The Design of a Highly Coincident Microphone Array for Stereo and Surround Sound

Gabriel Zalles

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

Advisor: Agnieszka Rogisnka
Reader: Juan Bello

May 14, 2018

Dedicated to my family.

ACKNOWLEDGMENTS

I wish to gratefully acknowledge my thesis committee for their insightful comments and guidance. I also want to specifically Charlie Mydlarz, Yigal Kamel and all the other students who were part of my journey.

PREFACE

This thesis regards the evaluation and analysis of a microphone capable of recording spatial audio which can be used to create realistic auditory experiences in virtual reality (VR). With advancements in silicone chip manufacturing and general optimization in hardware design, VR has slowly started to become a ubiquitous way for almost anyone to enjoy: journalistic pieces, musical content, games, or films, in a truly immersive manner. Spatial audio microphones, such as the one described in this thesis, are especially attractive due to their ability to seamlessly, and naturally, encode source location information without relying on meta data associated to audio objects. This method of encoding audio information is especially appealing in specific contexts where there are a very large number of audio sources or when it is unfeasible to set up a large number of microphones, due to the topological features of the location were these sound sources are located. Additionally, microphones of this type allow recording engineers to pick from a variety of recording techniques after recording including techniques used to develop high-quality surround sound mixes.

TABLE OF CONTENTS

Page

LIST OF TABLES . . . viii

LIST OF FIGURES . . . ix

ABBREVIATIONS . . . xi

GLOSSARY . . . xii

ABSTRACT . . . xiv

CHAPTER 1. INTRODUCTION . . . 1
- 1.1 Scope . . . 3
- 1.2 Significance . . . 4
- 1.3 Research Question . . . 5
- 1.4 Assumptions . . . 6
- 1.5 Limitations . . . 6
- 1.6 Delimitations . . . 6
- 1.7 Summary . . . 8

CHAPTER 2. REVIEW OF RELEVANT LITERATURE . . . 9
- 2.1 Musical Motivations . . . 9
  - 2.1.1 From Late Baroque to the 20th Century . . . 9
  - 2.1.2 ElectroAcoustic Music . . . 10
- 2.2 Technical Overview . . . 11
  - 2.2.1 Introduction . . . 11
  - 2.2.2 Forefathers, Pioneers & Foundational Technologies . . . 12
    - Bell & Edison . . . 12
    - Stereo, Blumlein & MS Encoding . . . 13
    - Michael Gerzon . . . 14
  - 2.2.3 Ambisonics . . . 15
    - Ambisonic Encoding . . . 15
    - Native Ambisonic Arrays and SoundField Microphones . . . 20
    - Multi-Channel Ambisonic Decoding . . . 23
    - ITDs, ILDs and Spectral Cues . . . 24

Page

HRTFs and Binaural Decoders . . . 26
Individualized Versus Generic HRTFs . . . 28
Research Involving MEMS Arrays . . . 30
Simulation Research . . . 32
Comparative Experiments . . . 33
SpHEAR . . . 35
Prior Work by Author . . . 36
2.3 Summary . . . 37

CHAPTER 3. FRAMEWORK AND METHODOLOGY . . . 38
3.1 Mic Design & Assembly . . . 38
3.1.1 Capsule Selection . . . 38
MEMS vs ECM . . . 38
Picking a MEMS Microphone . . . 40
3.1.2 PCB Design . . . 41
3.1.3 CAD Specifications and Requirements . . . 43
3.2 Objective Measurements . . . 44
Polar Plots . . . 45
Frequency Response . . . 46
EIN, AOP and Dynamic Range . . . 47
3.3 Study Design . . . 47
3.3.1 Hypotheses . . . 50
3.3.2 Population . . . 51
3.3.3 Variables . . . 52
3.4 Summary . . . 53

CHAPTER 4. RESULTS & ANALYSIS . . . 54
4.1 Part I - Single Factor ANOVAs . . . 54
4.1.1 Total . . . 54
4.1.2 Freedom From Noise . . . 55
4.1.3 Dynamic Range . . . 56
4.1.4 Tonal Quality . . . 57
4.1.5 Overall Quality . . . 58
4.1.6 Spatial Impression . . . 59
4.2 Results - Part II . . . 60
4.2.1 Stimulus + Microphone Repeated Measures ANOVA . . . 60
4.2.2 Question + Microphone Repeated Measures ANOVA . . . 62

CHAPTER 5. DISCUSSION . . . 65

APPENDIX A. ADDITIONAL FIGURES - OBJECTIVE MEASUREMENTS 67

APPENDIX B. ADDITIONAL FIGURES - POPULATION . . . 69

APPENDIX C. CONSENT FORM . . . 71

LIST OF REFERENCES 73

LIST OF TABLES

Table		Page
4.1	Anova: Single Factor - Total	55
4.2	Freedom From Noise - ANOVA - Single Factor	56
4.3	Dynamic Range - ANOVA - Single Factor	57
4.4	Tonal Quality - ANOVA - Single Factor	58
4.5	Overall Quality - ANOVA - Single Factor	59
4.6	Spatial Impression - ANOVA - Single Factor	60
4.7	Stimuli + Microphone Repeated Measures ANOVA	61
4.8	Question + Microphone Repeated Measures ANOVA	62

LIST OF FIGURES

Figure Page

2.1 MS Encoding . . . 14
2.2 FOA Mic w/ Type I Labeling Scheme . . . 16
2.3 Spherical Coordinate System . . . 19
2.4 Native Ambisonic Array . . . 21
2.5 Regular Decoder Loudspeaker Set-Up . . . 25
2.6 Cone of Confusion . . . 27
2.7 BACCH Binaural Microphone . . . 29
3.1 Top Port & Bottom Port MEMS . . . 40
3.2 ICS-40720 Landing Pattern . . . 42
3.3 40720 PCB 6.35mm Diameter . . . 42
3.4 Form2 . . . 44
3.5 ICS-40720 FOA Mic Polar Response . . . 45
3.6 MEMS Vs. Ambeo Frequency Response Raw . . . 46
3.7 NVSonic Head Tracker . . . 48
3.8 Degrees of Freedom in VR. . . . 49
4.1 Inter-quartile Range - Box Plot . . . 63
4.2 Means/Variance - Bar Graph . . . 64
A.1 Sennheiser Ambeo VR Mic Polar Response - Single Capsule . . . 67
A.2 Frequency Response Pre/Post Filtering Vs. Ambeo . . . 68
B.1 Questionnaire - Age - Responses . . . 69
B.2 Questionnaire - Hours Music Per Day - Responses . . . 69

Figure Page

B.3 Questionnaire - Experience VR - Responses 70
B.4 Questionnaire - Experience 3D Audio - Responses 70

ABBREVIATIONS

FOA	first order ambisonics
HOA	higher order ambisonics
MEMS	micro-electronic mechanical systems
HRTF	head related transfer function
BIR	binaural impulse response
BRIR	binaural room impulse response
ITD	interaural time difference
ILD	interaural level difference
SNR	signal-to-noise ratio
PCB	printed circuit board
CAD	computer-aided design
ECS	electret condenser microphone
ASIC	application specific integrated circuit

GLOSSARY

Head-tracker — headphone mounted accelerometer and gyroscopic system for real-time binaural decoding of ambisonic content with dynamic head movement interaction.

Electret capsule — electrostatic capacitor-based microphone.

Stereophonic sound — in contrast to monophonic audio, stereophonic playback systems allow for the manipulation of sound source location via psycho acoustic principles.

Binaural microphone(s)

- a dummy head such as the KU 100 by Neumann, which consists of a mannequins head with microphones embedded at each ear canal.
- a set of in-ear microphones, such as the Bacch-BMs, which resemble earbuds but contain microphones rather than speakers allowing us to capture personalized HRTFs.
- other related technologies such as 3Dio microphones containing multiple binaural systems used for interpolation.

Transaural reproduction — an advanced stereophonic reproduction system which employs cross-talk cancellation filters in order to segregate left and right channels without headphones or ear buds.

Periphony	the recording and reproduction of a full sphere of sound directions covering the whole of 3 dimensional acoustical space. (Gerzon (1980))
Quadraphonic sound	equivalent to 4.0 surround sound systems which employ two additional speakers placed behind the listener(s).
Median plane	plane demarcating the left and right sections of one's head or body.
Transverse plane	plane demarcating the top and bottom sections of one's head or body.
Frontal plane	plane demarcating the frontal and posterior sections of one's head or body.
Holophony	also known as wave field synthesis (WFS), a 3D audio method with aims to recreate a wave front by using a wall of microphones and speakers.
Isotropic	in the context of ambisonics, it refers to the idea that sounds from all directions are treated equally, in contrast to some surround sound techniques[1].

[1]In which rear channels are designated only special effects

ABSTRACT

This thesis regards the creation and evaluation of a First Order Ambisonics (FOA) MicroElectronic Mechanical Systems (MEMS) enabled microphone featuring increased capsule coincidence. Four ICS-40720s were surface mounted on custom made printed circuit boards (PCBs). Subsequently, these four PCBs were arranged in a tetrahedral configuration using a 3D printed model. The form factor of these systems allow us to examine whether there can be improvements towards localization, and a general sense of immersion, when capsules are more coincidentally[2] co-located.

Objective features such as polar plots and frequency response, from former experiments, are also presented to aid with the analysis and justify design criteria. The resulting microphone was subjectively compared against a professional FOA microphone using a head-tracker during binaural reproduction. Subjective responses from voluntary participants in the NYU Music Technology department were collected.

The subjective experiment undertaken consisted of a sound attribute evaluation in which our MEMS enabled FOA microphone was compared with to the Sennheiser Ambeo VR Mic, an industry standard in the FOA category. Subjects were presented with a number of different musical stimuli featuring a range of different genres and instruments.

Results show that the MEMS ambisonic microphone provides extremely rich localization information. Unfortunately, these system also suffer from noise, which

[2]Coincidence referring here to the proximity of multiple transducers

draw the listeners out of the experience and degrade the tonal balance. A number of different statistical methods were use to analyze subjects' responses.

Chapter 1

Introduction

The aim of this thesis is to evaluate if modern MicroElectronic Mechanical Systems (MEMS) capsules can perform, in a FOA context, as well or better than their electret counterparts. Despite the plethora of ambisonic research, and the proliferation of MEMS capsules in commercial hardware devices, recording engineers seemed to have dismissed these systems as a means of sound capture due to their, admittedly, limited dynamic range. The aim of this thesis is to evaluate whether these systems' form factor, namely with respect to their reduced size, provide benefits over other solutions in first order ambisonics (FOA) capture and reproduction.

Immersive audio, as a field in general, is becoming increasingly relevant in the 21st century as a result of the dropping cost of hardware manufacturing and the high availability of mobile devices capable of performing the intense computations required to experience virtual environments with true spatial audio. Disciplines such as: medicine, education, and even journalism, are already being impacted by the rise of VR (Mennecke et al. (2007)). Immersive audio, a name often given to 3D or VR audio, deals with the problem of capturing and reproducing sounds, as realistically as possible, in order to create or reinforce a sense of immersion. These immersive experience can, thus, be used to: expedite patient recovery (Lamson

(2002)), cement educational material (Burdea Grigore and Coiffet (1994)), or deliver superior journalistic experiences (De la Peña et al. (2010)).

The field of immersive audio can be loosely divided into the different techniques that are used for the capture and reproduction of auditory experiences including: wave field synthesis (WFS), binaural audio, object based audio (OBA), or ambisonics. These techniques are also often implemented in a complementary fashion in attempts to overcome the psycho-acoustic, or logistical limitations, associated with each. Mathematically, some of these methods can be considered as differing interpretations of the same underlying principles. [1].

Colloquially known as just surround sound, today's immersive audio systems have evolved beyond what is commonly understood by even the most informed audiophiles[2]. Despite the commercial success of 5.1 surround sound systems for home entertainment, few consumers are actually aware of the greater possibilities of spatial audio sound systems, which allow for sounds to emanate from above or even below the listener. Optimizing these systems requires providing the ideal auditory experience for all listeners and not just those situated inside the best listening position, or *sweet spot*. For many years this has been accommodated by non-isotropic[3] systems which use rear channels for special sound effects, allowing patrons to focus on cinematographic content. With the rise of new entertainment formats, such as VR, AR and MR[4], it is becoming increasingly important to provide immersive sound experiences that treat sounds equally from all directions.

Object-based audio (OBA), ambisonics, and WFS are just some of the techniques attempting to solve this and many other problems. Some of these techniques have already seen commercial success in commercial theatres. Modern music events, by artist such as Alt-J[5], which attempt to provide audiences with

[1]Such is the case for spherical holophony and ambisonics. (Zotter (2009))
[2]From the Greek, in love with audio
[3]Not treating sounds from all directions equally.
[4]Virtual, Augmented and Mixed Reality.
[5]http://www.nme.com/blogs/nme-blogs/alt-j-forest-hills-stadium-new-york-immersive-sound-2249446

immersive sound experiences, are consequently emerging as interest in this technology increases. The development and commercialization of cross-talk cancellation systems for transaural reproduction (Choueiri (2008)) as a means for immersive audio capture and reproduction further demonstrate how quickly the field is expanding.

While all of the aforementioned techniques, such as OBA or WFS, have virtues of their own, this thesis will focus particularly on research undertaken in the field of ambisonics: a capture and reproduction method dedicated to capturing *sound fields* via microphone arrays[6]. In order to introduce the reader to this subject, this thesis will begin with an overview of the historical background pertinent to the development of ambisonics, which includes a synopsis of the composers who have shaped the way we understand the role of space in music. The key terminology required to understand the minutiae of the research will be introduced during section 2.2. The literature review will then expound the importance of this project, and contextualize it, by citing vital research. Finally, the methodology, results, analysis, and discussion sections, pertinent to the research conducted for this thesis, will be presented.

1.1 Scope

While there has been a plethora of research done in the field of immersive audio over the last few decades, little seems to be known about the specific research question this thesis addresses. Dabin, Ritz, and Shujau (2015) describe a similar project in which they showed that, while systems such as the one proposed herein can provide highly accurate localization at high frequencies, they might also suffer from poorer directivity at lower frequencies. It should be noted that the MEMS system selected in that that experiment differed not only in terms of SNR but also

[6]Or synthesizing them via the encoding of pseudo-independent audio signals

in frequency response. Chapter 2 will elaborate upon Dabin's research and present various other experiments similar to the one undertaken for this thesis.

Objective and subjective methods outlined in chapter 3 will also present some key features of this work which help differentiate it from research performed in the past. Special care will be taken to frame this research in the context of other ambisonics research while adding emphasis to research undertaken in the 21st century with similar features to the research proposed herein.

1.2 Significance

With the surge of interest by consumers in virtual reality (VR), companies and universities have been developing and marketing technologies that focus on providing realistic content to be used as educational material (Kaufmann, Schmalstieg, and Wagner (2000)), simulation based training (Gallagher et al. (2005)) or even psychiatric treatment (Parsons and Rizzo (2008)).

While it remains to be seen the degree to which providing rich auditory experiences is a key to optimizing any or all of these use cases, we remain hopeful that further research will show the significance of 3D audio, allowing the field to further expand. One exemplary study by A. Rizzo (2002) describes the impact that VR could have on people with disabilities. We can easily imagine the emotional impact it would have on a paraplegic patient, to, for example, listen to a choir *in* his or her childhood church, with all the detail they remember.

While VR experiences have become increasingly sophisticated, there seems to be no standard method yet to guide a users attention during narrative experiences. This reported problem has extensively been investigated by authors such as A. A. Rizzo et al. (2000). The format in which films are consumed today make it easier for viewers to know what to pay attention to. In contrast to two-dimensional cinematographic experiences, VR experiences, consumed via a Head Mounted Display (HMD), give the audience members the autonomy to gaze in any direction

at any time, often at the cost of missing what directors or developers wanted them to see. Spatial audio is not only a convenient narrative tool, in this respect, but also elegant in its correspondence to our natural human behavior.

To date, there exist only a handful of microphones on the market capable of capturing ambisonics. Many of these, unfortunately, still remain inaccessible to a lot of people due to their high price point. Part of the motivation behind this thesis is to help the reader understand more about the technology behind some of these ambisonics arrays, facilitating future research, and, allowing for further development of solutions designed with affordable manufacturing processes in mind.

1.3 Research Question

While the focus of this thesis is to explore whether MEMS capsules can improve spatial audio fidelity in a FOA context due to their reduced size, which allow increased capsule coincidence, it is important to remember that more factors beyond just capsule coincidence can affect this. Directivity response, and overall capsule audio capture quality are among some of these factors. Most notably, in this thesis, we will underline distinctions in signal-to-noise ratio (SNR) specifications in MEMS versus electret capsules resulting in the potential under-performance of MEMS based FOA systems[7].

In order to determine whether or not this new highly coincident MEMS FOA mic can convincingly capture FOA, identical audio scenes were captured, in a controlled environment, with a professional FOA mic and our MEMS mic. These auditory stimuli were reproduced with a binaural renderer, along with a head-tracker, and presented to each subject. Objective measurements of a MEMS enabled microphone similar to the one proposed herein are also reported.

[7]MEMS capsules generally suffer from a lower SNR due to thermal noise limits (Kim and Lee (2015)).

1.4 Assumptions

We assume, for this study, that all participants used during the gathering of subjective data answered all questions in a deliberate and thoughtful manner. In order to partially validate results a number of statistical methods were used during analysis. It should be noted that there is a reasonable degree of expectation that subjects did answer deliberately given their background and field of research. Additionally, we assume that all subjects self-reporting healthy hearing were honest and, thus, we do not compensate for any potential hearing loss at any point during our listening experiments. Participants reporting partial or complete hearing loss were deemed inadmissible.

1.5 Limitations

This study was limited by the number of participants the author was able to perform the test on in the allotted time. Only a small pool of subjects was tested for the purposes of this experiment. While the subjects were all students devoted to the field of music and audio, a larger sample size would have likely improved the results of this experiment. Additionally, only a single binaural decoder was used for this experiment. The decoder did not allow for personalized HRTFs, therefore generic ones had to be used. Other binaural renderers are available which allow personalized HRTFs to be used, unfortunately, time did not allow for capture of these since it it widely known that high-quality HRTF measurements are a time intensive process. The details regarding the deficits of generic HRTFs, and the general concept of HRTFs will be outlined in chapter 2.

1.6 Delimitations

A single professional FOA microphone was selected for comparison, instead of multiple ones, due to: the nearly identical geometry of the two available FOA

microphones (Core Sound TetraMic and Sennheiser Ambeo), and their similar SNR capsule specifications (75dB and 76dB respectively (Vinkel (2017))). Related ambisonic methods such as native arrays (E. Benjamin and Chen (2005)) or encoding of pseudo-independent audio signals (Neukom (2007)) were not compared. While techniques such as these could have theoretically been used, we focus herein strictly on FOA microphone arrays in tetrahedral configurations to simplify the analysis of our results.

Monophonic audio experiences were not introduced in our subjective experiment. It was deemed unnecessary to compare mono files with FOA due to the large disparity between formats. No HOA microphones were tested due to their unavailability. Given the small pool size of subjects no control group was used in this research. A single experimental trial was all that was possible given the amount of time provided and the number of subject willing to participate. A more robust and comprehensive experiment could include a dedicated control group and multiple experimental trials.

In Gerzon (1975), the author describes the possibility of using encoding filters to compensate for the impossibility of perfect capsule coincidence. With the help of Angelo Farina, an ambisonics researcher at the University of Parma, the equations from Gerzon's paper were translated into a Matlab function that performed this compensation. Ultimately, due to work presented by (Bates et al. (2017)), we decided not to use these filters, as these were shown to have little to no effect on localization while adding unwanted high frequency boosting.

We also decided to compare the MEMS microphone against the Ambeo, and not our formerly designed MEMS microphone (Zalles et al. (2017)), in order to determine if the SNR and polar response differences were enough to out-weight the theoretical spatial quality increase due to the capsule coincidence. Our former design featured capsule coincidence equivalent to the Ambeo. The results of that experiment showed that the main problem with the MEMS system, in subjective assessments, was the high-frequency boost caused by Helmholtz resonances

associated with MEMS systems. This issue was mitigated in this experiment via a Matlab filtering system. Details describing the frequency response of the MEMS system and this filtering system are described in 3.2.

1.7 Summary

This chapter provided the scope, significance, research question, assumptions, limitations, delimitations, definitions, and other background information for the research project. The next chapter provides a review of the literature relevant to this thesis.

Chapter 2

Review of Relevant Literature

This chapter provides a review of the literature relevant to this thesis. The literature review section will be divided into two parts: the musical motivations section, and the technical overview section. The intention is that this organizational structure will give the reader a sense for the evolution of this technology from both the humanities and scientific perspective.

2.1 Musical Motivations

In this section of the literature review we will discuss the interaction between the arts and science, and how the work of composers throughout history have motivated the development of new technologies that satisfy these artists' creative visions. Namely, the first part of this literature review will present to the reader how audio technologies were shaped by the creative contributions of composers and musicians from various eras.

2.1.1 From Late Baroque to the 20th Century

Spatial music refers to compositions which employ spatial attributes as a means of creativity. One of the earliest examples of it can be traced back to the biblical times. Antiphons, which translated from Greek means opposite voice, is a

type of chant which features call and response patterns. These chants, which originally used psalms as the subject matter, were performed by people of the Jewish faith in the Middle East. Often the two sections of the choir dedicated to alternating verses would be side by side, but as music evolved, spatial positioning of singers and instruments became a stylistic feature for composers to experiment with.

One of the early published works which used space as a compositional tool comes from *maestro di cappella*, Adrian Willaert, who used a technique called *cori spezzati*, which translates to separated choir. His 1550 piece inspired later generations of Venetian composers such as Willaert's prodigy Andrea Gabrielli. Composers in England, galvanized by this movements, began, in the sixteen-hundreds, composing similar pieces. In honor of Queen Elizabeth, up to 40 separate vocal parts in 'eight 5-voice choirs' were called for to celebrate her 40th birthday (Zvonar (1999)). The high point for spatial music during the baroque era was Orazio Benevoli's *Festival Mass* in 1628, which called for 53 parts plus two organs and a *basso continuo*, or continuous bass.

Many years later, spatial antiphony would return to the forefront during the Tuba Mirum[1] section of Hector Berlioz's *Requiem* (1837). Berlioz's called for four separate brass ensembles to enter from four different points, one from each cardinal direction. Gustav Mahler's *Symphony No. 2* (1895) also employs off-stage brass ensembles. Composers such as Luigi Russolo, Charles Ives and Henry Brant would later continue this tradition, during the 20th century. Brant's *Voyage Four* (1963) called for three conductors on stage, violins on one side balcony, violas and celli on another balcony, basses on the floor level at the rear, woodwinds and a few strings on the rear balconies, and even a few performers in the audience!

2.1.2 ElectroAcoustic Music

The invention of sound recording, radio and telephony brought with them new era of musical experimentation. Notably, we have Theremin's ensemble

[1] 'Hark, the trumpet'

performances, which are considered one of the first examples of multi-channel speaker music. In 1948 Pierre Schaeffer, at Radiodiffusion-Television Francaise (RTF), presented some of his early works created with disk recorders, which he dubbed *musique concrete*. Multitrack recorders were not available at the time, so Schaeffer, in collaboration with Pierre Henry, had to use multiple mono tape decks in order to create their musical visions. The 4-channel speaker system used for the presentation of their piece was arranged in a tetrahedral configuration with two front speakers, one posterior speaker and one overhead speaker.

On the other side of the pond, at the same time, John Cage and his colleagues from The Project for Magnetic Tape were experimenting with tape splicing methods in music composition. *William's Mix* (1952), one of Cage's most famous pieces, and one of the first examples of chance operation for randomized selection of sound, calls for 8 equidistant speakers, fed by 8 mono tape machines. Other important works from the time include Earle Brown's *Octet* and Morton Feldman's *Intersection*.

2.2 Technical Overview

2.2.1 Introduction

As we discussed during section 2.1, a number of composers have throughout the course of time, in one way or another, experimented with the spatial properties of sound as a creative means for musical expression. Unfortunately, due to the widespread unavailability of immersive audio system, not only then but now, only a few lucky individuals allowed the pleasure of attending a live performance, have had the privilege of enjoying the works of these composers the way they were intended to. While recordings of these happenings do exist (Theile (2001)) most people still lack the proper equipment to adequately reproduce these recordings.

This leads us to our next section of this work, which describes research surrounding the topic of ambisonics, a technology which is quickly becoming the

standard for binaural 3D audio reproduction, and allowing more people than ever before to experience 3D audio. Before diving into the specifics of this capture and reproduction method we will lay the foundation upon which this technology was built by presenting to the reader a number of influential figures which paved the way for where we are today.

2.2.2 Forefathers, Pioneers & Foundational Technologies

Bell & Edison

Alexander Graham Bell, widely accredited for inventing the telephone, in large part for his successful patenting of the technology 1876, had a profound effect on music technology. Electronic music, as we know today, would have never been possible without telephony (Grosvenor and Wesson (2016)): a technology which completely revolutionized music experiences, and created the possibility for amplification and musical recordings. In some of his early works, far before musical recordings were a reality, Bell proposed alternate ways in which his device could be used, including: recording, amplification and the transmission of musical content over long stretches of distance.

In 1877, just a year after Bell's telephone patent, a different iconic figure, Thomas Edison, would make the first successful mechanical musical recording. Unfortunately, it would take years of development for the first real electrical recording. Prior to this, recordings were made acoustically by indenting a piece of tinfoil or wax. Despite the development of electrical amplification, condenser microphones and radios, around the time of the first world war, it was not until 1925 that the first ever electrical recording took place (Chanan (1995)). Today, both the real-time delivery of spatial audio, and capacity to store these multi-dimensional sound experiences in digital formats, are not only a reality, but the technologies which enable them are continually becoming more sophisticated due to the tireless work of: researchers, students and companies around the world.

Stereo, Blumlein & MS Encoding

It is hard to imagine now that there was ever a time when monophonic audio was the standard. Alan, more than anyone perhaps, should be thanked for the commercial introduction of the stereophonic system, which has allowed us to understand the role of space in music. Alan Blumlein, along with Bell and Edison, for his contributions regarding reproduction and microphone systems, is considered one of the most influential figures in audio. Before Blumlein, most theatre reproduction systems were monophonic. Blumlein, unhappy with the way these monophonic reproduction systems misrepresented sound, as actors moved around on screen, coined and patented the term binaural sound: a process which aimed at reconciling image and sound by accounting for our dual-pinnae[2] anatomy (Alexander (2013)).

Prior to Michael Gerzon's[3] birth, in 1933, Alan Blumlein had patented the stereophonic recording technique known as Mid-Side (MS) (Dooley and Streicher (1982)). MS uses one figure-eight microphone and one omnidirectional microphone to, as the name suggests, capture the middle of a sound scene as well as side information. Traditionally, this technique used an omnidirectional microphone, aimed directly at the sound, as well as a coincidentally located figure of eight microphone, 90 off-axis, dedicated to capturing room ambience. Today, it is more common to see this technique employed with a cardioid microphone, in lieu of the omnidirectional center mic, as the null point in the posterior part of the microphone often results in a more controlled recording. Figure 2.1 shows an example of the MS work-flow [4].

With MS, by duplicating and phase inverting one of the copies of the side recording, audio engineers gain access to a flexible way to add stereo information in post-production (Dooley and Streicher (1982)). Ambisonics can be considered an extension of the MS technique as it also makes use of: a highly coincident

[2]Pinna(e): the external part of our ear.
[3]Inventor of ambisonics
[4]Image 2.1 sourced from *Next Level Sound Forum*

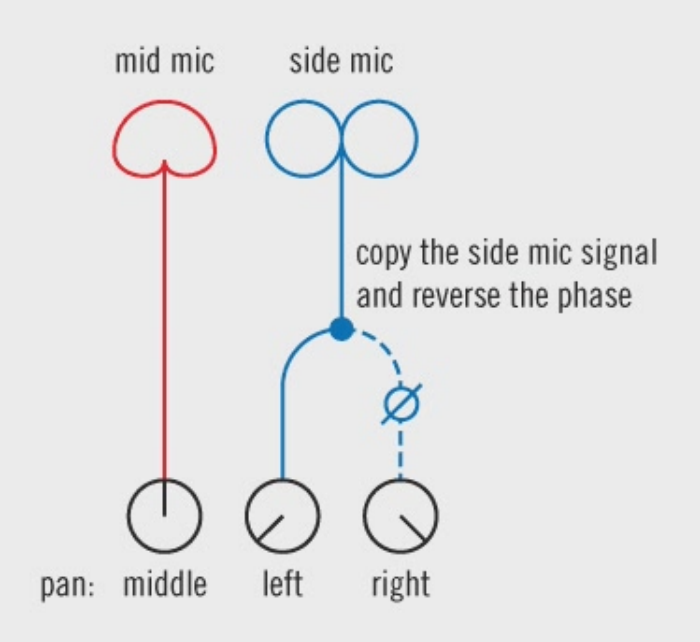

Figure 2.1.: MS Encoding

microphone array, and a signal matrix operation, for post production flexibility. Additionally, in both cases, the resulting audio is used for multi-channel reproduction [5]. (Malham (1999))

Michael Gerzon

Born on December 4th 1945, Michael Gerzon, from the University of Oxford, is widely credited, along with Peter Craven, with popularizing ambisonics technology (Thornton (2009)). By expanding Blumlein's technique, ambisonics can, not only record sounds arriving from all directions, as an omnidirectional microphone would, but encode directional information of sound sources as well. Decoding of these formats is designed to overcome limitations of stereo which provide listeners with sounds on strictly the horizontal plane. Part of the motivation behind Gerzon's work was figuring out the optimal way to successfully commercialize and improve upon quadraphonic sound, which, during the seventies, was predicted to become the

[5] Stereophonic sound is multi-channel in nature

new standard for audio reproduction. In a sense, the development of the tetrahedral array can be seen as a direct response to the struggles of quadraphonic sound, which never gained enough favour from the public to be considered a success.

In one of his seminal papers, used as the inspiration for the title of this thesis (Gerzon (1975)), Gerzon explains how these precisely coincident microphone arrays, while desirable, are unrealistic due of the practical inability to perfectly co-locate the multiple transducers required to perfect ambisonic encoding. This capsule separation, along with unsuitable polar responses, Gerzon explains, is one of main causes of poor image localization. To overcome this, the highest possible coincidence and least number of capsules required for point-source capture is selected in our methodology, since, according to Gerzon, this is the best practical approximation to a uniform covering of the sphere. Unlike Gerzon, today we have the benefit of being able to count on extremely small and robust transducers which can be used to increase the coincidence of microphone arrays, approximating Gerzons theoretical models more closely.

2.2.3 Ambisonics

Ambisonic Encoding

When referring to tetrahedral arrays, encoding refers to the sum and difference matrix, similar to Blumlein's MS technique, required to convert the raw A-format recording, captured by an FOA soundfield array, into B-format, the standard format used today for ambisonic reproduction. C, D and G-formats also exist but are less often mentioned.[6]

Knowing the order and orientation of the four capsules found in a soundfield array is the first step in being able to encode the A-format signals. Figure 2.2 shows an example of a FOA microphone with Type I labeling scheme[7]. It is important to notice that in the image the Front Left Up (FLU) capsule (channel I) is actually on

[6] Ortolani (2015) provides an overview of these other formats in section 1.3
[7] Type II schemes exist as well [FLD-FRU-BLU-BRD (DPA-4 uses this)]

the right. This is because during recording the microphone will actually be facing the direction of the musician, thus rotating the right and left capsules around the frontal plane.

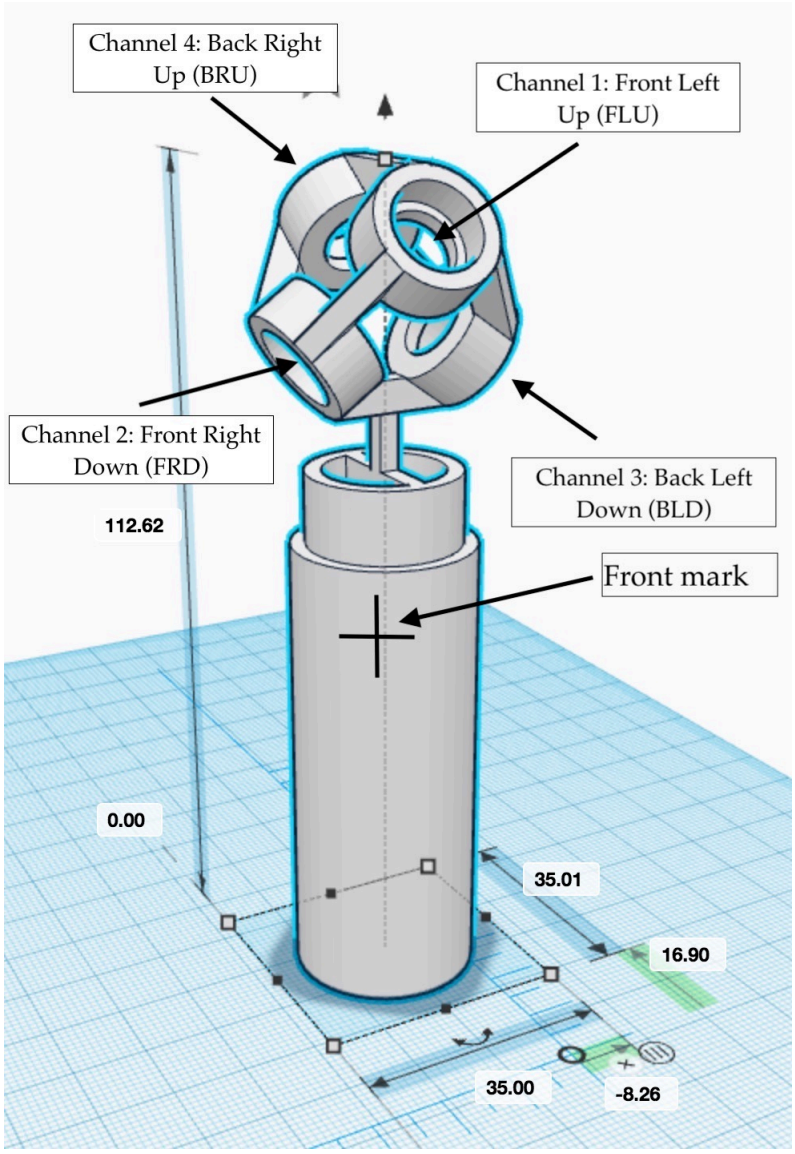

Figure 2.2.: FOA Mic w/ Type I Labeling Scheme

In order to convert these raw A-format signals into B-format the following method by Ortolani (2015) is employed. During the design of our microphone and encoding algorithm, naming scheme number I was employed due to it being the

same one the Sennheiser Ambeo VR Microphone uses. This allowed us to use a single Matlab script to process both sets of wave-forms instead of having to rely on any DAW based plug-ins, saving considerable time.

$$W = FLU + FRD + BLD + BRU$$
$$X = FLU + FRD - BLD - BRU$$
$$Y = FLU - FRD + BLD - BRU$$
$$Z = FLU - FRD - BLD + BRU$$

An intuitive way to derive equation 2.2.3 without referring to this work is to understand the underlying principle of FOA tetrahedral encoding. What we seek the accomplish is to matrix the four capsules in such a way that three virtual figure-8 microphones result from the following A-format signals. Each figure eight microphone has a positive and negative side. The positive side for the X-axis is in the front, on the left for the Y-axis and above for the Z-axis. Therefore: the X-axis microphone can be derived by summing both front microphones and subtracting the back microphones, the Y-axis microphone can derived by adding the left microphones and subtracting the right, and the Z-axis microphone is derived by subtracting the bottom microphones and adding the top.

The resulting four channels are then normalized and ordered based on the decoder's specifications. Additional normalization processes were applied to our stimuli in order to ensure equal amplitude during reproduction. Here we make a distinction between normalization schemes of spherical harmonics and the traditional normalization often talked about in the field.

For our traditional normalization, two steps were employed. Firstly, all four raw A-format signals were normalized by the global maximum of the four signals. This ensures that the outputs are proportionally weighted. After encoding is performed, via equation 2.2.3, the spherical harmonics (W, X, Y, Z) are once again normalized, this time using the local maximum.

Subsequently, the W channel is attenuated, as per the SN3D or MaxN normalization scheme. Both of these normalization schemes result in the same thing for FOA recording, specifically calling for a multiplication of the W channel by $sqrt(0.5)$.

Since the FB360 Spatial Workstation accepts both ACN and FuMA orderings, and both these orderings' normalization schemes were equivalent for FOA, it did not matter which one we selected. As a reference we provide these two orderings for the readers edification.

FuMa = [W X Y Z]

(*Usually paired with MaxN normalization.*)

ACN = [W Y Z X]

(*usually paired with SN3D normalization (also called AmbiX).*)

While this transformation into what is known as B-format signals is of primary importance in this literature, it should be noted that a number of other pattern configurations are possible; this makes ambisonics a powerful technique not just for spatial audio, but also for general purpose recordings as it grants engineers the flexibility of experimenting with different polar patterns even after the recording has taken place. For example, after deriving the Y-axis figure-8 microphone and W (omni) channel, a simple MS array can be constructed. Alternatively, the front/back (X-axis) figure-8 can be added with the W channel to create a cardioid mic. A slew of additional combinations can be experimented with.

Ambisonics encoding can also refer, as aforementioned, to the encoding of pseudo-independent audio signals. This technique also results in four spherical harmonics, but, in contrast to soundfield arrays, the positions of the instruments is a decision that can be made a posteriori. This associated technique is perhaps more common in the field of immersive audio as it takes advantage of common recording practices, entirely circumventing the necessity for ambisonic arrays of the type proposed.

Hollerweger (2005) provides the following formulas in this scenario:

$$W = 1/k \sum_{i=1}^{k} s_i[1/sqrt2]$$
$$X = 1/k \sum_{i=1}^{k} s_i[cos\phi_i cos\theta_i]$$
$$Y = 1/k \sum_{i=1}^{k} s_i[sin\phi_i cos\theta_i]$$
$$Z = 1/k \sum_{i=1}^{k} s_i[sin\theta_i]$$

Here S corresponds to our audio signal, k corresponds to the total number of signal, ϕ_i (phi) corresponds to the horizontal or azimuth angle, and θ_i (theta) corresponds to the vertical or elevation angle, based on the spherical coordinate system (depicted in figure 2.3).

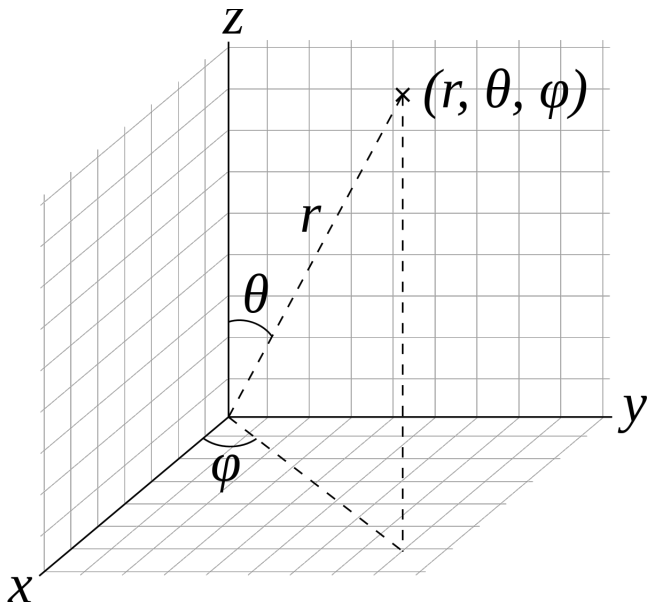

Figure 2.3.: Spherical Coordinate System

W, as always, corresponds to the zeroth order harmonic and X, Y and Z correspond to the 1st order harmonics in the front/back, left/right, and up/down axes accordingly. In practice, while this technique can yield good results, it is also prone to a number of issues. One issue which can occur is the distortion of spatial image as a result of one not possessing the necessary information required to position the musicians in their correct locations. Unless the recordings exhibits little to no cross-talk, or *bleeding*, the spatial image will be hard to accomplish as faint mirrors of various performers will be smeared throughout. A possible solution to this problem could be to record the musicians, individually, under free-field like conditions. This, in theory, could give us the flexibility to place musicians in any position we wanted or even automate their position freely. Unfortunately, a few issues arise from this:

- Free-field like recording settings are often inaccessible for most.
- Recording musicians in this manner would take additional time.
  - As opposed to tracking all musicians simultaneously, &,
- There would be a total lack of natural reverberation in the final output.
  - Which makes it more difficult for sources to be localized (Gerzon (1974)).

Native Ambisonic Arrays and SoundField Microphones

An alternate solution to the FOA tetrahedral array systems proposed by Gerzon is the *native* ambisonics array, sometimes called a Nimbus-Halliday array[8]. In contrast to a system of four cardioid capsules in a platonic solid configuration, which are used to encode the output of virtual microphones, these native systems jump straight into B-format, albeit, at the cost of capsule proximity. Namely, these systems combine the output of three figure-of-eight microphones plus an omnidirectional microphone, carefully positioned and aimed, to simulate the

[8] After Nimbus Records.

encoded output of a tetrahedral array. One of the advantages of systems like this is that they require no encoding equations and can be produced with readily available microphones. Figure 2.4 shows an example of a native ambisonic array [9].

Figure 2.4.: Native Ambisonic Array

Other native ambisonic arrays have been proposed in the past. Some arrays use four cardioid microphones pointed in the same dihedral angles used by soundfield arrays (Gerzon (1973)). Other systems, such as Geluso's Double MSZ, use two mid-side systems combined with a figure-of-eight microphone for height to reproduce B-format signals (Geluso (2012)).

E. Benjamin and Chen (2005) have presented a two-part paper in which a number of tetrahedral arrays were compared objectively and subjectively with native systems. They explain that one important feature dictating the performance of these systems is their direct/reverberant response, which should be flat in order for the reverberant field to be properly reproduced. They show that, while the polar response of the soundfield microphone deviates at high frequency, its diffuse-field

[9]Image 2.4 sourced from *ambisonic.info*

response compensates for it. In contrast, the native systems have great polar pattern but poor free-field and diffuse-field response. E. Benjamin and Chen (2005) conclude that it is not possible to equalize both the diffuse and free-field response to be flat, due to the fact that the 0th order channel rolls off at high frequencies but the 1st order channels *rise* in response at high frequency.

The second part of E. Benjamin and Chen (2005) involved recording these different systems and subjectively evaluating them. A few methodological processes, which are common to both this work and theirs, are noted. Firstly, the use of a comparative approach for subjective evaluation. By this we mean than rather than entirely randomizing the playback of stimuli, the two microphones were aligned and played one after the other, allowing for synchronous evaluation of attributes. Subjects, in our case, asked when they wanted us to switch microphones, since Reaper did not allow for a Graphic User Interface (GUI) to give subjects complete autonomy. Other authors have created GUI using MaxMSP which allow for binaural decoder VSTs[10] to run in the background. This approach will be taken in future experiments as it provides greater accuracy and autonomy for participants.

Secondly, the use of re-recorded stimuli can be seen in both the aforementioned research and the one here. This is pretty common in the field as it allows listeners to judge sound quality attributes without researchers having to worry about differences in music performance. It also allows us to more critically evaluate sound localization performance as was done in Bates et al. (2017) by recording stimuli with microphone systems one at a time. The exact height and distance from all walls was measured here to ensure our MEMS system and the Ambeo array were in the same position when recording our stimuli.

Finally, some similar results can be observed regarding the results of each subjective study which found that the native system composed of lavalier mics suffered from self-noise. This was likely due to the size of these capsules, which are on the smaller end of the spectrum. The native lavalier system, however, when

[10]Virtual Studio Technology

reproducing reverberant recordings, was found to be equivalent to that of the tetrahedral system. The effect of RT60 (reverberation time) and diffusion on MEMS FOA recordings is the subject of future reesearch.

Multi-Channel Ambisonic Decoding

One of the interesting features of ambisonics the recordings themselves is that they are decoupled from the playback system. In this respect ambisonics is considered layout agnostic. An ambisonic encoded sound field can be reproduced on any ambisonic decoding system (Hollerweger (2005)).

While ambisonics over loudspeakers still remains fashionable for some, it poses a serious accessibility problem for most since the smallest number of loudspeakers capable of accurately reproducing a first order ambisonic (FOA) sound field[11], the simplest sound field recording, periphonically, is four speakers (Gerzon (1973)).

This number increases even further when dealing with second and third order ambisonic sound fields, which promise wider sweet spots during loudspeaker reproduction (Spors and Ahrens (2007)). In general, the number of speakers L must always be equal to or larger than the number of ambisonic channels, N:

$$L >= N$$

The number of ambisonic channels, for different ambisonic orders can be found via the equation:

$$N = (M + 1)^2$$

Where M is the order of the ambisonic system.

[11]Name given to FOA or HOA recordings.

Additionally, traditional ambisonic decoders assume loudspeaker arrays to be regular[12], something which is rarely the case and often a limitation of a consumers listening environment (Hollerweger (2005)). For this reason, a lot of audio and electrical engineers have been shifting their attention to ways of improving ambisonic reproduction over headphones: a technology which, while presenting its own set of challenges, is enabling more people than ever to experience ambisonic audio.

Heller, Lee, and Benjamin (2008) provide a deeper look into decoders. Namely, the authors present a comprehensive overview of some of the psycho-acoustic principles used in the design of these systems. Their paper presents an objective evaluation of a number of decoders, as well as subjective studies revealing a preference for the *AmbDec* decoder by Adriaensen, consistent with predictions made using objective measures.

Heller et al. (2008) list the following features for a good decoder:

- Decoding matrix matched to the geometry of the loudspeaker array in use.
- Phase-matched shelf filters.
- Near-field compensation (NFC).

As the subject of ambisonic decoders qualifies as enough material to justify an entire secondary thesis, the reader is encouraged to explore the references in Heller et al. (2008) for more knowledge on the subject. Figure 2.5 shows an example of a *regular* loudspeaker configuration for ambisonic decoding. Speakers are found at either the vertices or edges of a convex polyhedron. Authors have studied compensation methods for non-regular environments with partial success.

ITDs, ILDs and Spectral Cues

The other family of decoders are called binaural decoders. In contrast to traditional decoders, binaural decoders work by imposing localization information

[12]Spherically arranged with the listener at the origin.

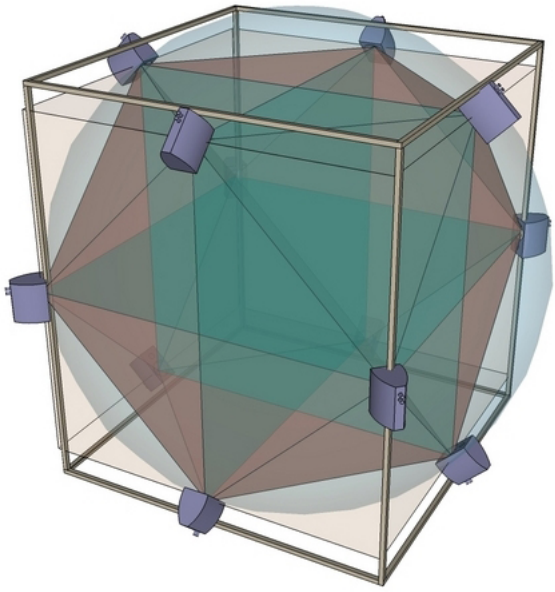

Figure 2.5.: Regular Decoder Loudspeaker Set-Up

unto sounds. These auditory cues are already present during loudspeaker decoding as they are also present in our everyday lives without the need for any synthetic enhancement.

The size and shape of our heads and ears, the distance between our ears, and the intricate design of our outer ear (pinnae), all affect how we interpret sounds. Luckily, all these things can also be quantified and exploited in other to simulate localization. Specific names have been given three primary auditory cues: interaural time differences (ITDs), interaural level differences (ILDs), and spectral cues.

ITDs correspond to the auditory time-of-arrival difference between the two ears. In the simplest case, we imagine a new sound manifesting itself at an unknown location. Unless this sound is directly in front, behind, above or below us, the sound will reach one ear before the other. Our brains are extremely adept at analyzing this information and using it to localize the location of the sound. This is a natural adaption we have developped over millenia to aid in our survival.

The ILD corresponds to the difference in volume with respect to each ear. ILDs are largely a function of the *head shadowing effect*, which in acoustics parlance

is described as diffraction, or the bending of sound around an object. Our subconscious mind understands that the amplitude of a sound is proportional to it's distance from our ears. The ILD is particularly important to determine the overall distance of a sound. The *direct-to-reverberant ratio* informs us about the distance of sounds that are particularly far away.

Perhaps the least known and most intriguing of these three cues are the spectral cues. Spectral cues, in a sense, are used to understand or attempt to explain all unknown phenomena of sound localization which cannot be explained by ILDs and ITDs. When a sound emerges from directly behind us, our pinnae act like filters which cut out high frequencies, letting us know that we should turn around. Spectral cues are also helpful in discriminating the elevation of sounds with identical ITDs and ILDs. These are just some examples of spectral cues in action.

One interesting phenomenon observed by acousticians is the *cone of confusion*. Consider two points directly opposite to each other on the circumference of the cone in figure 2.6. The ITD and ILD for these two points will be exactly the same and only the spectral differences created by the asymmetry of the head will help in differentiating the true position of the sound. A similar, related concept is that of front/back and up/down reversals, in which the same principles apply.

The Hass, or precedence effect, is also used sometimes to attempt to understand the underlying mechanisms involved with sound localization. The law of first wave-front, as it is sometimes called, says that humans perceive sounds as emanating from the first place we hear them come from within a certain time threshold (2-50 ms). This theory was derived to explain how humans can localize fused sounds in the presence of reverberation.

HRTFs and Binaural Decoders

Binaural decoders entail the most common ambisonic reproduction method today. This is in great part due to the availability of: headphones, smart-phones, and open-source software. Binaural decoders rely on sets of Head-Related Transfer

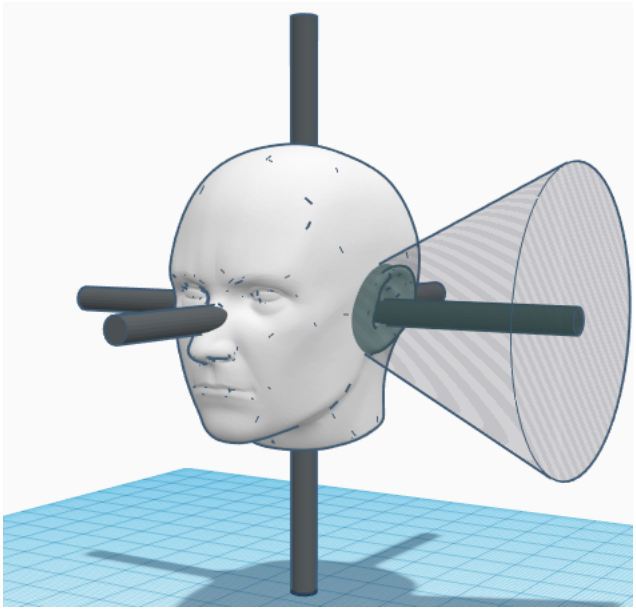

Figure 2.6.: Cone of Confusion

Functions (HRTFs), which contain ILDs, ITDs and spectral cues. Each HRTF acoustically describes, in its entirety, the linear-time invariant (LTI) system created by our physical anatomy (in terms of our heads and pinnae). It can also be defined as the frequency domain representation of a binaural impulse response (BIR).

These BIRs, acquired via deconvolution of binaurally recorded sine sweeps, can be used to virtually position sounds in space. By multiplying the resulting HRTFs with the frequency domain representation of a sound, given by applying the Fast Fourier Transform (FFT) unto the sound, one can give the impression that said sound emanates from any desired direction and distance. Alternatively, one can also convolve the BIRs and the dry sound in the time domain, effectively comprising the same operation, albeit at a cost of performance. Figure 2.7 shows an example of the type of binaural mic one might use to measure personalized HRTFs.

One thing to be understood in regards to binaural decoder is that the HRTFs used for localization do not need to change in real-time, they are static. Instead, when using a head-tracking, the soundfield is rotated and convolved with these static filters (McKeag and McGrath (1996)). This means that, for example,

for a 4-channel binaural decoder, only 4 HRTFs will be necessary. Naturally, this is a coarse approximation of a complex problem in which room asymmetries will be misrepresented since only a small number of filters are used.

Noisternig, Musil, Sontacchi, and Holdrich (2003) additionally propose the following optimizations for ideal binaural decoding:

- Shortening HRTFs filters down to just 128 taps[13]. This was shown to have little effect on localization performance. ((Sontacchi, Noisternig, Majdak, and Holdrich (2002a)) & (Sontacchi, Noisternig, Majdak, and Holdrich (2002b)))
- Using a mixed order system in which vertical directions are encoded at a lower order.
- Filtering virtual loudspeaker signal in the frequency domain.
- Using a recursive reverb network for room simulation. Early reflections and reverberant sound field simulation aid with source localization and an out-of-head experience.

It should be added that in non-free field conditions, these BIRs take on the title of Binaural Room Impulse Responses (BRIRs), as the reverberant nature of the room also becomes part of the LTI system. Free field-like conditions are often approximated via anechoic chambers, carefully designed rooms in which no sounds can enter or exit the room, and in which reverberations are not present.

Individualized Versus Generic HRTFs

The problem with HRTFs, as a means of decoding ambisonic signals, comes from their highly individualized nature: no two HRTFs are exactly alike (Hu, Zhou, Ma, and Wu (2008)). It is for this reason that much research and development is being done on the standardization of a cheap and effective HRTF measurement methods, as well as alternative solutions such as: using a default HRTF considered

[13]Number of delay lines.

Figure 2.7.: BACCH Binaural Microphone

to be an anthropomorphic average of most humans (Bernschütz (2013)); selecting from a database of HRTFs one that is similar to ours based on gender, height and weight (Algazi, Duda, Thompson, and Avendano (2001)); or using machine vision to synthesize HRTFs based on any observable features (Duraiswami et al. (2000)) - to name a few. Unfortunately, while many of these methods have yielded promising results, no individual solution has managed to become a standard yet. As a result the accessibility to solutions capable of providing users with personalized immersive audio content is poor at best.

While the specific subject of this thesis lies outside the scope of the individualized HRTF problem, it bears mentioning because it puts into perspective some of the limitations of any ambisonic research produced with improper HRTF measurements. Additionally, this provision gives an intuition into the lower-level systems that are at play when discussing any research which seeks to understand

the effects of parametric changes in ambisonic systems as they are related to their performance.

Research Involving MEMS Arrays

MEMS capsules today can be found in a myriad of consumer grade home electronics like: cellphones, home assistants and even TVs. Much like any traditional microphone these capsules translate acoustic events into electrical signals, allowing one to store or manipulate audio data at will. MEMS capsules come in both analog and digital formats, the latter of which contain analog-to-digital converters (ADCs), which discretize audio signal into samples based on the timing of a clocking signal provided by another device. In contrast, analog MEMS mics do not require a clock and simply send alternating current (AC) to a audio interface based on the response of a capacitor to air pressure fluctuations.

Research involving spatial audio and MEMS is limited. One of the most similar works to the one described herein was proposed by Dabin et al. (2015). The authors here propose two MEMS ambisonic microphones created with using 3D printing. A single-tier and a three-tier design are analyzed for Direction of Arrival (DOA) accuracy using simulated impulse response.

The author shows that capsule spacing can be optimally chosen based on microphone capsule sensitivity and required DOA accuracy. The three-tier design proposed records 3D sound for a given sub-band to achieve accurate DOA estimation. The authors provide the following equation to derive spatial aliasing:

$$f_{err} = c/2d$$

Where c is the speed of sound and d is the inter-capsule distance.

Traditional B-format microphones feature a 1.47cm (or 14.7mm) inter-capsule separation as proposed by Gerzon in 1978, achieving error free pressure gradients up to 11.6kHz. Our design with 6mm capsule spacing, in theory,

offers error free pressure gradients up to 28.5kHz. The maximum pressure difference reduces relative to wave length λ:

$$\Delta = 2d/\lambda$$

As aforementioned, this capsule spacing is allowed due to the form factor of our transducers. This same form factor results in poor SNR compared to most large diaphragm condenser microphones. Backman (2006) proposes using a large number of MEMS to 'improve SNR over minimum transducer configurations' and to 'provide precise polar pattern control over the entire audio bandwidth'. The author also illustrates the high-frequency effects due to finite transducer spacing as well as scaling this system in order to tune for optimum balance between dynamic and frequency range. The author concludes by proposing 3-Dimensional arrays for ambisonics using MEMS but explaining that Higher-Order components tend to be too noisy in MEMS systems, so, in lieu, *virtual microphones* with adjustable polar patterns should be used.

Kissner and Bitzer (2016) also detail some of the limitations of MEMS systems for microphone arrays. Namely, the authors compared the static noise floor and polar pattern exhibited by single and parallel MEMS microphone configurations with a conventional electret condenser mic (ECM). Their results suggest that 'direct parallel circuits' of MEMS microphones allows further reductions of the noise floor close to the theoretical value of 3dB SPL per doubling of number of microphones while maintaining omnidirectionality below 5 kHz.

Alexandridis, Papadakis, Pavlidi, and Mouchtaris (2016a) also developed a MEMS system with DOA estimation, this time using digital MEMS, with good objective results. Their system, much like that of Kissner and Bitzer (2016), was 2-Dimensional and functioned using properties of beamforming, rather than ambisonic encoding. Their listening tests, with music, showed that digital MEMS can be used for to create spatial audio arrays with uncompromising audio quality.

Simulation Research

Given the rise in popularity of ambisonics related research, audio engineers and acousticians have developed methods which allow them to approximate the theoretical implications of a particular design prior to it's implementation. These open source solutions use mathematical processes along with acoustical theorems to replicate the potential performance of microphone arrays with differing number of capsules and geometries.

(E. M. Benjamin (2012)), uses this very same method to describe an improved high-frequency array by selecting an octahedral geometry. As he explains, FOA microphones have been shown to perform well up to a critical frequencies of 7.3kHz. Above this frequency, he explains, polar patterns tend to become progressively more distorted. In Benjamins paper, a number of simulation outputs are provided which depict the performance of a tetrahedral array with a conservatively small radius. By matrixing all four simulated polar responses, a theoretical response of the B-Format virtual microphones, over seven different bands, is constructed.

The simulations show that for a 1.47 cm radius, degradation begins to occur above 8 kHz. This, according to the author, occurs because of phase differences between capsules, which lead to destructive interference at higher frequencies. The same process is also used by Benjamin to simulate the response of a figure-of-eight virtual microphone, showing substantial degradation above 10 kHz.

While comparison with other simulations done by Benjamin show that, theoretically, decreasing the radius of the circumscribing sphere should increase performance at higher frequencies, it should be noted that the principal assumption in these simulations is that the polar response of the capsule, independent of the matrixed B-Format output, is of cardioid form. Given this assumption it is ill-considered to assume a similar fate for omnidirectional capsules such as the ones proposed in this methodology.

Subjects interested in simulations of this type are refered to http://spatialaudio.net/sofia-sound-field-analysis-toolbox-2/ or Archontis Politis's Github page[14].

Comparative Experiments

Bates et al. (2017) presents a two part paper in which different microphone arrays are compared. The experiment had a MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) design. Namely, they compare a monophonic omnidirectional microphone against a few FOA mics, and an Eigenmike (HOA Microphone by MH Acoustics) and the Zoom H2n (horizontal only ambisonic soundfield recorder[15]).

A novel approach for instantaneous calculations of intensity vectors $I_{x_{i,j}}$, $I_{y_{i,j}}$ and $I_{z_{i,j}}$, as well as sound field energy of B-format recordings is also proposed here. The output of sixteen speakers is captured using the pink noise as the audio for testing. The azimuth, elevation and distance for each speaker is measured. The audio is split using the bark scale, a psycho-acoustical scale in which equal distances correspond to perceptually equal distances[16]. Namely:

$$I_{x_{i,j}} = (W_{i,j}X_{i,j})/\sqrt{2}$$
$$I_{y_{i,j}} = (W_{i,j}Y_{i,j})/\sqrt{2}$$
$$I_{z_{i,j}} = (W_{i,j}Z_{i,j})/\sqrt{2}$$
$$E_{i,j} = (W_{i,j}^2/2) + (X_{i,j}^2 + Y_{i,j}^2 + Z_{i,j}^2)/4)$$

Where i represents a frequency band and j represents a time sample.

Then using the intensity vectors, the magnitudes of source signal in both azimuth and elevation angles can be derived:

[14] https://github.com/polarch
[15] Embrace Cinema Gear used to make an ambisonic system using the H2N
[16] Corresponding to 24 critical bands

$$mag_{a_{i,j}} = \sqrt{I^2_{x_{i,j}} + I^2_{y_{i,j}}}$$
$$mag_{e_{i,j}} = \sqrt{I^2_{z_{i,j}}} + mag_{a_{i,j}}$$

Azimuth θ and elevation ϕ angles are then derived following:

$$\theta_{i,j} = \arctan I_{y_{i,j}}/I_{x_{i,j}}$$
$$\theta_{i,j} = \arctan I_{z_{i,j}}/mag_{a_{i,j}}$$

Using these formulas, and an additional diffuseness (12), estimates for the azimuth and elevation angles for each of the 16 loudspeakers could be found.

The overall results determined that the best localization performance was accomplished by the HOA system but the best overall performance was accomplished by a FOA system (the Soundfield MKV). Elevation localization was poor for the MKV but it was no worse that the timbral performance of the Eigenmike.

This is important as it dispels the myth that HOA are inherently superior to FOA ones. In general, it was found that trade-offs between localization and audio quality must be taken into account when designing these systems.

In part II of this research the DPA 4006 monophonic reference microphone was replaced with the Sennheiser Ambeo. It was found that the Ambeo performed on par in terms of directionality with the Eigenmike. The localization performance of this system could be in part explained by its exceptional low-frequency directivity factor, as measured in Zalles et al. (2017).

Other notable works include Hemingson and Sarisky (2009), who also created a DIY microphone and compared it to industry standards. Due to the use of

different capsules by the author, which are more similar to conventional FOA arrays, in depth discussion for this study is not provided here.

SpHEAR

While simulation based evaluations are certainly helpful, once these simulations are acquired, it is important to objectively examine the quality of actual hardware. One project that seeks to help researchers evaluate ambisonic microphones comes from the Center for Computer Research in Music and Acoustics (CCRMA), in Stanford. (Lopez-Lezcano (2016)) describes a set of parametric computer-assisted designs (CAD) created with the open-source software OpenSCAD. These models can be used to 3D print ambisonic microphone shells for quick prototyping.

Lezcano's research allows independent researchers and university groups to investigate the effect of manipulating various parameters of sound field microphones with relative ease. 3D printers have, in the last few years, as a function of their ever decreasing price-point, become a staple of the open-source community. These printers use plastic filament to turn 3D rendered computer models and into solid objects. Using algorithms as a method for rendering these models allows us to change: capsules diameter, sphere radius or number of capsules, with incredible ease for faster prototyping.

This work is particularly important to increase availability and accessibility to ambisonics since it allows anyone with a 3D printer to make their own cases at home. The author also provided electronics schematics, calibration software and assembling instructions.

Unfortunately, due to the differences in geometry between ECMs and MEMS, we decided that rather than use Lezcano's designs it would be simpler to resize the old models created for our initial MEMS experiment (Zalles et al. (2017)). Part of the motivation to include this in our literature is to show the breadth of

researchers that are looking for affordable manufacturing solutions to increase availability of these sophisticated systems.

Prior Work by Author

In (Zalles et al. (2017)), a FOA microphone was constructed, quantitatively analyzed and subjectively evaluated. In this research, the authors involved with the work constructed a microphone using CAD 3D printed models and analog MEMS capsules. The dimensions of the tetrahedral array for that experiment were a function of the radius of our custom PCB, which had a larger radius than our new design (from 12.5mm to 6.35mm diameter). The initial radius of the PCB was mostly a function of ease of assembly. These dimensions were also loosely based on the dimensions the Sennheiser Ambeo, our point of comparison in both that experiment and this one.

The quantitative measurements for this first iteration of the project were done under anechoic conditions. A rotating platform, dubbed Automatic Rotating Microphone Mount, or ARM^2, integrated with ScanIR (Boren and Roginska (2011)), was developed and employed in the measurement process. In a similar fashion to previous research (Hemingson and Sarisky (2009)), a comparative evaluation of a professional and amateur microphone was performed.

In our case, the experiment was conducted via a survey and took advantage of a web-based ambisonic binaural decoder, which gave the authors the ability to deploy the assessment globally. While this was an effective solution, it restricted our ability to dictate experimental conditions such as noise-levels and reproduction methods. The findings were reported based on the type of binaural reproduction method used by subjects (headphones or earbuds) and any subject who had not used either was discarded.

Preliminary findings showed that while MEMS capsules were capable of quite aptly reproducing the sound field, their omnidirectional response, over multiple frequency bands, had a negative effect on accurate sound field reproduction. The

most salient quality of the capsules resulting in decreased performance, according to subjects, was the high-frequency boost above 10 kHz. This is a commonly reported characteristic of MEMS capsules which occurs due to the Helmholtz resonance of the semi-open system and is a function of inner dimensions of the capsule (Neumann Jr (2003)).

An added problem that we had in our evaluation was the degree of separation occurring when subjects are asked to move around the sound field using traditional computer interfaces, such as mouses or keyboards, as opposed to using their natural head movements. To alleviate this problem several authors have designed open-source head tracking solutions that can be employed in the assessment of binaural systems. (Romanov et al. (2017)), describes a low-cost, low-latency, high-quality head tracking system which works with open-source digital audio workstations (DAWs) such as Reaper. A similar system is used herein.

In the evaluation of our new microphone we opted against using the web-based decoder and survey. This was in part due to due to our desire to use a head-tracker enabling subjects to have a more natural testing experience, as well as limiting the number of independent variables such as noise levels and headphone model.

2.3 Summary

This chapter provided a review of the literature relevant to this thesis. The next chapter provides the framework and methodology to be used in the research project.

Chapter 3

Framework and Methodology

This chapter provides the framework and methodology to be used in the research study. For ease of reading this chapter will be separated into three sections. Section 3.1 will outline the design and assembly of our MEMS enabled FOA system. Section 3.2 will describe methods used for gathering of objective measurements, such as polar response plots and frequency response, and the results. Section 3.3 will detail the design of a subjective experiment.

3.1 Mic Design & Assembly

3.1.1 Capsule Selection

MEMS vs ECM

Despite having been around for over twenty years (Scheeper, Van der Donk, Olthuis, and Bergveld (1994)) MEMS microphones still remain unpopular among hi-fi recording hardware engineers. Over the course of that time their design and performance has increased dramatically in large part to adapt to the needs of consumers. Unfortunately, while their performance has become comparable to that of ECMs, MEMS have only really found success in mobile devices and home assistants.

Most modern ambisonic microphones, such as the Sennheiser Ambeo VR mic, or the designs proposed by Lopez-Lezcano (2016), rely on electret condenser microphones (ECMs) to capture soundfields. Electret condenser microphones are largely identical to MEMS microphones with the main difference being how much smaller MEMS capsules are, due to the manufacturing processes which are used to create them which occur at a microscopic level.

While ECMs do offer certain advantages, namely with respect to SNR, MEMS capsules are not only approximating SNR performance rapidly, but are also capable of offering a large number of additional advantages. Some of the advantages, described by Weigold, Brosnihan, Bergeron, and Zhang (2006), include:

- The ability to be surface mounted via a re-flow soldering process.
- Better performance density than ECMs[1].
- Less sensitivity to temperature variations.
- More uniform part-to-part frequency response than ECMs.

Another large advantage of MEMS systems, for companies developing hardware, is their ability to be more cost-effectively mass manufactured, reducing the individual cost per capsule at no loss to performance given these systems' reliable tolerances.

This part-to-part consistency makes them extremely attractive for arrays of any type in which a large number of sensors must all behave, ideally, exactly the same. Additional features constantly being monitored to ensure the quality of these systems are: SNR and sensitivity tolerance, all which affect the performance of a FOA array.

Finally, as pointed out by Alexandridis, Papadakis, Pavlidi, and Mouchtaris (2016b), their ability to include analog-to-digital converters (ADC) and amplifiers inside small packages make them extremely important for the future generation of

[1] performance density refers to performance with relation to overall size.

connected devices. Bit-streams from capsules within arrays, for example, can be multiplexed and transmitted wirelessly, providing consumers with a convenient way to listen and record high-definition spatial audio. Audio streams can also be sent to digital signal processing (DSP) chips which perform any desired modifications, such as filtering or phase shifting, to the incoming sounds. Kissner and Bitzer (2016) are among of the people who have proposed combining multiple MEMS in series to combat SNR deficits.

Picking a MEMS Microphone

Nowadays, there exist a number of companies in the MEMS market, all of which offer a variety of MEMS mic models. Each of these companies offer various selections of microphone capsules in both analog and digital packages. Some of these designs offer a bottom port design, while other offer a top port design. For this project a bottom ported design was used simply because the best performing MEMS capsule of the analog type. Figure 3.1 shows an example of each[2].

Figure 3.1.: Top Port & Bottom Port MEMS

A number of additional considerations were taken into account when selecting a capsule. The first regarded whether to use an analog or digital MEMS. Given the complexity of operating protocols associated with digital MEMS systems, such as I^2S or Pulse Density Modulation (PDM), and because analog MEMS systems offer slightly improved SNR [3], it was deemed preferable to select a MEMS

[2]Image 3.1 sourced from *EDN.com*
[3]Digital MEMS suffer from quantization noise.

element of the analog variety. Implementing digital MEMS systems is intended for future research. An analog MEMS system was selected due to it's improved SNR when compared to digital systems.

From the large number of possible analog MEMS available, the ICS-40720 was chosen. This was largely due to it being one of the best capsules of its type, matched only by the ICS-40730, which offers a 4dB SNR improvement (making it equivalent to the AT2020, a popular large diaphragm condenser microphone). The ICS-40720 provides a differential output, which translates into balanced signals, but can also be used in a single-ended mode. Differential outputs allow audio interfaces to apply common mode rejection (CMR) schemes to signals via associated circuitry. This technique takes advantage of the effects of noise, such as electro-magnetic interference (EMI), on bi-polar signals. Anything common to both signals will get cancelled at the receiving end via a differential amplifier (Gray, Hurst, Meyer, and Lewis (2001)).

3.1.2 PCB Design

Once the specific capsule to be used was selected, the specification sheet of the component was used to design the PCB. The PCB software Eagle, by Autodesk, was used for this process. The specification sheet provided the dimensions of the component as well as distance between pads. Figure 3.2 shows some of the specifications provided by the manufacturer[4].

Using this landing pattern and other specifications provided by the manufacturer, a number of Gerber files, containing all the instructions required to manufacture the PCB, was created. Four through-holes were added around the mounting location for the MEMS and connected via traces to the terminals of the ICS-40720. These four copper plated through holes would serve as soldering points for our cabling. Figure 3.3 shows an image from Eagle of the PCB's various layers including a silkscreen, board outline and solder mask.

[4]Figure 3.2 sourced from the *ICS-40720 Spec Sheet*

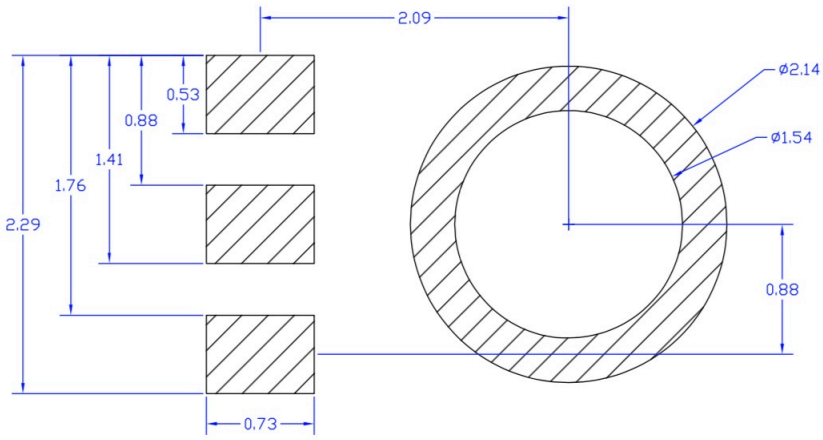

Figure 3.2.: ICS-40720 Landing Pattern

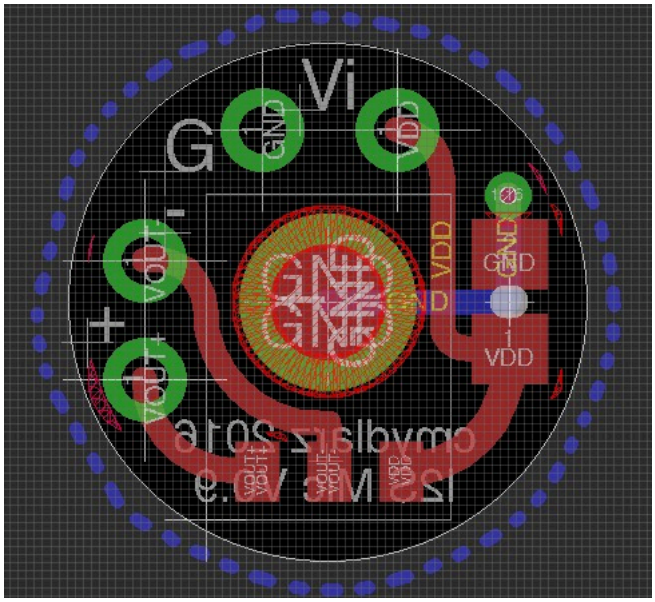

Figure 3.3.: 40720 PCB 6.35mm Diameter

The PCB design files, generally in Gerber format, an open ASCII[5] vector format for 2D binary images, consists of many layers which describe, for the manufacturer, how the board should be constructed. Today, it is common to use two layers of copper during PCB design since these have become as cheap to produce as single copper layer boards and allow for smaller PCB designs. The

[5]American Standard Code for Information Interchange

boards consist of: a relatively thick layer of fiber glass[6], copper traces, which can be additively or subtractively created, a solder mask, which protects the copper and prevents bridges, and a silkscreen, which is used to label components, allowing board assembly sans schematics. The PCB was miniaturized to allow greater proximity between capsules in the final design. The diameter of each PCB was reduced from 12.5mm to 6.35mm.

3.1.3 CAD Specifications and Requirements

In order to position the four PCBs in the appropriate dihedral angle of 70.53 degrees (Frank, Zotter, and Sontacchi (2015)), a CAD model was altered from our previous designs created on SolidWorks[7]. A number of additional considerations had to be taken into place when re-designing the model for this project.

In our previous work, (Zalles et al. (2017)), the final 3D printed model did not consider printing tolerances or have the necessary tensile strength[8]. This lead to difficulties when mounting the PCBs in place or even when fitting the microphone into a microphone clip, as the force used to push the microphone on the clip could result in snapping parts.

With the newer, smaller model, this problem was further exacerbated. Not only was the 3D printed model more fragile, but the PCBs had decreased in size. For these reasons, a different method of 3D printing was chosen for this project which could provide better resolution and durability.

During the creation of our first design, a stereolithographic approach to printing was employed via the Stratasys Mojo printer at NYU's 3D printing studio at LaGuardia Place. Initial prototypes of the MEMS soundfield mic version 2 were made using the same procedures but it became evident after a few trials that the print quality would not be sufficient for this project.

[6]Typically FR4, a flame retardant type of epoxy fiberglass.
[7]A CAD modeling software
[8]Material resistance to tension

After experimenting with a few different printers and talking to staff, the Form2 printer at NYU's MakerSpace in Brooklyn was selected. The Form2 uses Selective Laser Sintering (SLS), a rapid prototyping process, to allow generation of complex 3D parts by solidifying successive layers of powder material on top of each other (Kruth (1991)).

In order to do this, a mechanically controlled platform within the Form2 is lowered unto a pool of powder inside the printers build chamber. At each layer of the model the laser is instructed which points to sinter thus creating stacking cross sections of the model; the laser below the pool of liquid powder resin solidifies one cross-section of the model at a time. Post-processing the print, by soaking in alcohol and curing, ensures maximum performance. Figure 3.4 shows the Form2 in action.

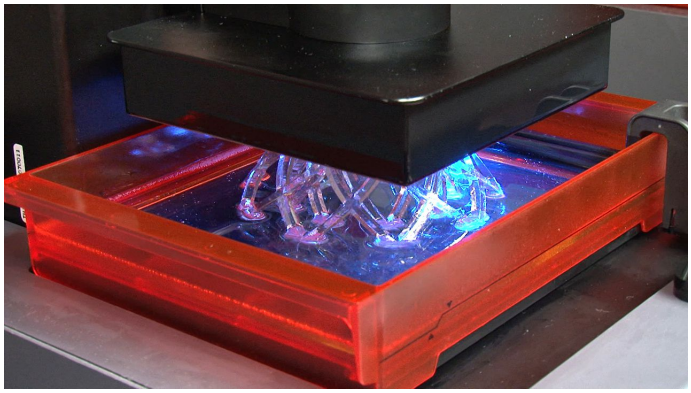

Figure 3.4.: Form2

3.2 Objective Measurements

During Zalles et al. (2017), a number of objective measurements were acquired to help further understand the specific interaction between hardware features and subjective feedback. In this section we report those measurements as well as other features of our MEMS system that might help the reader understand some of the differences between this system and others.

Polar Plots

Polar plots are a common specification of microphones. They describe the sensitivity of a microphone capsule to sounds arriving from a horizontal angle θ at specific frequencies. In order to create these polar plots a stepper motor combined with ScanIR and an Arduino UNO (plus a shield) were used. Namely, ScanIR was modified to allow for automatic measurements. The program sends a sine sweep through a speaker which is recorded by a single capsule, then commands the stepper to move the microphone 1.8°. A total of 200 steps gives us the sensitivity around 360°.

As one can observe from figure 3.5, the response of the microphone is for the most part omnidirectional. Above 4kHz the capsule begins to exhibit some directional characteristics due to the PCB blocking sounds arriving from behind. In contrast figure A.1 (found in the appendix) shows the Ambeo polar response, which features perfect cardioid response at multiple frequency bands.

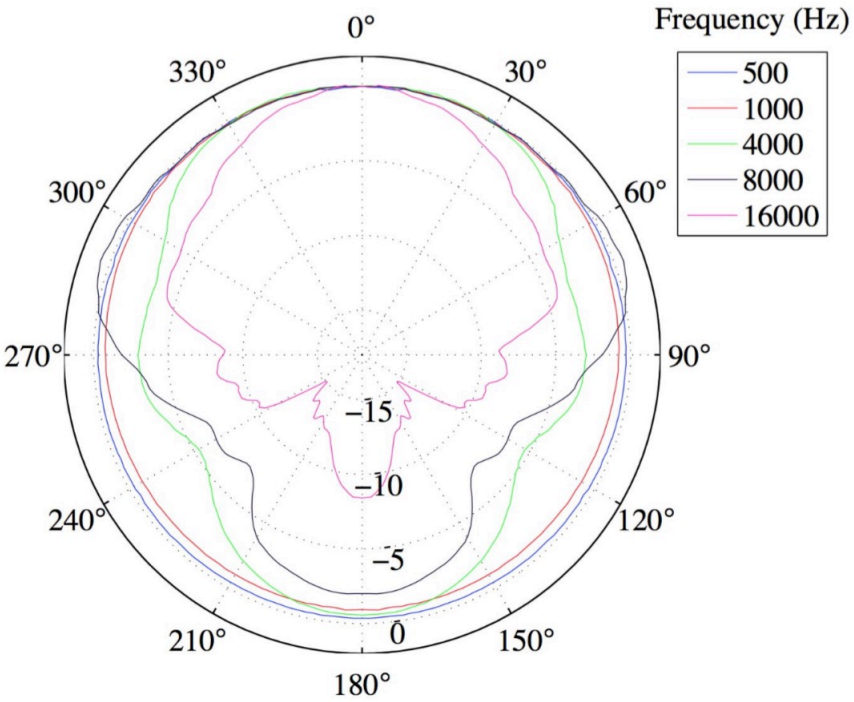

Figure 3.5.: ICS-40720 FOA Mic Polar Response

Frequency Response

As discovered in our previous research, one of the big problems of MEMS capsules is their exacerbated high-frequency response above 10kHz. This behavior is shown in the specification sheet of the ICS-40720 and is a known behavior of MEMS capsules. The enclosure which houses the amplifying circuit and transducer element of the MEMS capsule serves as a Faraday cage protecting the system against electromagnetic and radio interference. Unfortunately, due to it's design, this enclosure also creates resonances which cause the system to over-emphasize frequencies above 10kHz.

In order to mitigate the effect of the Helmholtz resonance created by the enclosure a Matlab filtering script was implemented. The filter was created using the Signal Processing toolbox in Matlab and consisted of a Finite Impulse Response (FIR) low-pass filter of order ten with a pass-band frequency of 10kHz.

Figure 3.6 shows the frequency response of the two capsules. The appendix contains additional figures depicting the response of the filter created and it's effect on the resulting output of the MEMS capsule.

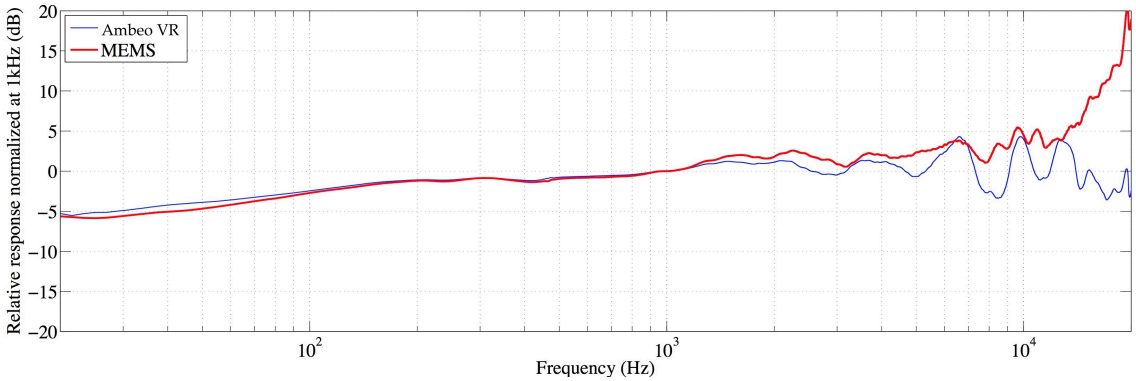

Figure 3.6.: MEMS Vs. Ambeo Frequency Response Raw

EIN, AOP and Dynamic Range

The Equivalent Input Noise (EIN) & Acoustic Overload Point (AOP) determine the effective dynamic range of each capsule. The calculated dynamic range of these systems was calculated given the available information provided by the manufacturers. It was determined that all in all the dynamic range difference was 12dB (with the Ambeo having a dynamic range of 112dB and the ICS-40720s having a dynamic range of 100dB).

As already mentioned, this great difference in dynamic range can be compensated for in part by adding the output of multiple MEMS elements in parallel.

It should be noted that the values presented here have not been verified objectively using measurements. The values reported are those provided by the manufacturer but Sennheiser does not clarify if the EIN provided is the total EIN or the per capsule EIN. Part of our future work will be to make measurements such as this as well as evaluating optimal MEMS configurations which reach the optimal balance between MEMS capsule coincidence and dynamic range.

3.3 Study Design

A subjective study was designed in order to evaluate the performance of our MEMS microphone. As reported by several authors across studies ((Minnaar, Olesen, Christensen, and Moller (2001)), (Noisternig et al. (2003)), (Mackensen et al. (2000))), an important factor for successful sound localization during binaural decoding is the use a head-trackers, which modify the sound scene based on the head movement. Part of the importance of head-trackers derives from the synchronized spectral cue modulations which occurs as function of the listener orientation helping one disambiguate hard-to-localize sources.

For this experiment, the participants were outfitted with a DIY[9] head-tracker designed by Tomasz Rudzki, Warsaw University of Technology graduate, and author of several papers for the Polish Academy of Sciences.

The head-tracker consists of a Arduino micro-controller and an MPU9250, an accelerometer and gyroscope unit that interfaces with the ATmega32U4 on board this micro-controller's breakout board. The firmware, containing the code necessary for the chip to interpret data sent from the MPU9250, was uploaded to the ATmega32U4. The resulting device was used to interface with the DAW Reaper, via the *OSC Bridge* MacOSX application Tomasz designed. A pair of DT990 PRO Beyerdynamic headphones were used to mount the head-tracker on. Figure 3.7 shows the head-tracker[10].

Figure 3.7.: NVSonic Head Tracker

The DAW Reaper was selected due to it being the only digital audio workstation (DAW) that allows communication via Open Sound Control, an TCP/IP protocol used by synthesizers and other musical hardware interfaces,

[9]Do-it-yourself
[10]Image 3.7 from *NVSonic.com*

created at CNMAT (Berkeley). OSC allows communication between the head-tracker and FB360 plug-in. The bridge is used to relay messages from the micro-controller to Reaper which uses yaw, pitch and roll information to modulate, in real-time, the output of each virtual speaker used for the binaural decoding of the FOA audio.

Despite the MPU950's 9-axis design, which allows for six degrees of freedom (6DoF), this particular head-tracker was configured to provide only yaw, pitch and roll information to the binaural decoder. Schörkhuber, Hack, Zaunschirm, Zotter, and Sontacchi (n.d.) explores the possibility of ambisonic networks and interpolating between soundfields. Systems such as this could extend ambisonics from three degrees of freedom (3DoF) to six degrees of freedom (6DoF). The three added degrees of freedom correspond to the possibility of forward, sideways and vertical movement of the entire body, irrespective of head movement.

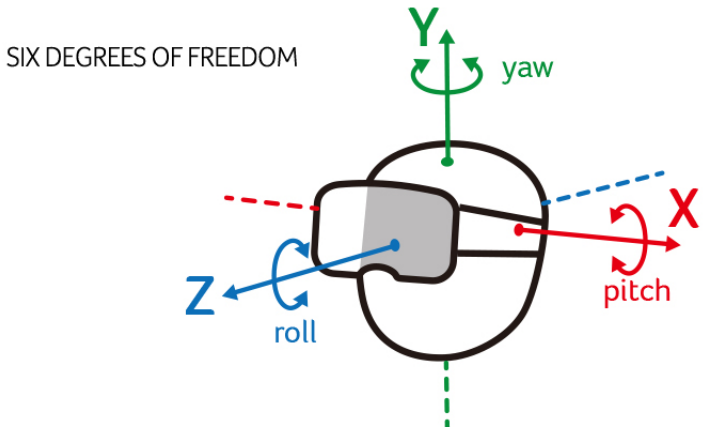

Figure 3.8.: Degrees of Freedom in VR.

Other researchers such as Gardner (1997) are exploring this problem from a transaural perspective. Cross-talk cancellation filters together with cameras tracking the location and orientation of the listeners head are already being sold by companies such as *Theoretica Applied Physics* as well. Authors such as Malham (1999) have researched how HOA can widen the, *sweet spot*, during ambisonic

reproduction, allowing partial or total sound field navigation with accurate or tolerable localization.

3.3.1 Hypotheses

Our main hypothesis concerned the effect of capsule coincidence in the MEMS FOA microphone, and whether it is possible to recreate a soundfield with greater spatial impact than other FOA arrays, which feature spaced capsules, despite SNR and polar response deficits in MEMS transducers. Namely, our hypothesis was that subjects would respond positively to the increase in coincidence and rate stimuli recorded with the MEMS system higher than the Ambeo.

While the independent variable in our experiment is the microphone used to record our stimuli, because participants were subjected to multiple stimuli, and asked to rate stimuli using a number of different attributes, we can also consider these as independent factors in our analysis. Attempts will be made to determine the interaction these addition factors had on subjects responses to the questions posed.

Six single-factor Analysis of Variance (ANOVAs) tests, as well as two 2-factor ANOVAs with replication were performed in order to provide deeper insight into responses provided by subjects. The ANOVA uses the distribution of data to determine, within a certain margin of error, the likelihood that two or more means are equivalent. Namely, if the reported F value reported is greater than the F-critical score, we will reject the null hypothesis and determine that the two means are different.

The null hypothesis is denoted as H_0. The null hypothesis for all ANOVAs is that the means of our groups are, statistically speaking, the same. The alternative H_α is that the means differ.

Six total single-factor ANOVAs were done for the first part of the analysis:

- **Total:** ANOVA comparing the overall score of both microphones by aggregating along stimuli and attribute.
- **Freedom from noise:** ANOVA comparing the freedom from noise attribute responses between microphones.
- **Dynamic range:** ANOVA comparing the dynamic range attribute responses between microphones.
- **Tonal quality:** ANOVA comparing the tonal quality attribute responses between microphones.
- **Overall quality:** ANOVA comparing the overall quality attribute responses between microphones.
- **Spatial impression:** ANOVA comparing the sptial impression attribute responses between microphones.

Two additional 2-factor Repeated Measures ANOVA were performed. Namely:

- **Microphone + Stimuli:** determining interactions between stimuli and microphone.
- **Microphone + Attribute:** determining interactions between attributes used and microphone.

3.3.2 Population

The population for this study consisted of 11 NYU students from the field of Music Technology. The students were all from different backgrounds, ages, genders and nationalities. Given the small subject pool we consider all subjects equally and did not sample of subgroup of this population as representative of a large group. It could be argued however that the population presented is representative of the

larger population of international audio students given the diversity of the NYU academic environment. Chapter B, in the appendix, features a number of additional figures regarding subjects' self reporting experience with VR, 3D audio and tendency to listen to music. All subjects reported healthy hearing.

3.3.3 Variables

Our independent variable is the microphone used during the re-recording of our stimuli. In order to create our 5 stimuli both microphones were situated in studio E at NYU in front of a reproduction system. Five songs were then played back from said system and re-recorded, one at a time, with the microphones at the same position and height. This is consistent with the approach taken by Zalles et al. (2017).

In contrast to our earlier research, a stereo system was used for the capture of our stimuli this time. During our former subjective study in 2017 a 5.1 listening set-up was used for the collection of stimuli. Part of the motivation behind this is to allow subjects to more discretely localize sounds during the experiment. A diffuse sound field, while potentially more enjoyable, would have less impact on spatial impression for subjects during decoding. Using a narrow spatial sound source range, it was conceived, that subjects would have an easier time noticing the effect of the binaural decoder on the recordings.

The dependent variable is the subject's responses to question presented during the experiment. The definition for the attributes presented are as follows:

- Freedom from noise: the lowest amount of noise is desired.
- Dynamic range: this is the range between the maximum level and the noise. The maximum level would be limited by comfort levels and the minimum by the ambient noise.
- Tonal quality: rich tonal quality is free from the distortions of peaks and dips in the response over frequency. Poor tonal quality has an uneven frequency

response that may cause certain notes to be lost and others to be unintentionally accentuated.

- Overall quality: [this attribute was not defined].
- Spatial Impression: a rating of envelopment and immersiveness.

The first three attributes were taken from McCarthy (2012). The fourth attribute was undefined and the fifth attribute was a self-created definition.

3.4 Summary

This chapter provided the framework and methodology to be used in the research study. The next chapter provides the Results & Analysis.

Chapter 4

Results & Analysis

For sake of structure and readability the section will be divided into Parts I and II. Part I will give the results of performing six single factor ANOVAs for each of the five attributes evaluated. Part II will give the results of two additional Repeated 2-Factor ANOVAs performed in order to evaluate any further effect stimuli or attributes had on subjects' responses. The α value used for all ANOVAs was 0.05. P values of 0.01 might be smaller than 1% but are reported as 0.01 for legibility of table data.

4.1 Part I - Single Factor ANOVAs

4.1.1 Total

Results from the ANOVA performed by aggregating scores accross all stimuli and attributes for each group reveals a very strong preference for the Ambeo system. The average score for the MEMS was found to be 5.3 while the average for the Ambeo was found to be 7. The following 5 ANOVAs will show the attribute-specific means and their significance. We start by reporting freedom from noise scores aggregating all stimuli.

Table 4.1: Anova: Single Factor - Total

Anova: Single Factor - Total

SUMMARY

Groups	Count	Sum	Average	Variance
MEMS	275	1451	5.276363636	3.60947578
Ambeo	275	1933	7.029090909	2.247325813

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	422.4072727	1	422.4072727	144.2450341	0.01	3.9
Within Groups	1604.763636	548	2.928400796			
Total	2027.170909	549				

4.1.2 Freedom From Noise

Table 4.2 gives the results for our freedom from noise question. As we can see the F-value derived was found to be extremely significant with a P-value rounded to 0%.

One possible way to explain this is the added sensitivity of our MEMS array. The balanced signal from each capsule corresponds a sensitivity of -38dBV. In contrast the Ambeo reports a sensitivity of -30dBV. One of the subjects in fact reported that he perceived the MEMS output as being more *honest*. This added sensitivity likely resulted in picking up much of the outside noise that bled into the room causing confusion among subjects who were incapable of separating self-noise and background noise in their scores.

Another possible explanation is the interference of electromagnetic and radio frequencies. The breakout board used to operate our design was unfortunately exposed to interference. We believe this might be part of the reason the self-noise of our system was more pronounced.

Finally, while fluctuations in direct current output from our battery supply were mitigated by including a 1uF capacitor in our PCB design, as recommended by

the manufacturer, the Power Supply Rejection Ratio (PSRR) might have been poorer than that of the Ambeo (no rating was found for this). Naturally, the higher EIN of the MEMS, caused by thermal noise limits, also meant that quieter sections of the stimuli were prone to noticeable self-noise created by the array in tandem with our interface (Behringer U-Phoria UMC404HD).

Table 4.2: Freedom From Noise - ANOVA - Single Factor

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
MEMS	55	232	4.218181818	1.914478114
Ambeo	55	372	6.763636364	2.48013468

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	178.1818182	1	178.1818182	81.09102053	0.01	3.929011484
Within Groups	237.3090909	108	2.197306397			
Total	415.4909091	109				

4.1.3 Dynamic Range

The results for this question were found to be very similar that those for freedom for noise. While we were expecting to find the exact same results it should be noted that the average for both group went up, despite the fact that noise and dynamic range are closely related. This leads us to believe that noise is often rated more harshly than dynamic range, an attribute that some might consider more subjective than noise.

Table 4.3: Dynamic Range - ANOVA - Single Factor

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
MEMS	55	272	4.945454545	1.45993266
Ambeo	55	388	7.054545455	2.163636364

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	122.3272727	1	122.3272727	67.51756179	0.01	3.929011484
Within Groups	195.6727273	108	1.811784512			
Total	318	109				

4.1.4 Tonal Quality

The tonal quality responses do not seem to match what was expected from the specifications from the manufacturer. While both microphones have extended frequency response (up to 20kHz), and despite the MEMS system being equalized for Helmholtz resonance, the F score shows that the Ambeo system outperforms the MEMS array with statistically significant results. A possible reason for this is that the tonal quality of the MEMS system was degraded due to the EIN which the MEMS suffer from. Follow-up questions with participants revealed that indeed, the noise from the capsules degraded their perception of tonal quality. Other subjects reported that the low frequency response of the MEMS was inadequate, which might be explained by it's more limited frequency response (75hZ-20kHz). Another possible explanation could be the non-linear nature of noise caused by the PSRR found in these capsules which caused high-frequencies to be louder than lower ones.

Table 4.4: Tonal Quality - ANOVA - Single Factor

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
MEMS	55	264	4.8	3.792592593
Ambeo	55	387	7.036363636	1.628282828

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	137.5363636	1	137.5363636	50.74322981	0.01	3.929011484
Within Groups	292.7272727	108	2.71043771			
Total	430.2636364	109				

4.1.5 Overall Quality

Extremely significant results were found for the overall quality question. This is no surprising as we could have inferred these results based on responses from tables 4.2, 4.3 & 4.4. These criteria naturally had a profound impact on the way subjects rated these systems in terms of overall quality. It should be noted that possible improvements in score might have been found had overall quality been defined as including spatial panning attributes of the microphones, something which is often disregarded by listeners due to the rarity of subjective experiments involving head-tracked ambisonic audio.

Table 4.5: Overall Quality - ANOVA - Single Factor

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
MEMS	55	255	4.636363636	2.346801347
Ambeo	55	399	7.254545455	1.600673401

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	188.5090909	1	188.5090909	95.5087001	0.01	3.929011484
Within Groups	213.1636364	108	1.973737374			
Total	401.6727273	109				

4.1.6 Spatial Impression

The results from the spatial impression question are perhaps the most promising. Despite the imperfections in directionality exhibited by the MEMS capsules, as seen in their polar plot, subjects noted that the localization effect of the system was extremely noticeable, even to an unrealistic degree. The results from our analysis show that this response had the closest match in performance out of all the attributes, further emphasizing the point that designing these systems with this type of capsule requires a compromise between localization performance and overall sound quality.

Some subjects reported that they rated the MEMS system lower on the 10 point scale because their overall sense of envelopment was lower as a result to the extreme localization of sources. This perhaps could also be explained by the diffuse to free-field response differences of both systems (the analysis of which is also the subject of future work). It is possible that while the directionality of the system was very good, the overall enveloping effect was lost due to the systems inability to pick up on reverberation, due to it's poorer dynamic range.

Table 4.6: Spatial Impression - ANOVA - Single Factor

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
MEMS	55	298	5.418181818	4.173737374
Ambeo	55	374	6.8	2.459259259

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	52.50909091	1	52.50909091	15.83269036	0.01	3.929011484
Within Groups	358.1818182	108	3.316498316			
Total	410.6909091	109				

4.2 Results - Part II

4.2.1 Stimulus + Microphone Repeated Measures ANOVA

The column F-values on table 4.7 show that the mean values are not the same, demonstrating that the stimuli had an effect on the results. From the sums of responses we notice that Stimulus V had the best scores. This can be explained by the genre of the music for stimuli V, which was of the jazz/popular nature, as opposed as the other stimuli which were classical or jazz. A more in depth look into Stimuli V showed that there was no statistically significant difference in means for the spatial impression question (the p value there was .5) when looking only at stimuli V. This reveals that the performance of our system is condition-dependent.

One possible explanation for this could be the lower dynamic range found in popular music. Given the poorer dynamic range of the MEMS system is was appropriate to record this genre with this microphone. All other stimuli, unfortunately, exhibited some noticeable noise which caused the scores of the system to drop.

Further analysis of stimuli V showed that the MEMS system performed on par with the Ambeo in questions I and III (p = .06, p = .2). This further demonstrates that subjects considered noise as an independent measure of quality. Despite equivalent means the overall quality mean values were found to be different with a p = 0 value.

Table 4.7: Stimuli + Microphone Repeated Measures ANOVA

SUMMARY	Stimulus I	Stimulus II	Stimulus III	Stimulus IV	Stimulus V	Total
MEMS						
Count	55.00	55.00	55.00	55.00	55.00	275.00
Sum	267.00	238.00	263.00	258.00	295.00	1,321.00
Average	4.85	4.33	4.78	4.69	5.36	4.80
Variance	2.65	3.26	2.77	2.48	2.75	2.85
Ambeo						
Count	55.00	55.00	55.00	55.00	55.00	275.00
Sum	386.00	364.00	377.00	378.00	415.00	1,920.00
Average	7.02	6.62	6.85	6.87	7.55	6.98
Variance	1.57	2.24	2.27	1.96	1.96	2.07
Total						
Count	110.00	110.00	110.00	110.00	110.00	
Sum	653.00	602.00	640.00	636.00	710.00	
Average	5.94	5.47	5.82	5.78	6.45	
Variance	3.27	4.05	3.58	3.40	3.53	
ANOVA						
	SS	df	MS	F	P-value	F crit
Sample	652.37	1.00	652.37	272.80	0.01	3.86
Columns	56.30	4.00	14.07	5.89	0.01	2.39
Interaction	0.66	4.00	0.17	0.07	0.99	2.39
Within	1,291.35	540.00	2.39			
Total	2,000.67	549.00				

Table 4.8: Question + Microphone Repeated Measures ANOVA

Anova: Two-Factor With Replication

SUMMARY	Q1	Q2	Q3	Q4	Q5	Total
MEMS						
Count	55.00	55.00	55.00	55.00	55.00	275.00
Sum	273.00	286.00	291.00	279.00	322.00	1,451.00
Average	4.96	5.20	5.29	5.07	5.85	5.28
Variance	3.48	2.72	4.47	3.29	3.87	3.61
Ambeo						
Count	55.00	55.00	55.00	55.00	55.00	275.00
Sum	366.00	395.00	390.00	403.00	379.00	1,933.00
Average	6.65	7.18	7.09	7.33	6.89	7.03
Variance	3.19	2.19	1.60	1.71	2.43	2.25
Total						
Count	110.00	110.00	110.00	110.00	110.00	
Sum	639.00	681.00	681.00	682.00	701.00	
Average	5.81	6.19	6.19	6.20	6.37	
Variance	4.03	3.42	3.83	3.76	3.39	

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Sample	422.41	1.00	422.41	145.92	0.01	3.86
Columns	18.88	4.00	4.72	1.63	0.17	2.39
Interaction	22.65	4.00	5.66	1.96	0.10	2.39
Within	1,563.24	540.00	2.89			
Total	2,027.17	549.00				

4.2.2 Question + Microphone Repeated Measures ANOVA

The values in the columns row of the ANOVA analysis show that the question had no effect on the averages. This shows that no definition or attribute was rated more highly or lower than others.

Figure 4.1 shows a Box Plot with the upper and lower inter-quartile ranges for each question and group. Figure 4.2 shows the means and variance of the microphone's performance divided by stimuli.

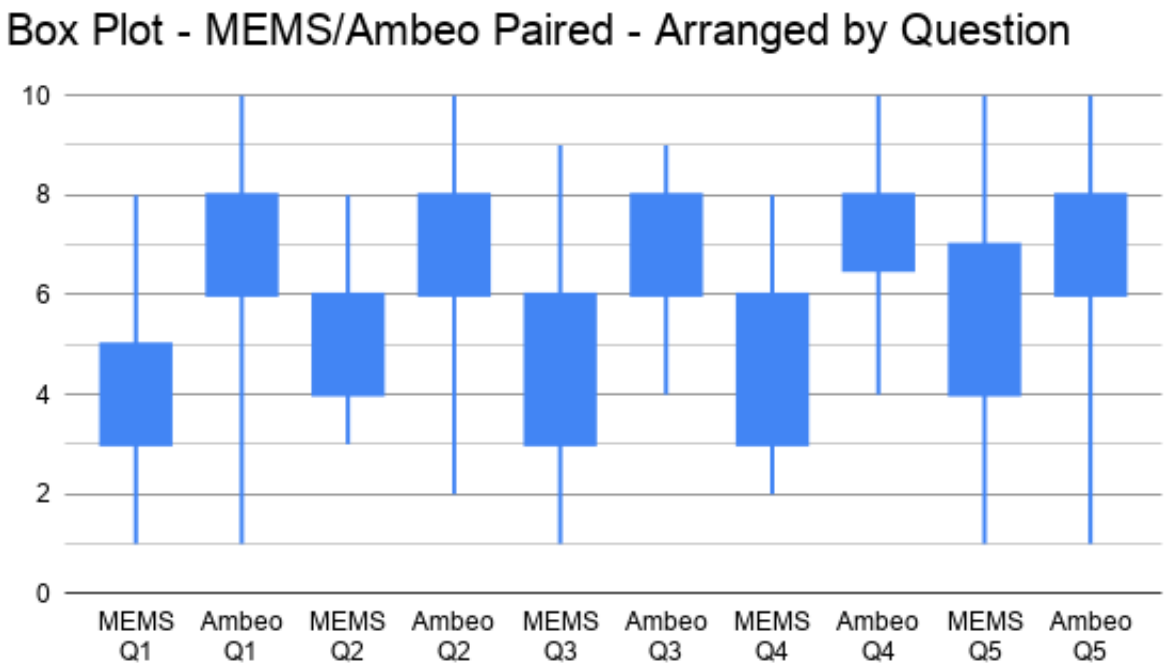

Figure 4.1.: Inter-quartile Range - Box Plot

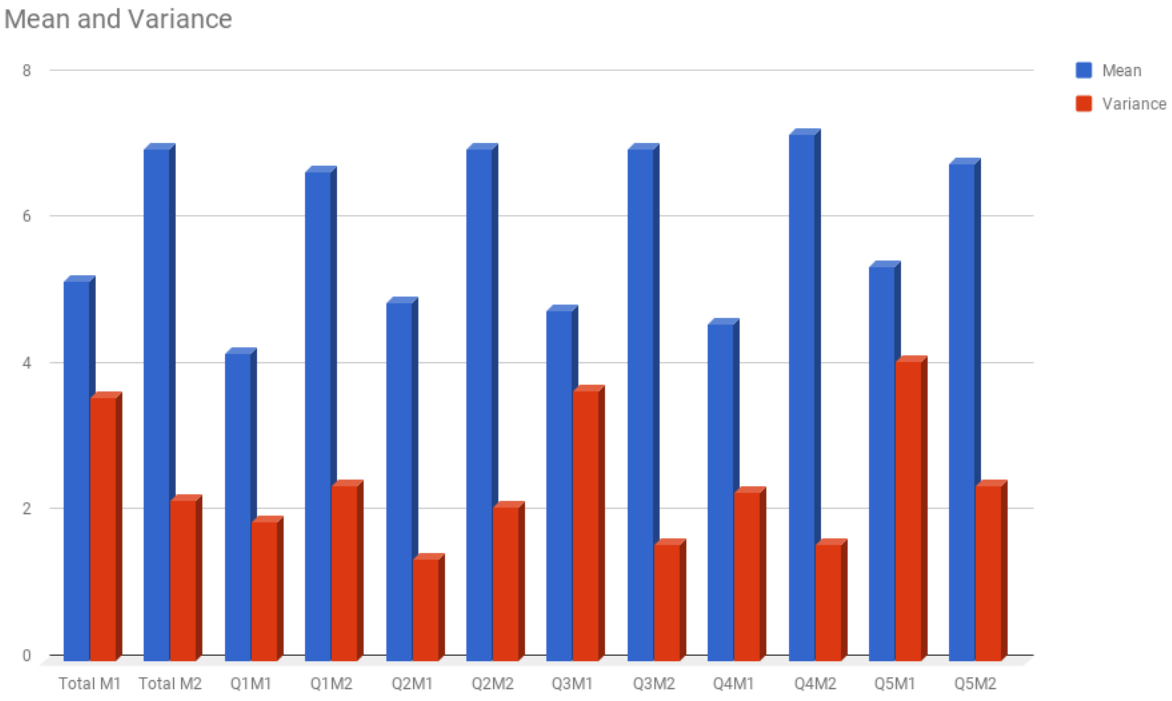

Figure 4.2.: Means/Variance - Bar Graph

Chapter 5

Discussion

This thesis evaluated the efficacy of MEMS transducers as replacements for ECMs in FOA arrays. The microphone constructed was subjectively evaluated, and objective calculations and measurements were provided to help understand the results of our experiment. A number of statistical analyses were performed on subjective responses given by 11 participants which revealed trade-offs between dynamic range and attributes of envelopment.

The constructed system, which employed four ICS-40720s, was shown to have intense localization attributes despite yielding omnidirectional polar capsules. The results of polar plot measurements gathered for a previous experiment revealed a semi-directional capsule response above a limiting frequency. The new microphone system was not measured due to the very similar nature of these two devices. Polar response measurements for this new design are the subject of future work.

The result of 6 single factor ANOVAs with two groups, equivalent to a T-test, showed that the Ambeo outperformed the MEMS system in every category. The spatial impression question showed the closest performance between both systems with only a 10% increase in performance in the Ambeo. Subjects reported that the extreme panning/localization experienced with the MEMS system was too aggressive to be considered realistic. Future work will examine the localization

performance of this system to determine if extreme capsule coincidence, resulting in an f_{err} above 20kHz actually degrade the ability to accurately localize sources.

Additional repeated measure ANOVAs showed that while questions had no effect or interaction for groups, the stimuli affected the results of our analysis. A deeper look into Stimuli V showed that the MEMS performed on par with the Ambeo in three out of five categories. This leads us to believe that stimuli featured limited dynamic range skew the results towards system incapable of handling these greater dynamic ranges.

In conclusion, the MEMS system was indeed able to spatialize sources even though the capsules were not cardioid as the literature specifies. It was also considered to be more *honest* due to it's sensitivity wish greater than that of the Ambeo VR Mic. Despite these features the limited SNR of the system caused the tonal and overall quality of the system to be rated far lower than expected. As mentioned before, part of the discrepancy could be explained by the lack of isolation in the Studio where the stimuli were recorded.

Future work will attempt to re-evaluate this system using MaxMSP as a GUI for randomized testing in conjunction with out head-tracker. Stimuli should be recorded in a quieter setting and de-noising algorithms can be employed to determnie the effective performance of these systems under these new conditions. Measurements involving the frequency response of 0 and 1 order spherical harmonics will also be performed, as well as directivity effects at various frequency bands.

APPENDICES

Appendix A

Additional Figures - Objective Measurements

Figure A.1 shows shows the polar response of one of the Ambeo VR mic capsules.

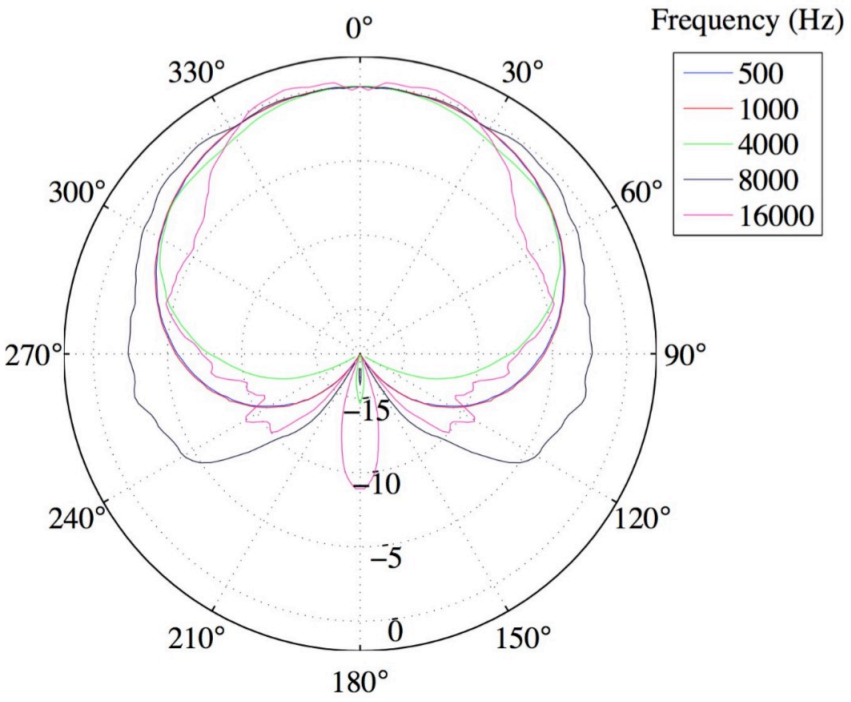

Figure A.1.: Sennheiser Ambeo VR Mic Polar Response - Single Capsule

Return to section 3.2.

Figure A.2 shows the frequency response of the MEMS before and after filtering and the response of the Ambeo mic overlayed.

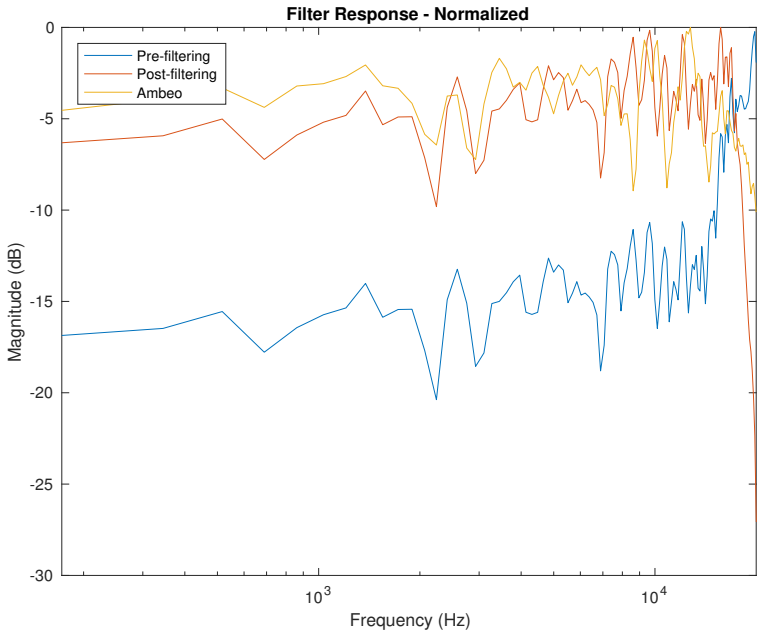

Figure A.2.: Frequency Response Pre/Post Filtering Vs. Ambeo

Appendix B

Additional Figures - Population

How old are you?

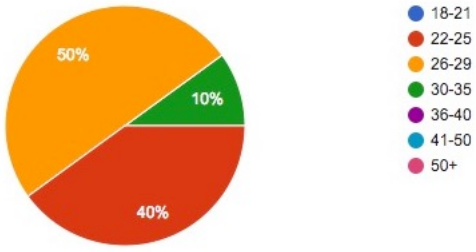

Figure B.1.: Questionnaire - Age - Responses

How many hours of music do you listen a day?

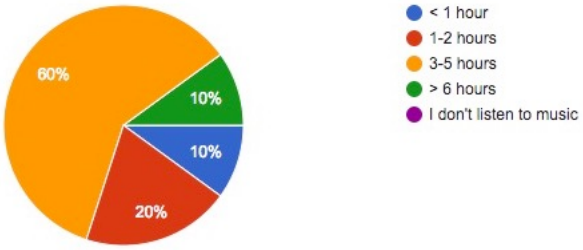

Figure B.2.: Questionnaire - Hours Music Per Day - Responses

Do you have experience with VR?

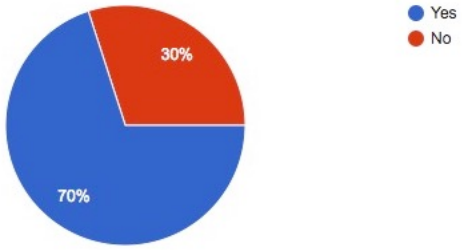

Figure B.3.: Questionnaire - Experience VR - Responses

Do you have experience with 3D audio?

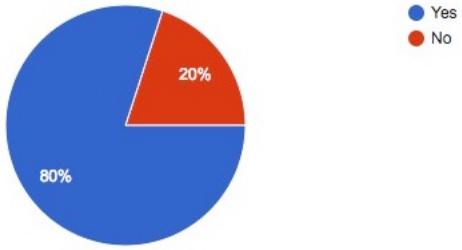

Figure B.4.: Questionnaire - Experience 3D Audio - Responses

Appendix C

Consent Form

Agreement to Participate

Consent Form for IRB Study #IRB-FY2018-1505

You have been invited to take part in a research study regarding the applicability of micro-electronic capsules in virtual reality microphone arrays. This study will be conducted by Gabriel Zalles, STEINHARDT - Music & Performing Arts Professions, Steinhardt School of Culture, Education, and Human Development, New York University, as a part of his Master's Thesis. Their faculty sponsor is Professor Agnieszka Roginska, Department of STEINHARDT - Music & Performing Arts Professions, Steinhardt School of Culture, Education, and Human Development, New York University.

If you agree to be in this study, you will be asked to do the following: —Evaluate the quality of various recording on multiple features. —Complete a questionnaire about your background (age, gender, education, etc.).

Participation in this study will take 30 minutes.

There are no known risks associated with your participation in this research beyond those of everyday life.

Although you will receive no direct benefits, this research may help the investigator understand the applicability of micro-electronic capsules in virtual reality microphone arrays.

Confidentiality of your research records will be strictly maintained by assigning code numbers to each participant so that data is never directly linked to individual identity.

If there is anything about the study or your participation that is unclear or that you do not understand, if you have questions or wish to report a research-related problem, you may contact Agnieszka Roginska at (212) 998-4500, roginska@nyu.edu, 82 Washington Square E, New York, NY 10003, or the faculty sponsor, Agnieszka Roginska at (212)998-5141, roginska@nyu.edu, 82 Washington Square E, New York, NY 10003.

For questions about your rights as a research participant, you may contact the University Committee on Activities Involving Human Subjects, New York University, 665 Broadway, Suite 804, New York, New York, 10012, at ask.humansubjects@nyu.edu or (212) 998-4808. Please reference the study # (IRB-FY2018-1505) when contacting the IRB.

You may request a copy of this consent document to keep.

LIST OF REFERENCES

LIST OF REFERENCES

Alexander, R. (2013). *The inventor of stereo: The life and works of alan dower blumlein*. Focal Press.

Alexandridis, A., Papadakis, S., Pavlidi, D., & Mouchtaris, A. (2016a). Development and evaluation of a digital mems microphone array for spatial audio. *2016 24th European Signal Processing Conference (EUSIPCO)*, 612-616.

Alexandridis, A., Papadakis, S., Pavlidi, D., & Mouchtaris, A. (2016b). Development and evaluation of a digital mems microphone array for spatial audio. In *Signal processing conference (eusipco), 2016 24th european* (pp. 612–616).

Algazi, V. R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001). The cipic hrtf database. In *Applications of signal processing to audio and acoustics, 2001 ieee workshop on the* (pp. 99–102).

Backman, J. (2006, May). Miniature microphone arrays for multi-channel recording. In *Audio engineering society convention 120*. Retrieved from http://www.aes.org/e-lib/browse.cfm?elib=13485

Bates, E., Dooney, S., Gorzel, M., O'Dwyer, H., Ferguson, L., & Boland, F. M. (2017). Comparing ambisonic microphonespart 2. In *Audio engineering society convention 142.*

Benjamin, E., & Chen, T. (2005). The native b-format microphone. In *Audio engineering society convention 119.*

Benjamin, E. M. (2012). A second-order soundfield microphone with improved polar pattern shape. In *Audio engineering society convention 133.*

Bernschütz, B. (2013). A spherical far field hrir/hrtf compilation of the neumann ku 100. In *Proceedings of the 40th italian (aia) annual conference on acoustics and the 39th german annual conference on acoustics (daga) conference on acoustics* (p. 29).

Boren, B., & Roginska, A. (2011). Multichannel impulse response measurement in matlab. In *Audio engineering society convention 131.*

Burdea Grigore, C., & Coiffet, P. (1994). *Virtual reality technology*. London: Wiley-Interscience.

Chanan, M. (1995). *Repeated takes: A short history of recording and its effects on music*. Verso.

Choueiri, E. Y. (2008). Optimal crosstalk cancellation for binaural audio with two loudspeakers. *Princeton University*, 28.

Dabin, M., Ritz, C., & Shujau, M. (2015). Design and analysis of miniature and three tiered b-format microphones manufactured using 3d printing. In *Acoustics, speech and signal processing (icassp), 2015 ieee international conference on* (pp. 2674–2678).

De la Peña, N., Weil, P., Llobera, J., Giannopoulos, E., Pomés, A., Spanlang, B., ... Slater, M. (2010). Immersive journalism: immersive virtual reality for the first-person experience of news. *Presence: Teleoperators and virtual environments*, *19*(4), 291–301.

Dooley, W. L., & Streicher, R. D. (1982). Ms stereo: a powerful technique for working in stereo. *Journal of the Audio Engineering Society*, *30*(10), 707–718.

Duraiswami, R., Davis, L., Shamma, S. A., Elman, H. C., Duda, R. O., Algazi, V. R., ... Raveendra, S. (2000). Individualized hrtfs using computer vision and computational acoustics. *The Journal of the Acoustical Society of America*, *108*(5), 2597–2597.

Frank, M., Zotter, F., & Sontacchi, A. (2015). Producing 3d audio in ambisonics. In *Audio engineering society conference: 57th international conference: The future of audio entertainment technology–cinema, television and the internet.*

Gallagher, A. G., Ritter, E. M., Champion, H., Higgins, G., Fried, M. P., Moses, G., ... Satava, R. M. (2005). Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. *Annals of surgery*, *241*(2), 364.

Gardner, W. G. (1997). Head tracked 3-d audio using loudspeakers. In *Applications of signal processing to audio and acoustics, 1997. 1997 ieee assp workshop on* (pp. 4–pp).

Geluso, P. (2012). Capturing height: The addition of z microphones to stereo and surround microphone arrays. In *Audio engineering society convention 132.*

Gerzon, M. A. (1973). Periphony: With-height sound reproduction. *Journal of the Audio Engineering Society*, *21*(1), 2–10.

Gerzon, M. A. (1974). Surround-sound psychoacoustics. *Wireless World*, *80*(1468), 483–486.

Gerzon, M. A. (1975). The design of precisely coincident microphone arrays for stereo and surround sound. In *Audio engineering society convention 50.*

Gerzon, M. A. (1980). Practical periphony: The reproduction of full-sphere sound. In *Audio engineering society convention 65.*

Gray, P. R., Hurst, P., Meyer, R. G., & Lewis, S. (2001). *Analysis and design of analog integrated circuits*. Wiley.

Grosvenor, E. S., & Wesson, M. (2016). *Alexander graham*. New Word City.

Heller, A., Lee, R., & Benjamin, E. (2008). Is my decoder ambisonic? In *Audio engineering society convention 125.*

Hemingson, D., & Sarisky, M. (2009). A practical comparison of three tetrahedral ambisonic microphones. In *Audio engineering society convention 126.*

Hollerweger, F. (2005). an introduction to higher order ambisonic . *Florian Hollerwegers Website*.

Hu, H., Zhou, L., Ma, H., & Wu, Z. (2008). Hrtf personalization based on artificial neural network in individual virtual auditory space. *Applied Acoustics*, *69*(2), 163–172.

Kaufmann, H., Schmalstieg, D., & Wagner, M. (2000). Construct3d: a virtual reality application for mathematics and geometry education. *Education and information technologies*, *5*(4), 263–276.

Kim, B.-H., & Lee, H.-S. (2015). Acoustical-thermal noise in a capacitive mems microphone. *IEEE Sensors Journal*, *15*(12), 6853–6860.

Kissner, S., & Bitzer, J. (2016). Analysis of current mems microphones for cost-effective microphone arraysa practical approach. In *Audio engineering society convention 140.*

Kruth, J.-P. (1991). Material incress manufacturing by rapid prototyping techniques. *CIRP Annals-Manufacturing Technology*, *40*(2), 603–614.

Lamson, R. J. (2002, July 30). *Virtual reality immersion therapy for treating psychological, psychiatric, medical, educational and self-help problems.* Google Patents. (US Patent 6,425,764)

Lopez-Lezcano, F. (2016). The *sphear project, a family of parametric 3d printed soundfield microphone arrays. In *Conference on sound field control.*

Mackensen, P., Fruhmann, M., Thanner, M., Theile, G., Horbach, U., & Karamustafaoglu, A. (2000). Head tracker-based auralization systems: Additional consideration of vertical head movements. In *Audio engineering society convention 108.*

Malham, D. G. (1999). Higher order ambisonic systems for the spatialisation of sound. In *Icmc.*

McCarthy, B. (2012). *Sound systems: design and optimization: modern techniques and tools for sound system design and alignment*. CRC Press.

McKeag, A., & McGrath, D. S. (1996). Sound field format to binaural decoder with head tracking. In *Audio engineering society convention 6r.*

Mennecke, B., Roche, E., Bray, D., Konsynski, B., Lester, J., Rowe, M., & Townsend, A. (2007). Second life and other virtual worlds: A roadmap for research.

Minnaar, P., Olesen, S. K., Christensen, F., & Moller, H. (2001). The importance of head movements for binaural room synthesis..

Neukom, M. (2007, Oct). Ambisonic panning. In *Audio engineering society convention 123.* Retrieved from http://www.aes.org/e-lib/browse.cfm?elib=14354

Neumann Jr, J. J. (2003). Mems (microelectromechanical systems) audio devices-dreams and realities. In *Audio engineering society convention 115.*

Noisternig, M., Musil, T., Sontacchi, A., & Holdrich, R. (2003). 3d binaural sound reproduction using a virtual ambisonic approach. In *Virtual environments, human-computer interfaces and measurement systems, 2003. vecims'03. 2003 ieee international symposium on* (pp. 174–178).

Ortolani, F. (2015). Introduction to ambisonics. *Ironbridge Electronics*.

Parsons, T. D., & Rizzo, A. A. (2008). Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis. *Journal of behavior therapy and experimental psychiatry*, *39*(3), 250–261.

Rizzo, A. (2002). Virtual reality and disability: emergence and challenge. *Disability and rehabilitation*, *24*(11-12), 567–569.

Rizzo, A. A., Buckwalter, J. G., Bowerly, T., Van Der Zaag, C., Humphrey, L., Neumann, U., ... Sisemore, D. (2000). The virtual classroom: a virtual reality environment for the assessment and rehabilitation of attention deficits. *CyberPsychology & Behavior*, *3*(3), 483–499.

Romanov, M., Berghold, P., Frank, M., Rudrich, D., Zaunschirm, M., & Zotter, F. (2017). Implementation and evaluation of a low-cost headtracker for binaural synthesis. In *Audio engineering society convention 142.*

Scheeper, P., Van der Donk, A., Olthuis, W., & Bergveld, P. (1994). A review of silicon microphones. *Sensors and Actuators A: Physical*, *44*(1), 1–11.

Schörkhuber, C., Hack, P., Zaunschirm, M., Zotter, F., & Sontacchi, A. (n.d.). Localization of multiple acoustic sources with a distributed array of unsynchronized first-order ambisonics microphones.

Sontacchi, A., Noisternig, M., Majdak, P., & Holdrich, R. (2002a). An objective model of localisation in binaural sound reproduction systems. In *Audio engineering society conference: 21st international conference: Architectural acoustics and sound reinforcement.*

Sontacchi, A., Noisternig, M., Majdak, P., & Holdrich, R. (2002b). Subjective validation of perception properties in binaural sound reproduction systems. In *Audio engineering society conference: 21st international conference: Architectural acoustics and sound reinforcement.*

Spors, S., & Ahrens, J. (2007). Comparison of higher-order ambisonics and wave field synthesis with respect to spatial aliasing artifacts. In *In 19th international congress on acoustics.*

Theile, G. (2001). Multichannel natural music recording based on psychoacoustic principles. In *Aes 19 th international conference.*

Thornton, S. (2009). *Michael gerzon - audio pioneer.*

Vinkel, S. P. (2017). *Exploration of first-orderambisonics usage in vr concertexperiences.*

Weigold, J., Brosnihan, T., Bergeron, J., & Zhang, X. (2006). A mems condenser microphone for consumer applications. In *Micro electro mechanical systems, 2006. mems 2006 istanbul. 19th ieee international conference on* (pp. 86–89).

Zalles, G., Kamel, Y., Anderson, I., Lee, M. Y., Neil, C., Henry, M., ... Roginska, A. (2017). A low-cost, high-quality mems ambisonic microphone. In *Audio engineering society convention 143.*

Zotter, F. (2009). Sampling strategies for acoustic holography/holophony on the sphere. *NAG-DAGA, Rotterdam*, 1–4.

Zvonar, R. (1999). A history of spatial music. *Montreal: CEC*.