMD. In addition, these cross-platform applications do not need to be ported or re-compiled for multiple operating systems or devices but are instead designed to

work seamlessly across hardware platforms.

Spatial audio has become an increasingly popular subject in new media in the last decade, with major investments by companies such as Meta and Google. Unfortunately, this compositional dimension of music-making is not equally accessible to everyone, and, to some extent, it appears that merit is attributed to artists for ownership of resources and the quality of their tools. Our research involves investigating how spatial audio compositions can be created, captured, and disseminated using low-cost and open-source technologies.

Continuum is one of the projects we created that relies on open-source JavaScript (JS) libraries to accomplish this task. In the development of the work, we also learned about many of the possible drawbacks of working with WebXR frameworks over mobile spatial audio SDKs or game engines for developing with HMDs.

This section will outline the development of our composition and address some of the strengths and deficits of WebXR technologies. We will also discuss other WebXR projects that have made use of spatial sound, focusing on those featuring open-source and low-cost technologies. Lastly, we will address the ethical implications of working in this domain, and what criticisms one might level at a project of this nature.

Related Works

McArthur et al. [MvTGBKH21], in their 2021 paper, present a survey of 3D audio via the browser which covers the affordances and limitations of several libraries for spatializing audio on the web and provides some useful guidance for newcomers to the domain. McArthur discusses several libraries that might be of relevance to the reader such as the: Web Audio API (WAA), JSAmbisonics library [PPQ16], Resonance Audio SDK[41] [GAK+19], HOAST (Higher-Order Ambisonics STreaming) platform [DMKH+20], and many others. The purpose of the survey is not to categorize any one system as inherently superior, nor is there an evaluation of these, rather it hopes to

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

delineate the strengths and weaknesses of each, allowing the reader to make informed decisions. In the concluding remarks, the authors note some general principles to follow when developing for the web[42]:

- Keep media files as small as possible.
- Develop for a primary browser.
- Consider the size of your server if live streaming.
- Invest in a test set-up to simulate actual use.

In our projects concerning WebXR, we found these principles to be acutely pertinent. In Continuum, we opt for the Resonance Audio SDK for the spatialization of point sources via an A-Frame plug-in[43]. In this process, each sound source was encoded as individual mp3 files to reduce the load time of the scene and we set a limit of eight, on the number of stems each scene could have. We also decided to confine our models to glTF files, specifically those designed to have a low polygon count. Much like the authors of the aforementioned survey, we tested our experience on several different browsers and found that the behavior of the scene was not consistent across browsers[44]. While it is also important for us to test the behavior of the experience on various HMDs, it was more important that it worked on mobile devices, as we believe smartphones are more accessible to a global audience.

Another publication by Çakmak and Hamilton [ÇH20] describes the development of a VR experience pertinent to our discussion. In the abstract of the paper, the authors note one important deficit of WebXR technologies, which is the unfortunate reality that developing technologies can become deprecated or obsolete during the period of making the work. While we might be able to control for versions of

[42]The list was redacted and paraphrased.
[43]`https://github.com/mkungla/aframe-resonance-audio-component`
[44]We recommend Firefox for viewing Continuum.

libraries such as Resonance or A-Frame, browser changes cannot be predicted and are not always backward compatible. The only solution to this, we believe, is to not only specify in the documentation the browser that was used but the version as well. The paper provides a very reliable framework for creating multi-media WebXR experiences, however, many of the tools relied upon are not open-source, and some may be financially burdensome. The introduction to the paper also provides a quick, albeit important, glance at composers who made important contributions to the field of spatial music.

Jot et al. [JAHS21] discuss in their publication an approach for six-degree-of-freedom (6Dof) object-based interactive audio. This framework considers optimization while addressing the: position and orientation of the source and listener, velocity vectors contributing to Doppler shifts, distance-based attenuation, source directivity models, and efficient reverberation, which can contribute to localization based on early reflections. 6DoF corresponds to experiences in which listeners can move laterally in addition to the 3DoF provided by yaw, pitch, and roll (e.g., head rotation). Ambisonics has become a simple way to create spatial sound in 3DoF as linear rotation matrices exist allowing the basis functions upon which this system relies to be non-destructively transformed. The resulting virtual loudspeaker signals, after decoding, are filtered using binaural impulse responses, which contribute spectral and timing cues allowing us to localize sounds.

In Continuum, the ability to move laterally is moot, as we target mobile devices that can experience the work via a browser and cardboard HMD. This method was chosen due to its low cost and open-source nature. Cardboard experiences are generally only 3DoF, as the interactive components of the system are not ergonomically suited for 6DoF. Navigation in 6DoF is possible in the desktop version, however, the immersion level is lower in this configuration. Navigation is also possible from one scene to the next. Continuum counts currently with seven scenes which can be navigated using a fuse-based cursor, implemented in A-Frame. The cursor works by

casting a ray from the camera look direction and calculating intersection points with other objects in the scene, looking at the red cube in the scene for a few seconds begins audio playback[45], and looking at one of the circles in the scene, allows us to navigate to the next song[46]. See Figure 4.15 for a visual representation of the scene.

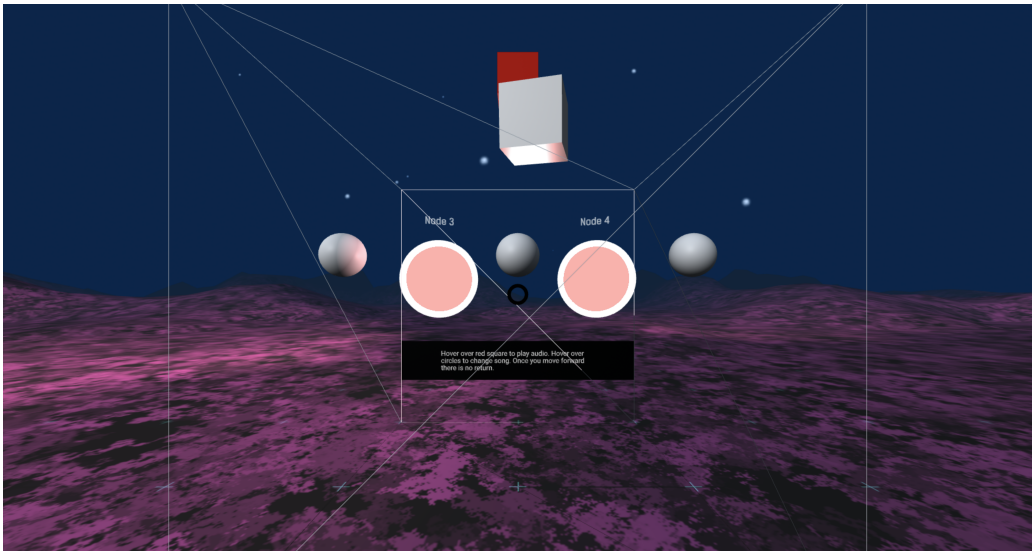


Figure 4.15: A sample image from Continuum depicting Scene 1

Simiscuka et al. [STV+22] also provide an important contribution to the topic of WebXR and spatial audio. The authors here conducted a Quality of Experience (QoE) evaluation using various audio quality levels to determine the merit of ambisonic audio in an artistic context. Specifically, the authors developed a 360-degree video player which was evaluated by 24 participants from various countries. The study showed that: "more participants agreed or strongly agreed that the clip with ambisonic audio improved the immersiveness of the experience" - when compared to stereo or mono. Enjoyment and quality were also rated as higher with footage that featured stereo or ambisonics. In every question but one, ambisonic performed better

[45]This interaction is required for the AudioContext() in the WAA to allow audio playback.
[46]Using the `link-controls` A-Frame component.

than stereo[47]. This validates the use of ambisonic libraries in WebXR experiences, as quality, engagement, enjoyment, and, various other attributes, were all improved by the use of the technology.

Description

The name Continuum comes from the ever-expansive nature of the universe, and the notion that time has no end or beginning in a world where infinite dimensions exist simultaneously, as has been proposed by numerous theorems. The piece is part of a larger initiative at our University called (Y) which defines itself as a creative community seeking to dissolve social barriers and create bridges between faculty, graduate and undergraduate students, staff, and alumni. During the COVID-19 pandemic, and in the aftermath of it, our academic institution remained partially shut down, which was part of the rationale for engaging with a WebXR project.

The idea for the piece came about over multiple meetings with undergraduate students, who volunteered their time to this endeavor, and the author, who has been directing the group since its inception. Many of the artistic and technical decisions resulted from the facilities and competencies of the author, who had previous experience with A-Frame[48] (e.g., the WebXR library chosen for this project) and is an amateur guitarist. The decision to use Resonance as the method for spatialization came from the author's larger research direction, which investigates how free and open-source software (FOSS) and low-cost hardware can be leveraged to disseminate, record, and synthesize spatial music. In this process we had already developed artistic projects with Omnitone and the WAA, thus it seemed logical to explore Resonance as an alternative tool.

The process we followed was rather simple; the first step was to create three original compositions, one for each layer of our binary tree. In total, there would

[47]The question being: "I believe that the immersive experience is comparable to a live opera".
[48]`https://aframe.io/`

end up being seven total scenes. Each of the nodes to the right of the one preceding it modulates to a musical key a fifth above, creating musical flow, while the nodes to the left would modulate to a fifth below. In order to define the color palette of each scene, we used the visible light spectrum as our template and decided to use longer wavelengths for nodes to the right, and shorter ones for nodes on the left[49]. The nodes at the top of the tree are also supposed to be darker while those lower are brighter. An example of a binary tree is depicted in Figure 4.16.

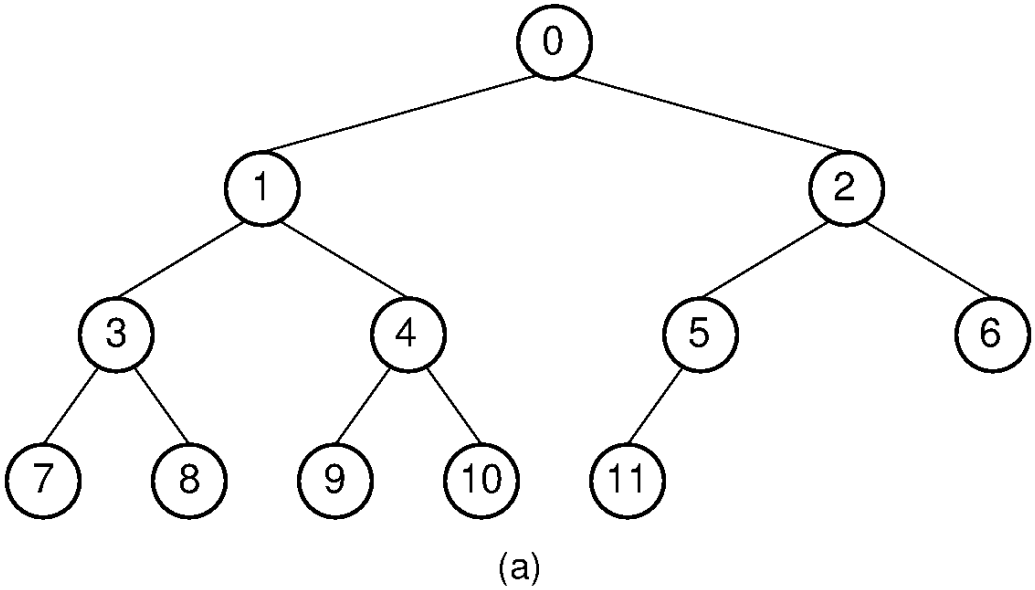


Figure 4.16: An example of a binary tree (data structure)

Each of the three compositions used for the remixes which populated each layer was composed in Ardour and then stems were exported for each of our collaborators to manipulate. These were transposed according to the requirements of the project by each team member, most of whom selected Ableton[50] as their tool for the remixing process. Each of these sessions subsequently was condensed into at most eight mp3

[49]Using the application `https://colorswall.com/` to facilitate this process.
[50]`https://www.ableton.com/`

files[51] which were assigned to eight sound objects in the A-Frame scene. Special care was taken to make sure each sound object was equidistant from the origin in order to preserve the balance of each song. We decided to use a small number of possible sonic elements to reduce loading times, as this is a pervasive problem with WebXR works.

In addition to the spatial audio present in each scene, we also wished to create some visual material to keep the listener engaged. To this end, we adopted another A-Frame component capable of creating visualizations from the sound spectra of musical elements[52]. A secondary mix, called a full mix, was piped to these visualization objects which react to a rendering of the musical material containing all the elements of the track. Furthermore, we chose to include glTF models and simple animations, based on A-Frame primitives, in the scene, to further accentuate and stimulate the audience.

Affordances and Limitations of WebXR

In contrast to native mobile XR applications built using iOS or Android XR SDKs, web applications require a stable and resilient network to operate smoothly. In contrast, standalone mobile applications, once packaged and downloaded, can be executed without having to rely on an internet connection or concerns about dropouts in connectivity. Another related format for XR experiences is the standalone application built using a game engine such as Unity or Unreal Engine. While these formats are generally more powerful, they also require specialized equipment and are considerably less accessible than mobile devices, which with the birth of the iPhone have become a ubiquitous technology around the world.

WebXR spatial audio technologies themselves, as noted by [MvTGBKH21] each have their affordances and limitations. McArthur et al. note three important criteria

[51] From mix busses.

[52] `https://www.npmjs.com/package/aframe-audio-analyser`

for the adoption of these tools: usefulness, ease of use, and documentation. Usefulness is defined by the authors as the number of barriers that the tool overcomes and the range of use cases. WebXR technologies intend to overcome a number of barriers which it seems in their current state they cannot fully satisfy. While the goal of these libraries is to create a system that can perform equally on mobile, desktop, and HMDs, the reality is that all these platforms yield vastly different results.

Using Continuum as a test case, we found that Safari, both on mobile and desktop did not behave as desired. Furthermore, in previous experiments with Omnitone[53], we found that iOS as a whole was unsuitable for distributing ambisonic audio via the browser. Likewise, with desktop browsers, our experience has been that security measures and API (e.g., application programming interface) implementations vary from browser to browser, making only some browsers compatible with the project. More so, these projects often require a large amount of Random Access Memory (RAM) and have distinct buffering mechanisms from project to project[54], making the performance heavily hardware-dependent.

Binaural audio developments in the last decades have made it possible for people to experience spatial music with considerably less hardware in a personalized setting, however, the audio quality when compared to speaker arrays is substantially reduced. Many of the SDKs created for spatialization in Unity, JS, or Virtual Studio Technologies (VSTs) rely on generic head-related transfer functions (HRTFs) which may take the listener time to adapt to. While HRTFs are becoming easier to measure, there still does not seem to be wider agreement regarding how formats such as SOFA [MZB+22] can be incorporated into WebXR technologies.

Unfortunately in the last few years, support for mobile HMDs such as the Google Cardboard or Daydream has been waning, as there is a greater financial incentive for companies to sell specialized equipment than to sell peripherals that can turn

[53] `https://github.com/GoogleChrome/omnitone`

[54] Some use adaptive streaming, while others download all assets prior to playback.

smartphones into capable virtual reality hardware. For most consumers, VR is still a luxury rather than a tool for work, which means a very small percentage of the population has adopted this technology. While most WebXR libraries support HMDs such as Vive or Oculus, the majority of our audience we suspect will explore the work on a desktop computer, while a small number will experience it with a Cardboard viewer.

Ethical Statement

While this project makes a great effort to address some of the ethical concerns regarding XR development from a socio-economic perspective, the piece is not without its faults. One of the primary concerns with this type of art-making is that it helps us further dissociate from our environment, which makes it easier to ignore the catastrophic changes that are being felt globally. The only defense to this claim is that the beauty of the virtual environment, albeit synthetic, attempts to inspire a sense of wonder through music, which we hope can be put to positive use in the real world. Furthermore, the elements comprising this work are multi-functional and have a relatively small footprint (e.g., the materials are relatively environmentally friendly).

While we seek to create methods for developing spatial music with low-cost hardware, the reality is that many of the systems used to create this piece are still prohibitively expensive to large swaths of the global population. Graphical processing units (GPUs) are heavily relied upon to process all the animations and models that make these works captivating, and while these have been getting cheaper, they remain out of reach for many people. Furthermore, while smartphones are fairly accessible devices, these experiences require state-of-the-art models with fast processors and advanced connectivity hardware.

4.5.3 DOAR (Bits)

DOAR is a virtual experience seeking to comment on our human desire for digital hoarding. We tend to think of data as abstract and therefore detached from the physical systems that maintain and support it. Unfortunately, there is a large cost associated with expanding and preserving these interconnected networks of information, not just on our bodies, but also on our planet. While the environmental cost might be easy to consider, the cognitive toll this excess of information has on our psyche is harder to quantify. DOAR (or Bits) seeks to connote how this seemingly boundless domain of information can cause the mind to become paralyzed. The experience is a paradox in that it uses vast amounts of information to make a statement about digital overindulgence; relative to the total amount of data currently being handled around the world, however, the information requirements of this project are negligible.

Evolution

This piece has had many different names over the years. The first time I made something similar to this was in NY in 2016 during Luke Dubois's class called MPIP (Multichannel Media Installation and Performance).

I worked on that project with an Iranian student (Makan Taghavi) and we used Cage's *Williams Mix* as inspiration. I remember hearing about the piece at UCSD when Tom Erbe showed us his realization of the work. This MPIP version used 700 samples manually gathered from Freesound.org and just a few images also manually gathered from online sources. The images were used as textures for various Jitter geometries.

A few years later I took another MAX class with Natascha Diels in 2018. That's where I started working on DOAR a patch that was a response to Poème Électronique. I wrote a MAX patch that used a large number of audio samples and a variety of

images, much like the original piece (e.g., Varese's work). I also incorporated other synthesis methods, some modern visual elements, and spatial audio. The piece is called Death of a Robot (DOAR) because it looks like what one might imagine a robot sees when it is dying.

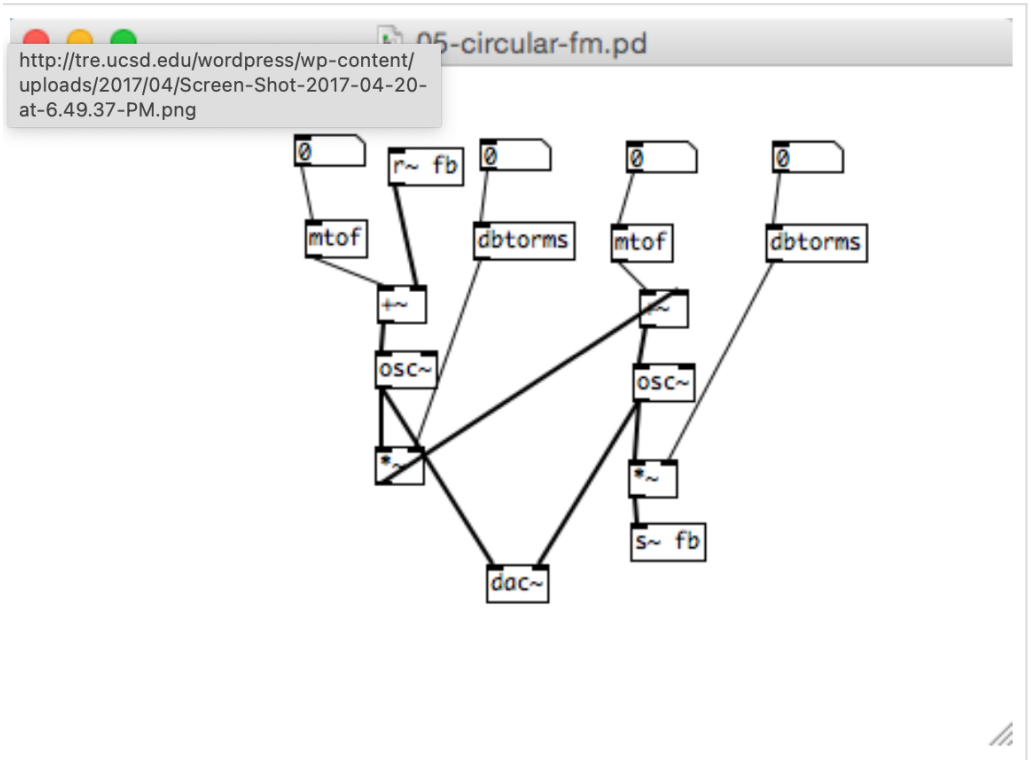


Figure 4.17: Feedback FM Patch by Tom Erbe in Pd

The final patch has many images being quickly changed (one every 100 ms) but in addition, it uses a particular shader to make the video look a lot more "glitchy". In addition to sampling many sounds from the FSDKaggle dataset, I also applied a feedback effect to these sounds and used the ICST ambisonic package in case I wanted to use multichannel audio in the future. At that point I decided I wanted to add some synthesized sound – to compliment the sampled sound – so I opted for the circular FM patch that Tom Erbe showed us in Spring 2019 during his graduate MAX course. In the patch, I change the values of three of the four parameters of the

algorithm every 100 ms to synchronize with the video. There is a third voice which is shaped by noise oscillating at different speeds which use [lores~] and a final voice which is simply a triangle wave that drones on throughout, intermittently changing frequency to create some movement. The four-voice instrumentation was inspired by previous counterpoint courses which introduced me to SATB-style[55] compositional approaches.

Later, I decided to move away from MAX and instead worked out more of the piece using FOSS. The sequencing of images was already done in MAX, but the new audio material is done in Pd. The VBAP external was used to create 8 channels that simulate an octophonic system (like the one in CPMC122 or CPMC365). The audio was then imported into an A-Frame scene and we used the native WAA (Web Audio API) methods to spatialize the signals. We used a multi-video component to create a grid of videos that were offset in their start time. This was inspired by the large multi-tile displays featured at Calit2. In order to reduce the amount of data we use FFMPEG to convert our 10-minute WAV files into 3-minute MP3 files and Handbrake to convert our 10-minute MPEG4 into a 3-minute MPEG4 with lower resolution (720p). Both of these tools are free and open-source.

Virtual Installation

The experience/virtual installation was then submitted to EDAM organized by IUPUI (Indiana University). The call specified live performances, presentations, or videos but I wanted to share true dynamic spatial audio, not binaural audio. Given the nature of the call, I needed to reframe the piece to fit the theme of the event. To that end, I decided to use it as an opportunity to talk about the ecological impact that "big data" has on our climate.

In this piece there is not a real formal attempt at generating direct symbiosis between the audio and the visual, however, the aesthetic principles applied to the

[55]Soprano, Alto, Tenor, Bass.

generation of one are also applied to the other, creating definite congruence. Namely, the approach to generating the video, and sound, was to sequence a large number of images and sound samples, in a rapid fashion. Additionally, we create the synchronization of audio and sound events via the use of a single metronome speed, further increasing the congruence between both mediums.

Another feature of the work that provides a direct link between the aural and visual is the use of visual landmarks to symbolize virtual speakers. These virtual speaker representations are included in an effort to facilitate the localization of sound sources which has been shown to be a bi-modal task. Naturally, there is a clear interaction between the movement of the user, which dictates what they see, and the timbre of the sound events, which are filtered to simulate a real acoustic environment.

In our piece, we sought to expand on the theme of scale by repurposing a dataset meant for machine learning applications into a musical oeuvre. The work uses 10k sound samples and hundreds of images as well as 3D audio via the Web Audio API. A command line tool called Bulksplash was used to scrape images with high resolution from an online repository. This allowed us to download about 4 GB of images (400 JPGs) in just a few minutes. The categories used for searching the database can be seen in Figure 4.18. The FSDKaggle set also has about 4GB which ends up being a little under 10,000 samples. On the scale of real Big Data problems or data science projects, this is in fact conservative. As an informed society, we are well aware that large companies operate thousands of servers each with thousands of terabytes of storage. The question is: *is bigger really better?*

4.6 Conclusion

This chapter has explored the viability of WebXR as a tool for the dissemination of spatial audio on the web. In Section 4.1 we introduced the reader to a brief history of virtual reality. Section 4.4 talked about the various JS libraries in the open-

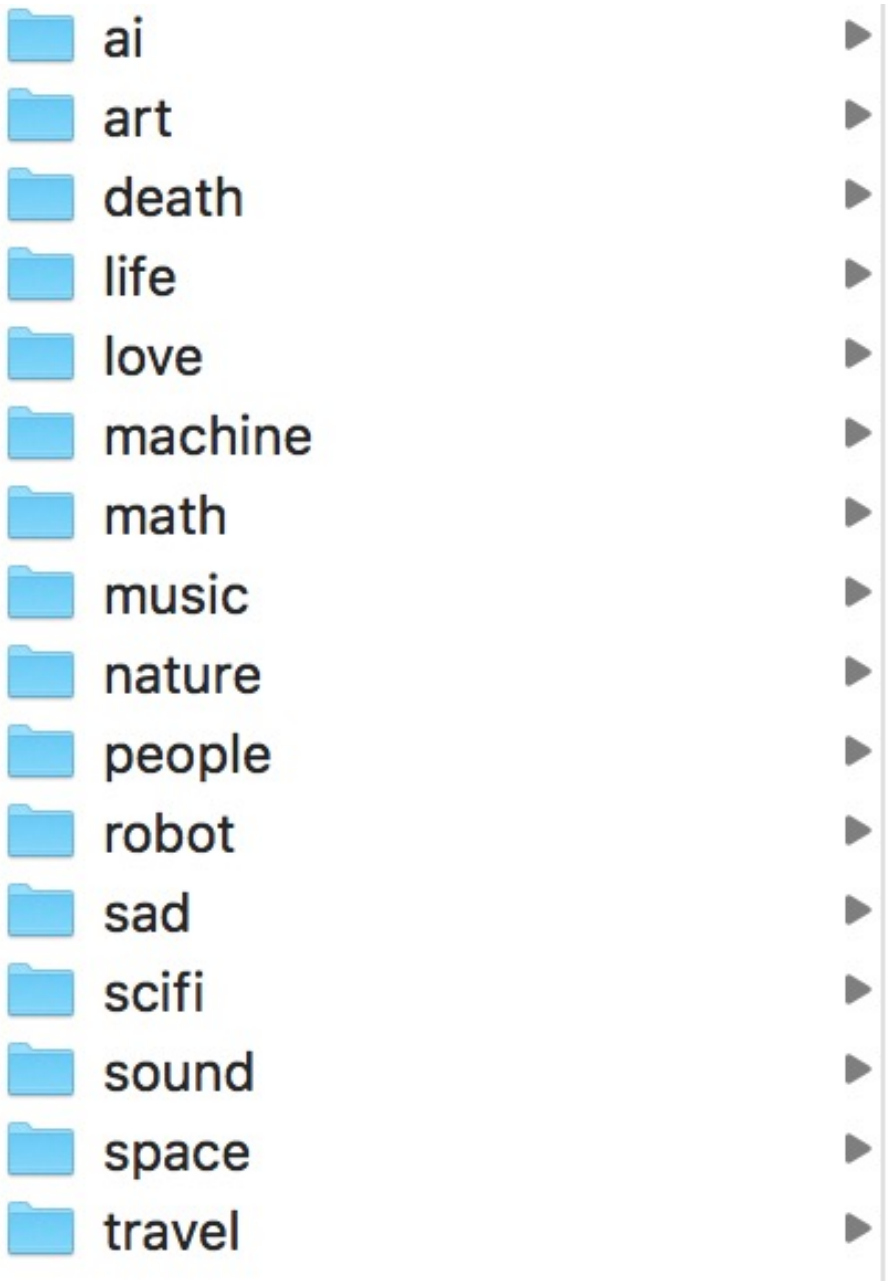


Figure 4.18: Categories used in Bulksplash CLI

source domain which we can leverage for this task. And, finally, in Section 4.5 we demonstrated the viability of these tools by presenting three different compositions

created with three different libraries.

As you have seen, there are several tools in the open-source domain that enable the binaural rendition of experiences in three and six degrees of freedom. Unfortunately, although spatial audio is possible on the web, there are many limitations that make its delivery complicated. The bandwidth, memory, and performance problems arise from unsupported formats or incompatible devices. That being said, smaller projects - featuring low-polygon models, seem to perform fine so long as data files stay small.

Whether elements are dynamically streamed or not, changes the nature of the delivery system entirely. Libraries like HOAST [DMKH$^+$20] specifically try to perform adaptive streaming to increase accessibility. HMDs are starting to become part of everyday life. Certain design problems are perhaps simply better approached in a virtual space. We believe, despite some of the obvious limitations of WebXR as a 3D audio distribution system, it is the most accessible option allowing anybody to share this type of music publicly.

Chapter 5

Conclusion

The motivation of this dissertation has been to demonstrate the possibility for marginalized communities to exploit FOSS towards creating and distributing spatial music. In the introductory chapter, we framed the motivation for this work by presenting a brief history of the domain, which featured composers such as Edgard Varèse and John Cage, in the electro-acoustic domain, and Henry Brant, in the acoustic domain. We gather from this brief history that space as a predominant feature of the compositional process has been in existence for hundreds of years, and has grown in importance over the recent decades. The introductory chapter also offered a brief account of the psychoacoustic principles that modern spatial audio systems capitalize on to deliver immersive musical experiences.

Chapter 2 explored the concept and principles of ambisonic encoding[1], decoding, and presented a collection of selected works, two of which are original works, and one of which is a collaboration with UC San Diego faculty (Lei Liang). All of these works were presented in formal academic and professional settings including conferences held internationally or at paid functions at our University. This chapter serves to provide a background for later chapters in which ambisonics is

[1] Also known as ambisonic synthesis.

also the main technology addressed. Section 2.2 outlines the general mathematics and concepts pertinent to the technology and provides one method of creating ambisonic audio: synthesizing it. Section 2.3 delineated the fundamental principles of ambisonic decoding which entails using a linear combination of spherical harmonics (e.g., B-format ambisonic signals) to generate speaker feeds. This isotropic manner of projecting sound in space represents one of the key features of ambisonic which makes it attractive for low-cost spatial music since playback can be adjusted for any number of speakers. The chapter culminates with descriptions of three selected works that employ this technique. The general principles applicable to physical ambisonic decoding over real speakers also elucidate the methodology and ideas pertinent to Section 4.4, which addresses binaural audio in the context of WebXR.

Chapter 3 addressed the topic of ambisonic recording via microphone arrays - specifically the development of a FOA and HOA microphone array. Section 3.2, the literature review, discusses four projects by other authors that have greatly informed and inspired the fabrication of our prototypes. Section 3.3 discusses early experiments by the author in the fabrication and objective evaluation of a FOA microphone array. The results of this research were published in a recent publication by the author in the AES [Zal19]. The last section of this chapter addresses the development and evaluation of a HOA microphone array. Much like our early prototypes this new design also features MEMS which allows us to record many channels directly with a low-cost MCU. These sections also describe the process of evaluating and calibrating the system, which entails acquiring a large set of IRs using an automated system [2]. This chapter offers an alternative methodology for creating spatial music, specifically attractive for acoustic composers who want to naturally record live performances.

Finally, Chapter 4 covered the concept of delivering spatial music using WebXR technologies. Section 4.1 discussed the general idea of WebVR, including a short description of the technology, its history, and its affordances and limitations. Sec-

[2]Currently powered by ScanIR [VGR19].

tion 4.4 talked about the methods used in contemporary binaural rendering systems which are commonly applied to WebXR experiences to provide immersive auditory experiences. Finally, Section 4.5 covered three WebXR projects executed by the author that exploit these systems to provide spatial music on the browser. Two of these works have been created in partnership with various undergraduate students at UCSD as part of a larger initiative to create community and fellowship between different academic levels at our University[3]. All of these works have been submitted in calls for music such as EDAM, NIME, or DAVAMOT. This chapter demonstrates how WebXR technologies can be adopted by artists of underprivileged backgrounds to create spatial music that can be experienced globally.

The aim of this dissertation has been to create a roadmap for computer music artists and researchers interested in ambisonic to navigate the complex and rich world of open-source technologies enabling the creation, dissemination, and documentation of spatial music. This framework we believe is especially relevant to marginalized communities which lack the infrastructure and resources to produce these rich sonic forms. In the context of pedagogy within these communities, this document also serves as the basis for course development, designed consciously using principles of equity, diversity, and inclusion. Our hope is that this dissertation can serve as the foundation for 21st-century arts and sciences courses in less-than-prosperous populations, further narrowing the expanding divide between those who have and have not.

[3]For more information see UCSD SElectOr

Appendix A

Bessel Functions

A.1 Cylindrical Bessel Functions of the First Kind

Bessel functions arise when finding solutions to Laplace's equation and the Helmholtz equation in cylindrical or spherical coordinates and are important in problems of wave propagation.

When n is an integer, these functions can be calculated by [oM]:

$$J_n(z) = \left(\frac{z}{2}\right)^n \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(n+k)!} \left(\frac{z}{2}\right)^{2k} \tag{A.1}$$

where,

n is the ambisonic order, and

z is a vector of frequency values.

However, spherical Bessel functions are easier to define as a function of half-integer order cylindrical Bessel functions. Luckily, the first two half-integer order cylindrical Bessel functions are elementary. They are given by:

$$J_{\frac{1}{2}}(z) = \sqrt{\frac{2}{\pi z}} \sin(z) \tag{A.2}$$

and,

$$J_{-\frac{1}{2}}(z) = \sqrt{\frac{2}{\pi x}}\cos(z) \tag{A.3}$$

using the recurrence formula:

$$J_{\nu-1}(z) + J_{\nu+1}(z) = \frac{2\nu}{z}J_{\nu}(z) \tag{A.4}$$

we can calculate other half-integer order cylindrical Bessel functions. For example, for $v = 1/2$:

$$\sqrt{\frac{2}{\pi \cdot z}}\cos(z) + J_{v+1}(z) = \frac{2(1/2)}{z}\left(\sqrt{\frac{2}{\pi \cdot z}}\sin(z)\right) \tag{A.5}$$

Isolating $J_{v+1}(z)$ gives:

$$J_{3/2}(z) = \frac{1}{z}\left(\sqrt{\frac{2}{\pi z}}\sin(z)\right) - \sqrt{\frac{2}{\pi z}}\cos(z). \tag{A.6}$$

Which further simplifies to:

$$J_{3/2}(z) = \frac{1}{z}\cdot\sqrt{\frac{2}{\pi}}\cdot\sqrt{\frac{1}{z}}\cdot\sin(z) - \sqrt{\frac{1}{z}}\cdot\sqrt{\frac{2}{\pi}}\cdot\cos(z). \tag{A.7}$$

Finally we obtain:

$$J_{3/2}(z) = \sqrt{\frac{2}{\pi}}\cdot\left(\frac{1}{z^{3/2}}\sin(z) - \frac{1}{z^{1/2}}\cos(z)\right). \tag{A.8}$$

Additional properties of Bessel functions can be used to find negative half-integer order solutions. Various authors use built-in functions to calculate Cylindrical Bessel functions. They can be found in MATLAB, for example.

A.2 Cylindrical Bessel Functions of the Second Kind

Provided the Cylindrical Bessel Functions of the First Kind have already been calculated, the Cylindrical Bessel Functions of the Second Kind can be calculated using:

$$Y_\nu(z) = \frac{J_\nu(z)\cos\nu\pi - J_{-\nu}(z)}{\sin\nu\pi} \quad \text{for } \nu \notin \mathbb{Z} \tag{A.9}$$

This equation is valid for values of v that are not part of the set $\mathbb{Z}$, which corresponds to real integers. We can use these two definitions to derive our Spherical Bessel Functions of the First and Second kind, as well as the Spherical Hankel functions of the First and Second Kind.

A.3 Spherical Bessel Functions

To calculate our Sphericak Bessel Functions we use

$$\begin{aligned} j_n(z) &= \sqrt{\frac{\pi}{2z}} J_{n+1/2}(z) \\ y_n(z) &= \sqrt{\frac{\pi}{2z}} Y_{n+1/2}(z) \\ h_n^{(1)}(z) &= j_n(z) + iy_n(z) = \sqrt{\frac{\pi}{2z}} \left[J_{n+1/2}(z) + iY_{n+1/2}(z)\right] \\ h_n^{(2)}(z) &= j_n(z) - iy_n(z) = \sqrt{\frac{\pi}{2z}} \left[J_{n+1/2}(z) - iY_{n+1/2}(z)\right] \end{aligned} \tag{A.10}$$

$j_n(z)$ and $y_n(z)$ represent the spherical Bessel functions of the first and second kind respectively. $h_n^{(1)}(z)$ and $h_n^{(2)}(z)$ define the Hankel functions of the first and second kind respectively.

Alternate solutions to the four spherical Bessel functions can also be found in the Appendix of [ZF19]. Implementations of these can also be found in Politis's "Array-Response-Simulator"[1] or the SOFiA toolbox used in Moschner et al.[MDLP20]. In our research we Bessel functions appeared when discussing Near-Field Compensation (NFC), Radial Filter Design for HOA encoding, and a particular VST application (e.g. Kronlachner's widening algorithm).

A.4 Practical Spherical Bessel Functions

Generally in periphonic ambisonic research we are interested in spherical Bessel functions of the first and second kind, which we can use to calculate Hankel functions applicable to NFC and Radial filter design. Here we give the spherical Bessel functions up to 3OA, using simple trigonometric functions, which we find suitable for most recording and reproduction purposes:

$$\begin{aligned} j_0(x) &= \tfrac{\sin x}{x} \\ j_1(x) &= \tfrac{\sin x}{x^2} - \tfrac{\cos x}{x} \\ j_2(x) &= \left(\tfrac{3}{x^2} - 1\right) \tfrac{\sin x}{x} - \tfrac{3\cos x}{x^2}, \\ j_3(x) &= \left(\tfrac{15}{x^3} - \tfrac{6}{x}\right) \tfrac{\sin x}{x} - \left(\tfrac{15}{x^2} - 1\right) \tfrac{\cos x}{x} \end{aligned} \tag{A.11}$$

The relationships above are for Spherical Bessel Functions of the First Kind.

$$\begin{aligned} y_0(x) &= -j_{-1}(x) = -\tfrac{\cos x}{x} \\ y_1(x) &= j_{-2}(x) = -\tfrac{\cos x}{x^2} - \tfrac{\sin x}{x}, \\ y_2(x) &= -j_{-3}(x) = \left(-\tfrac{3}{x^2} + 1\right) \tfrac{\cos x}{x} - \tfrac{3\sin x}{x^2}, \\ y_3(x) &= j_{-4}(x) = \left(-\tfrac{15}{x^3} + \tfrac{6}{x}\right) \tfrac{\cos x}{x} - \left(\tfrac{15}{x^2} - 1\right) \tfrac{\sin x}{x} \end{aligned} \tag{A.12}$$

While the second equation refers to Spherical Bessel Functions of the Second Kind.

[1] Found here (Accessed: May 14, 2021).

Bibliography

[20] Valerio Saggini . Alexandra stepanoff, Sep 2020.

[aEW] Andyj at English Wikipedia. File:how cinerama is projected.gif - wikimedia commons. `https://commons.wikimedia.org/wiki/File:How_Cinerama_is_projected.gif`. (Accessed on 04/08/2021).

[Arc] Politis Archontis. Spherical harmonic transform library. `http://research.spa.aalto.fi/projects/sht-lib/sht.html`. (Accessed on 04/23/2021).

[Arn59] Denis Arnold. The significance of" cori spezzati". *Music & Letters*, pages 4–14, 1959.

[Ben12] Eric M Benjamin. A second-order soundfield microphone with improved polar pattern shape. In *Audio Engineering Society Convention 133*. Audio Engineering Society, 2012.

[Ber12] Cyrille-Paul Bertrand. The pepsi-cola pavilion, osaka world's fair, 1970. 2012.

[Bla97] Jens Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT press, 1997.

[Bla11] Manuella Blackburn. The visual sound-shapes of spectromorphology: an illustrative guide to composition. *Organised Sound*, 16(1):5–13, 2011.

[BPSW11] Benjamin Bernschütz, Christoph Pörschmann, Sascha Spors, and Stefan Weinzierl. Sofia sound field analysis toolbox. In *Proceedings*

of the International Conference on Spatial Audio (ICSA), pages 7–15, 2011.

[Cag61] John Cage. Experimental music. *Silence: Lectures and writings*, 7:12, 1961.

[CDCJ18] Elliot K Canfield-Dafilou, Eoin Callery, and Christopher Jette. A portable impulse response measurement system. In *15th Sound and Music Computing Conference*, pages 172–176, 2018.

[ÇH20] Cem Çakmak and Rob Hamilton. od: Composing spatial multimedia for the web. *Journal of the Audio Engineering Society*, 68(10):747–755, 2020.

[Cho71] John M Chowning. The simulation of moving sound sources. *Journal of the audio engineering society*, 19(1):2–6, 1971.

[Cle] Patrick Clenet. Ven basil matin 082005 - basilica di san marco - wikimedia commons. `https://commons.wikimedia.org/wiki/Basilica_di_San_Marco#/media/File:Ven_basil_matin_082005.JPG`. (Accessed on 01/19/2021).

[Coc] Alex Cochrane. File:playstation dualsense controller.png - wikimedia commons. `https://commons.wikimedia.org/wiki/File:Playstation_DualSense_Controller.png`. (Accessed on 02/16/2021).

[Con] Conversion to minimum phase. `https://ccrma.stanford.edu/~jos/filters/Conversion_Minimum_Phase.html`. (Accessed on 05/06/2021).

[CRPGT+19] María Cuevas-Rodríguez, Lorenzo Picinali, Daniel González-Toledo, Carlos Garre, Ernesto de la Rubia-Cuestas, Luis Molina-Tanco, and Arcadio Reyes-Lecuona. 3d tune-in toolkit: An open-source library for real-time binaural spatialisation. *PloS one*, 14(3):e0211899, 2019.

[Dav01] Davepape. File:cave crayoland.jpg - wikimedia commons. `https://commons.wikimedia.org/wiki/File:CAVE_Crayoland.jpg`, August 2001. (Accessed on 04/08/2021).

[Dix06] Steve Dixon. A history of virtual reality in performance. *International Journal of Performance Arts & Digital Media*, 2(1), 2006.

[DLAP19] Damian T Dziwis, Tim Lübeck, Johannes M Arend, and Christoph Pörschmann. Development of a 7th order spherical microphone array for spatial audio recording. *45. Deutsche Jahrestagung für Akustik*, 2019.

[DM04] Jérôme Daniel and Sébastien Moreau. Further study of sound field coding with higher order ambisonics. In *Audio Engineering Society Convention 116*. Audio Engineering Society, 2004.

[DMKH+20] Thomas Deppisch, Nils Meyer-Kahlen, Benjamin Hofer, Tomasz Latka, and Tomasz Zernicki. Hoast: A higher-order ambisonics streaming platform. In *Audio Engineering Society Convention 148*. Audio Engineering Society, 2020.

[EA] Evan-Amos. File:sony-playstation-move-controller.png - wikimedia commons. `https://commons.wikimedia.org/wiki/File:Sony-PlayStation-Move-Controller.png`. (Accessed on 02/16/2021).

[Ein48] Alfred Einstein. Music in the romantic era. 1948.

[Ele91] Richard Elen. Whatever happened to ambisonics. *AudioMedia Magazine, Nov*, 1991.

[Erz] Erzalibillas. File:entrega de las llaves a san pedro (perugino).jpg - wikipedia. `https://en.wikipedia.org/wiki/File:Entrega_de_las_llaves_a_San_Pedro_(Perugino).jpg`. (Accessed on 02/09/2021).

[Fila] File:ibm anechoic chamber.jpg - wikimedia commons. `https://commons.wikimedia.org/wiki/File:IBM_Anechoic_chamber.jpg`. (Accessed on 01/19/2021).

[Filb] File:igb 006055 visore stereoscopico portatile museo scienza e tecnologia milano.jpg - wikipedia. `https://en.wikipedia.org/wiki/File:IGB_006055_Visore_stereoscopico_portatile_Museo_scienza_e_tecnologia_Milano.jpg`. (Accessed on 01/24/2021).

[Filc] File:psconcer.jpg - wikimedia commons. `https://commons.wikimedia.org/wiki/File:Psconcer.jpg`. (Accessed on 01/21/2021).

[Fild] File:types of biwa, japanese traditional instrument.jpg - wikimedia commons. https://commons.wikimedia.org/wiki/File:Types_of_Biwa,_Japanese_traditional_instrument.jpg. (Accessed on 01/19/2021).

[FM99] Jörg Fliege and Ulrike Maier. The distribution of points on the sphere and corresponding cubature formulae. *IMA Journal of Numerical Analysis*, 19(2):317–334, 1999.

[GAK+19] Marcin Gorzel, Andrew Allen, Ian Kelly, Julius Kammerl, Alper Gungormusler, Hengchin Yeh, and Francis Boland. Efficient encoding and decoding of binaural sound with resonance audio. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

[GPL18] Raimundo González, Joshua Pearce, and Tapio Lokki. Modular design for spherical microphone arrays. In *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2018.

[Guy] Martin Guy. File:ircam 4x.jpg - wikimedia commons. https://commons.wikimedia.org/wiki/File:IRCAM_4X.jpg. (Accessed on 01/21/2021).

[HA17] Christoph Hohnerlein and Jens Ahrens. Spherical microphone array processing in python with the sound field analysis-py toolbox. *Fortschritte der Akustik–DAGA 2017*, pages 1033–1036, 2017.

[Har97] Maria Anna Harley. An american in space: Henry brant's" spatial music". *American Music*, pages 70–92, 1997.

[HDSC+17] Huseyin Hacihabiboglu, Enzo De Sena, Zoran Cvetkovic, James Johnston, and Julius O Smith III. Perceptual spatial audio recording, simulation, and rendering: An overview of spatial-audio techniques based on psychoacoustics. *IEEE Signal Processing Magazine*, 34(3):36–54, 2017.

[Hei61] Morton Heilig. File:sensorama patent fig5.png - wikimedia commons. https://commons.wikimedia.org/wiki/File:Sensorama_patent_fig5.png, Jan 1961. (Accessed on 04/08/2021).

[HF20] Matteus Hemström and Anton Forsberg. A comparison of webvr and native vr: Impacts on performance and user experience, 2020.

[HLB08] Aaron Heller, Richard Lee, and Eric Benjamin. Is my decoder ambisonic? In *Audio Engineering Society Convention 125*. Audio Engineering Society, 2008.

[Ian] Iannis xenakis - electronic works 2 - hibiki hana ma; polytope de cluny (2008, cd) — discogs. `https://www.discogs.com/Iannis-Xenakis-Electronic-Works-2-Hibiki-Hana-Ma-Polytope-De-Cluny/release/1557104`. (Accessed on 01/14/2021).

[Ini14] Inigo.quilez. Spherical harmonics - wikipedia. `https://en.wikipedia.org/wiki/Spherical_harmonics`, 14 May 2014. (Accessed on 02/19/2021).

[JAHS21] Jean-Marc Jot, Rémi Audfray, Mark Hertensteiner, and Brian Schmidt. Rendering spatial sound for interoperable experiences in the audio metaverse. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pages 1–15. IEEE, 2021.

[JEP13] Craig T Jin, Nicolas Epain, and Abhaya Parthy. Design, optimization and evaluation of a dual-radius spherical microphone array. *IEEE/ACM transactions on audio, speech, and language processing*, 22(1):193–204, 2013.

[Joh19] Mathias Johansson. Vr for your ears: Dynamic 3d audio is key to the immersive experience by mathias johansson· illustration by eddie guy. *IEEE spectrum*, 56(2):24–29, 2019.

[Jon91] Mark Jones. 20th century composers. *Psychiatric Bulletin*, 15(7):442–445, 1991.

[Kal] Kaldari. File:the horse in motion high res.jpg - wikipedia. `https://en.wikipedia.org/wiki/File:The_Horse_in_Motion_high_res.jpg`. (Accessed on 02/09/2021).

[KD15] Gavin Kearney and Tony Doyle. An hrtf database for virtual loudspeaker rendering. In *Audio Engineering Society Convention 139*. Audio Engineering Society, 2015.

[Ken95] Gary S Kendall. A 3-d sound primer: directional hearing and stereo reproduction. *Computer music journal*, 19(4):23–46, 1995.

[Kro14] Matthias Kronlachner. Spatial transformations for the alteration of ambisonic recordings. *M. Thesis, University of Music and Performing Arts, Graz, Institute of Electronic Music and Acoustics*, 7, 2014.

[LaV16] Steven LaValle. Virtual reality. 2016.

[LL16] Fernando Lopez-Lezcano. The* sphear project, a family of parametric 3d printed soundfield microphone arrays. In *Audio Engineering Society Conference on Soundfield Control*, 2016.

[LL18] Fernando Lopez-Lezcano. The* sphear project update: the tinysphear and octathingy soundfield microphones. In *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2018.

[LL19] Fernando Lopez-Lezcano. The sphear project update: Refining the octasphear, a 2nd order ambisonics microphone. 2019.

[LZ15] Stefan Lösler and Franz Zotter. Comprehensive radial filter design for practical higher-order ambisonic recording. *Fortschritte der Akustik, DAGA*, pages 452–455, 2015.

[Man13] Peter Manning. *Electronic and computer music*. Oxford University Press, 2013.

[MDB06] Sébastien Moreau, Jérôme Daniel, and Stéphanie Bertet. 3d sound field recording with higher order ambisonics–objective measurements and validation of a 4th order spherical microphone. In *120th Convention of the AES*, pages 20–23, 2006.

[MDLP20] Oscar Moschner, Damian Dziwis, Tim Lübeck, and Christoph Pörschmann. Development of an open source customizable high order rigid sphere microphone array. In *Audio Engineering Society Convention 148*. Audio Engineering Society, 2020.

[MDMF+18] Leo McCormack, Symeon Delikaris-Manias, Angelo Farina, Daniel Pinardi, and Ville Pulkki. Real-time conversion of sensor array signals into spherical harmonic signals with applications to spatially

localized sub-band sound-field analysis. In *Audio Engineering Society Convention 144*. Audio Engineering Society, 2018.

[MIC+13] Piotr Majdak, Yukio Iwaya, Thibaut Carpentier, Rozenn Nicol, Matthieu Parmentier, Agnieszka Roginska, Yôiti Suzuki, Kankji Watanabe, Hagen Wierstorf, Harald Ziegelwanger, et al. Spatially oriented format for acoustics: A data exchange format representing head-related transfer functions. In *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.

[MM95] David G Malham and Anthony Myatt. 3-d sound spatialization using ambisonic techniques. *Computer music journal*, 19(4):58–70, 1995.

[Moo82] F Richard Moore. The computer audio research laboratory at ucsd. *Computer Music Journal*, 6(1):18–29, 1982.

[Muh15] Muhanna A Muhanna. Virtual reality and the cave: Taxonomy, interaction challenges and research directions. *Journal of King Saud University-Computer and Information Sciences*, 27(3):344–361, 2015.

[MvTGBKH21] Angela McArthur, Cobi van Tonder, Leslie Gaston-Bird, and Andrew Knight-Hill. A survey of 3d audio through the browser: practitioner perspectives. In *2021 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, pages 1–10. IEEE, 2021.

[MW19] Charlie Middlicott and Bruce Wiggins. Calibration approaches for higher order ambisonic microphones. Audio Engineering Society, 2019.

[MZB+22] Piotr Majdak, Franz Zotter, Fabian Brinkmann, Julien De Muynke, Michael Mihocic, and Markus Noisternig. Spatially oriented format for acoustics 2.1: Introduction and recent advances. *Journal of the Audio Engineering Society*, 70(7/8):565–584, 2022.

[NAS06] NASA. File:head-mounted display and wired gloves, ames research center.jpg - wikimedia commons. `https://commons.wikimedia.org/wiki/File:Head-mounted_display_and_wired_gloves,_Ames_Research_Center.jpg`, May 2006. (Accessed on 04/08/2021).

[Net] Nettings. File:naive ambisonic square decoder example.png - wikimedia commons. https://commons.wikimedia.org/wiki/File:Naive_Ambisonic_Square_Decoder_Example.png. (Accessed on 02/18/2021).

[NZDS11] Christian Nachbar, Franz Zotter, Etienne Deleflie, and Alois Sontacchi. Ambix-a suggested ambisonics format. In *Ambisonics Symposium, Lexington*, page 11, 2011.

[O'c11] James O'callaghan. Soundscape elements in the music of denis smalley: negotiating the abstract and the mimetic. *Organised Sound*, 16(1):54–62, 2011.

[oM] Encyclopedia of Math. Bessel functions - encyclopedia of mathematics. https://encyclopediaofmath.org/wiki/Bessel_functions. (Accessed on 05/14/2021).

[P+16] Archontis Politis et al. Microphone array processing for parametric spatial audio techniques. 2016.

[Pap] Dave Pape. File:vpl eyephone and dataglove.jpg - wikimedia commons. https://commons.wikimedia.org/wiki/File:VPL_Eyephone_and_Dataglove.jpg. (Accessed on 02/15/2021).

[PLS12] Nils Peters, Trond Lossius, and Jan C Schacher. Spatdif: Principles, specification, and examples. In *9th Sound and Music Computing Conference, Copenhagen, Denmark*, 2012.

[PPQ16] Archontis Politis and David Poirier-Quinot. Jsambisonics: A web audio library for interactive spatial sound processing on the web. In *Interactive Audio Systems Symposium*, 2016.

[Puc01] Miller Puckette. New public-domain realizations of standard pieces for instruments and live electronics. In *ICMC*. Citeseer, 2001.

[Sel14] Zachary Seldess. Miap: manifold-interface amplitude panning in max/msp and pure data. In *Audio Engineering Society Convention 137*. Audio Engineering Society, 2014.

[SGK76] Naraji Sakamoto, Toshiyuki Gotoh, and Yoichi Kimura. On-out-of-head localization-in headphone listening. *Journal of the Audio Engineering Society*, 24(9):710–716, 1976.

[SH16] Dieter Schmalstieg and Tobias Hollerer. *Augmented reality: principles and practice*. Addison-Wesley Professional, 2016.

[Sha] SharkD. Spherical coordinate system - spherical coordinate system - wikipedia. `https://en.wikipedia.org/wiki/Spherical_coordinate_system#/media/File:Spherical_coordinate_system.svg`. (Accessed on 02/19/2021).

[Sin] Singular value decomposition (svd) and pseudoinverse. `https://www.johndcook.com/blog/2018/05/05/svd/`. (Accessed on 05/06/2021).

[Sma97] Denis Smalley. Spectromorphology: explaining sound-shapes. *Organised sound*, 2(2):107–126, 1997.

[Sto96] Karlheinz Stockhausen. Helikopter-streichquartett. *Grand Street*, (56):213–225, 1996.

[STV+22] Anderson Augusto Simiscuka, Mohammed Amine Togou, Rohit Verma, Mikel Zorrilla, Noel E O'Connor, and Gabriel-Miro Muntean. An evaluation of 360° video and audio quality in an artistic-oriented platform. In *2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pages 1–5. IEEE, 2022.

[Sut65] Ivan Sutherland. The ultimate display. 1965.

[SZH18] Christian Schörkhuber, Markus Zaunschirm, and Robert Höldrich. Binaural rendering of ambisonic signals via magnitude least squares. *Fortschritte der Akustik–DAGA*, pages 339–342, 2018.

[Teu07] Heinz Teutsch. *Modal array signal processing: principles and applications of acoustic wavefield decomposition*, volume 348. Springer, 2007.

[The] Theramin-alexandra-stepanoff-1930 - theremin - wikipedia. `https://en.wikipedia.org/wiki/Theremin#/media/File:Theramin-Alexandra-Stepanoff-1930.jpg`. (Accessed on 01/14/2021).

[Ven14] Jakob Vennerød. Binaural reproduction of higher order ambisonics-a real-time implementation and perceptual improvements. Master's thesis, NTNU, 2014.

[VGR19] Julian Vanasse, Andrea Genovese, and Agnieszka Roginska. Multichannel impulse response measurements in matlab: An update on scanir. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, 2019.

[VTL10] Andrea Valle, Kees Tazelaar, and Vincenzo Lombardo. In a concrete space: Reconstructing the spatialization of iannis xenakis' concret ph on a multichannel setup. In *Proceedings of the Sound and Music Computing Conference (SMC-2010)*, 2010.

[Wei] Eric W. Weisstein. Associated legendre polynomial – from wolfram mathworld. `https://mathworld.wolfram.com/AssociatedLegendrePolynomial.html`. (Accessed on 02/19/2021).

[wik20a] Expo 58, Dec 2020.

[wik20b] Graphic notation (music), Dec 2020.

[Wri05] Matthew Wright. Open sound control: an enabling technology for musical networking. *Organised Sound*, 10(3):193–200, 2005.

[Zal19] Gabriel Zalles. Effects of capsule coincidence in foa using mems: Objective experiment. In *Audio Engineering Society Convention 147*. Audio Engineering Society, 2019.

[ZBB05] Pavel Zahorik, Douglas S Brungart, and Adelbert W Bronkhorst. Auditory distance perception in humans: A summary of past and present research. *ACTA Acustica united with Acustica*, 91(3):409–420, 2005.

[ZF12] Franz Zotter and Matthias Frank. All-round ambisonic panning and decoding. *Journal of the audio engineering society*, 60(10):807–820, 2012.

[ZF19] Franz Zotter and Matthias Frank. *Ambisonics: A practical 3D audio theory for recording, studio production, sound reinforcement, and virtual reality*. Springer Nature, 2019.

[ZPN12] Franz Zotter, Hannes Pomberger, and Markus Noisternig. Energy-preserving ambisonic decoding. *Acta Acustica united with Acustica*, 98(1):37–47, 2012.

[Zvo99] Richard Zvonar. A history of spatial music. *Montreal: CEC*, 1999.

[Zvo00] Richard Zvonar. An extremely brief history of spatial music in the 20th century. *Surround Professional Magazine*, 296, 2000.