CAPTURING SPATIAL AUDIO FROM ARBITRARY MICROPHONE ARRAYS FOR BINAURAL REPRODUCTION

BY

MICHAEL FRIEDMAN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Adviser:

Professor Douglas Jones

ABSTRACT

Spatial audio has typically been recorded from specialized microphone arrays that are too expensive and unwieldy to incorporate into today's consumer devices. Consumer devices, such as mobile devices, typically use cheap microphones with strict limitations on array geometry. Therefore, in order to capture spatial audio on such a device, we must be able to work with the given device geometry. To make spatial audio possible on a wide range of devices, a method for capturing the best spatial audio from an arbitrary array geometry is needed. In this thesis we propose several methods for capturing spatial audio from an arbitrary array for reproduction via headphones or binaural cross-talk cancellation. A technique for designing filters that minimize the reconstruction error of the soundfield captured at the array relative to the head related transfer function is described. Our techniques are compared with the current state of the art in spatial audio, Ambisonic recording and reproduction. Additionally, case studies of several microphone arrangements capable of fitting a mobile device geometry are examined. Their efficacy for use in a spatial audio system is discussed. It is demonstrated that such restricted geometries are capable of capturing compelling spatial audio. In addition, it is shown that given the reconstruction techniques proposed in this thesis, performance is equal to Ambisonics when an Ambisonic array is used, and potentially superior to Ambisonics when a more flexible array is employed.

To my parents, for their unconditional support.

ACKNOWLEDGMENTS

I would like to sincerely thank Research In Motion (now Blackberry) for their financial support of my research. Working with Blackberry opened my eyes to the practical implications of my work, which I may not have seen otherwise.

I would also like to thank my adviser, Doug Jones, for bringing me into his lab, where I have had the opportunity to learn and make many new friends. His direction of my work has helped to shape my own approach to solving problems.

A final thank you to my labmates, who have been a constant source of discussion, exposing me to new ideas and helping me get out of the box when I am stuck.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Motivation

With the introduction of the MEMS microphone, there has been a growing interest in employing multi-microphone techniques on today's mobile devices. While many of these techniques are dedicated to speech and call quality, such as environmental noise reduction and echo cancellation, there is an interest in expanding the entertainment capabilities of mobile devices as well. One such application is recording and playing back spatial audio.

Spatial audio is the presentation of sound over speakers or headphones such that the listener perceives the sound as located in space. In this thesis we will examine methods for recording and reproducing spatial audio on a mobile platform. With an unprecedented ability for user generation of multimedia content, the mobile platform represents a logical choice for expanding the spatial audio user base. Traditional spatial audio recording has been the domain of scientists and hobbyists with specialized hardware. Spatial recording on a mobile device would be an excellent vehicle for advancing the technology by bringing attention from a much larger audience. Mobile devices also present a unique set of challenges that must be overcome in order to create a successful mobile spatial audio system, and thus this work is an attempt at addressing some of them.

The demand for an immersive experience is one of the driving forces in current audio technology. The first stab at this was stereophonic recording and playback. With two speakers and something as simple as a recording of a ping-pong ball, users could experience a sense of "being there" like never before. As the idea of immersion in audio evolved, the natural goal became the recreation of a real auditory scene. The idea is to recreate as accurately as possible the experience of being at the location of the microphone during

recording. Certainly the experienced listener can attest that while stereo could hint at this, it did not have the ability to recreate a full 3D experience, such as a live performance in a concert hall. While there was a resurgence in immersive audio with 5.1 and 7.1 home theater, various limitations have led to the systems being used for the creation of artificial scenes, such as those mixed by hand for movies, and subsequently a spatially accurate audio recreation has not achieved popularity.

The usefulness of spatial sound is not simply limited to the reproduction of music or the soundtrack of a Hollywood movie; it can also enhance everyday experiences such as home movies, or the multitude of user created videos on the Internet. Mobile devices are a good host for spatial audio arrays because they could allow users to create their own spatial audio content. Existing methods for content creation have been limited to professional users. Creating a synthetic audio scene requires a painstaking process of placing audio sources in a virtual environment and controlling their movement, a process that is akin to animation. Recording real spatial audio scenes, on the other hand, is relatively simple, but requires expensive, specialized microphone arrays and the ability to record many audio channels at once. The integration of inexpensive MEMS elements into a mobile device, however, would solve both the cost and the recording issues.

Recreating an environment in a room with speakers is challenging since we cannot undo the effect of the room itself on the playback experience. Additionally, we cannot easily control for the position of speakers inside each user's room. Mobile devices present an ideal platform for content delivery, since they eliminate these playback challenges. Users commonly use headphones, which provide a controlled environment for spatial processing that does not exist for speaker arrays. With known mobile speaker geometries, spatial audio could also be delivered via near-field binaural crosstalk cancellation techniques, avoiding, to a large extent, the influence of the room. Thus the mobile platform represents a controlled environment for which it is easier to develop solutions.

1.2 Requirements

Spatial audio is typically recorded from arrays that are not conducive to integration with mobile devices because they require many microphones of a directional nature and a geometry that is not planar. In order for an array to meet mobile manufacturing requirements, it must use as few microphones as possible, fit into existing device geometries, and use only omnidirectional sensors. These requirements are due to cost, existing mobile device geometry, and manufacturing considerations, respectively [1]. We are proposing a system that can meet these requirements for spatial audio recording, as well as playback from a mobile device.

Additionally, any processing done on the array must be possible to perform in real-time, real-world conditions. This means the processing cannot be overly sensitive to differences in sensors or exact microphone placement, two concerns that, given cost, are likely to exist.

1.3 Prior Methods for Recording Spatial Audio

1.3.1 Binaural Recording

Currently there are several methods for recording spatial audio. The first and most direct method, binaural recording, or recording directly at the ear canals, works well and yields high-quality audio [2]. The advantages of binaural recording are that it is simple, inexpensive, and robust. By recording directly at the ears, the cues needed for spatial playback are directly applied to the audio without the need for additional processing. This transparent technique results in spatial audio that sounds very natural and is arguably the pinnacle for recreating an acoustic space over headphones. Drawbacks to binaural recording are that it is inflexible and that it requires a human head.

The inflexibility arises from the fact that binaural audio playback is restricted to headphones and that the spectral characteristics as well as the head movements of the individual wearing the array are "hard-coded" onto the audio. The recording of individualized spatial filters, known as head-related transfer functions (HRTFs), for every consumer is still impractical at this time, which minimizes their benefits, but the technology for employ-

ing head-tracking on a mobile device is quickly becoming a reality. Alone or in conjunction, the front-facing camera or integration of an inertial measurement unit (IMU) into headphones would allow head movements to be tracked and thus enable the spatial scene to be updated accordingly. Updating spatial audio due to head movements has been shown to greatly increase the realism of the presentation. It provides an ability to interact with the environment that is attractive in virtual reality applications.

1.3.2 Ambisonics

The most widespread technique for recording spatial audio in the academic literature is Ambisonics. Ambisonic arrays are popular because they allow symmetrical beampatterns to be steered in any direction in three-dimensional space. Current Ambisonic techniques, which will be discussed in more detail later in the thesis, do not fit our needs because they require directional microphones and a somewhat large (relative to the thickness of a cellphone or tablet) geometry.

Current research in the area of spatial audio is directed towards larger arrays with more microphones, numbering as high as several hundred sensors in a single array [3, 4, 5]. The goal of the current research is to develop greater spatial acuity so that the directional information may be recovered more accurately. Our research, on the other hand, takes the opposite approach. Our goal is to find a method of recording and reproducing spatial audio from the smallest number of microphones and the simplest playback scheme. This will require using just a few omnidirectional microphones, the ability to reconstruct spatial audio from an arbitrary array design, and the ability to play back the reconstructed audio over headphones.

1.3.3 Other Techniques

There are a host of other techniques designed to provide a sense of spatial realism. Most of these techniques were intended for loudspeaker playback and so we will not delve too deeply into them here. The first technique is, of course, stereo. The need for a sense of envelopment resulted in stereo's quick replacement of monophonic sound. The promise of even greater realism with

the introduction of quadraphonic sound was a failure, however. This was due to the increased complexity of the system as well as a non-standardized playback scheme. Quadraphonic sound is a reminder that our systems must be simple and robust to a variety of user equipment in order to reach viability. Home theaters have extended quadrophonics to 5.1 or 7.1, but maintain the problem of controlling for the individual acoustics and layout at each consumer's home.

Finally there are cross-talk cancellation methods which seek to apply binaural sound without headphones [6]. While cross-talk cancellation is possible, it is limited to a frequency range generally below 1.5 kHz and performs better in anechoic spaces, as cross-talk cancellation cannot account for room reflections.

1.4 Our Methods

The basic idea behind spatial audio is to employ headphones or speakers to recreate for the user what it would have been like if they were in the same position as the microphone array. In practice, this means we must recover the directional audio from a microphone array, then play it back to the listener so that it is perceived as coming from the same direction as the original audio. If it is possible to recover the directional audio with a fine enough spatial precision, then it should be possible to recreate the original experience convincingly.

In order to recover the directional audio, it is possible to draw on a large literature on beamforming. Naively, we want to form the beampattern with the best spatial acuity possible, then, for the case of headphone playback, convolve the audio recovered from that direction with the appropriate head-related impulse response (HRIR) in order to place it in space. The problem with this scheme is that for microphone arrays with a limited number of sensors as well as small physical apertures, it is not possible to attain the spatial acuity needed to create the beampatterns without large overlapping regions. This overlap results in summations and cancellations which distort the intended spatial signal presented to the listener.

In order to manage this error in a systematic way, we propose a least-squares method for minimizing the reconstruction error relative to the HRTF.

Using this method, we can get the best performance possible from a given arrangement of sensors. Comparison of various straightforward arrays with Ambisonics will be performed. Finally the least-squares method will be extended to a weighted version, which allows the user to specify regions of angular importance, giving greater control over playback.

This thesis is organized as follows:

- Chapter 2 provides an overview of the psychoacoustic principles relevant to spatial audio. In addition, the development of spatial filters is discussed.
- Chapter 3 describes Ambisonics, the current state of the art in spatial audio recording and playback.
- Chapter 4 covers the basics of beamforming and develops a spatial audio recording and playback scheme based on delay-and-sum or superdirective beamforming.
- Chapter 5 describes the error incurred in the formulation of Chapter 4 and presents a method for minimizing this reconstruction error in the frequency-domain in a least-squares sense.
- Chapter 6 addresses possible circularity issues in the least-squares frequency-domain solution and describes a method for instead formulating the solution in the time-domain with additional memory and computational requirements.
- Chapter 7 directly compares the results of the above algorithms.
- Chapter 8 provides background on gradients and how they can be used to design arrays with advantageous qualities for spatial audio.
- Chapter 9 presents a simple method for spatial audio from two omnidirectional microphones. This provides a good option for situations when more complex arrays are not possible due to cost or geometric concerns.

CHAPTER 2

HUMAN LOCALIZATION AND THE HEAD-RELATED TRANSFER FUNCTION

2.1 Fundamentals of Human Localization

In his seminal work, *Spatial Hearing*, Blauert defines "localization" as the determination of the direction and distance of an auditory event [7] by a listener. These events, which are distinct from sound (mechanical vibrations), are the perceptual aspect of human hearing. The goal of a spatial audio system is to induce auditory events as accurately as possible in order to recreate a real environment, or synthesize an artificial environment, using sound signals generated from loudspeakers or headphones.

The key to an accurate synthesis is the recreation of the cues responsible for human localization. The fundamental cues, *interaural time difference* (ITD), *interaural phase difference* (IPD), and *interaural level difference* (ILD), were first investigated by Lord Rayleigh in pioneering experiments at the end of the 19th century [8, 9]. We call these the fundamental cues because they are highly robust and determine the absolute azimuth from center. He found that listeners had little difficulty identifying left from right for pure frequency tones where little level information exists.

ILD, which describes a difference in intensity between the two ears, was at first thought to be the only cue for localization. Rayleigh made a spherical model of ILD which predicted that at low-frequency there was very little level difference between the two ears, a claim that was backed up by measurements. It was found that for pure low-frequency tones, where little level information existed, the listeners still had no trouble telling left from right. This led to the discovery of ITD, a difference in onset time between the two ears, and IPD, a difference in phase of steady-state sinusoids.

The relationship between frequency and localization cues was discovered by Mills in his two classic papers on the *minimum audible angle* [10, 11].

The minimum audible angle is the smallest difference in angle between two sources that is perceivable before they sound as if they were one source. Mills sought to establish a relationship between *minimum audible phase* - the smallest perceivable phase difference of two steady-state sinusoids of the same frequency presented to the two ears, and *minimum audible intensity* - the smallest perceivable level difference between two steady-state sinusoids, and the minimum audible angle.

Mills found that the minimum audible phase difference agreed with the minimum audible angle below 1400 Hz and that both increased rapidly as frequency approached 1400 Hz. For frequencies above 1400 Hz, it was found that there was correspondence between the minimum audible intensity difference and the minimum audible angle. This crossover in cues at 1400 Hz is likely due to the fact that phase becomes ambiguous as frequency approaches 1400 Hz from below due to wraparound, and that level differences are very small below 1400 Hz due to the wavelength being much larger than the size of the head. The minimum audible angle will have important implications later during our discussion of spatial audio playback. In general, we would like to be able to sample spatially according to the minimum audible angle in order to make a system that is indistinguishable from reality. Unfortunately, due to the constraints on sensors and aperture on mobile devices, meeting this criterion will be impossible. We can, however, ensure a perceptually smooth solution by relating our solution to measurements according to the minimum audible angle, since large changes in the spatial filters will not occur at such an interval. In addition, head tracking updates are more perceptually transparent (free of clicks and jumps) when the resolution meets or exceeds the minimum audible angle.

Rayleigh noticed that while sounds to the left or right were never confused, for certain types of stimuli, notably pure tones, the subjects had difficulty telling front from back, whereas for complex sounds they had no difficulty. While initially ignored as merely a means of protecting the middle ear, it is now believed that it is the role of the *pinna*, or outer ear, to differentiate front/back and elevation ambiguities on the cone of confusion.[1] The pinna accomplishes this feat by spectrally coloring the incoming sound. When pure tones are presented, there is not enough information given by the spectral

[1] A surface of constant ITD, IPD, and ILD. See Figure 2.1.

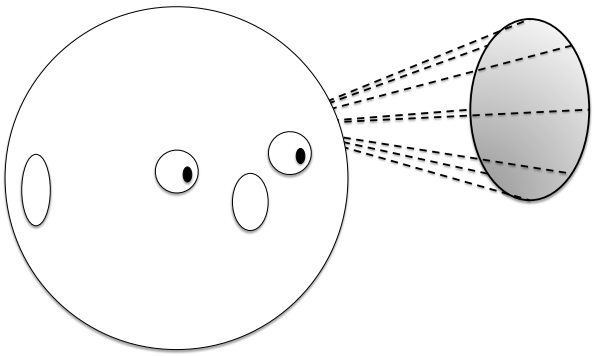


Figure 2.1: Illustration of a cone of confusion for a given angle. The circle (and every point on the external cone) represent a surface of constant ITD and ILD.

coloration at a single frequency in order to determine front and back.

Compared to the left/right cues of ITD, IPD, and ILD, the spectral coloration of the pinna is a relatively weak cue and can easily be confused. An alternative cue for front/back localization is the use of head movements (see Figure 2.2). In a series of ingenious experiments, Wallach [12],[13] used an array of loudspeakers to test which cue was dominant, head movement or spectral cues, due to the pinnae. These experiments were conducted by connecting a series of switches to the user's head, which would in turn select which speaker from the array was playing according to the head movements that were made. For example, a source directly above a listener does not move if they turn their head to the left and right. If Wallach wished to simulate a source directly above the listener, the speaker selected due to the head movements would always be the one directly in front of them. He found that, without fault, the subjects did not find the source to be in front of them, but instead above. Even though the speaker was directly in front of the user and they were obtaining pinnae cues to this effect,

> "In every case of a successful synthetic production the pinna factor is overcome by the cues procured by the head movement for here the perceived direction is quite different from the direction

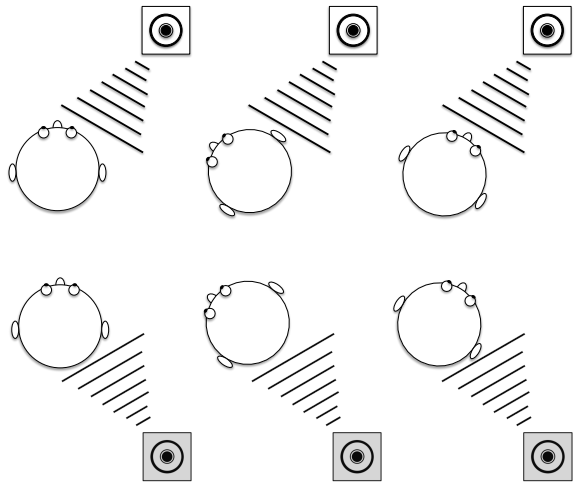


Figure 2.2: Illustration of how head movements resolve front/back confusion. The picture on the left depicts a sound incident on the right side of the subject. With the source in front, a head turn to the left increases the ITD and ILD, while a head turn to the right decreases the ITD and ILD. Conversely, if the source is behind the listener, a head turn to the left decreases the ITD and ILD, while a head turn to the right increases them.

from which the sound actually arrives at the head."

This points to the importance of developing a system that is capable of dynamically updating virtual sources due to head movements.

In addition to the role of head movements, Wallach also investigated what is known as the *law of the first wavefront*, or *precedence effect* [14]. In this paper, Wallach sought to explain why humans localize a sound only according to the source direction instead of being confused by the directions of later reflections. Humans hear only one source in a reverberant space, rather than many individual echoes.

Wallach found that when the same sound originated from two separate loudspeakers, the listener localized the sound as coming from the loudspeaker from which the sound was played first, even if the intensity of that sound was less than the delayed speaker. Once the delay between speakers became 70 ms, listeners perceived an echo instead of a single source coming from the first speaker. We will make use of this effect when describing a weighting scheme for manipulating the importance of angles of incidence in the chapter on least-squares filter-design.

2.2 The Head-Related Transfer Function and Virtualization

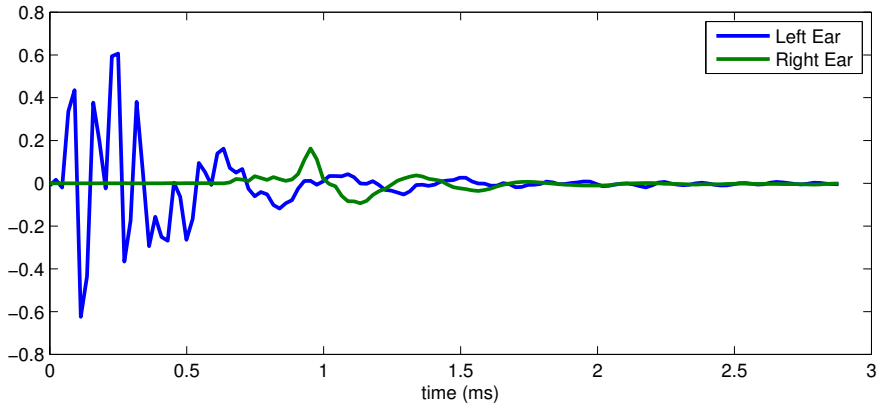
Early attempts at simulating spatial audio were performed by emulating the ILD and ITD between the ears with panning and delays. While this technique could effectively move sounds from left to right, and even to the sides of the listener, it cannot place sound in front or behind. Listeners hear the sounds as if they came from inside their heads. This technique is known as lateralization.

Many classic experiments on human localization made use of headphones because they provided a better method of controlling the environment and stimulus presented to the subject. These methods, however, were lateralization studies rather than localization. Wightman and Kistler, two psychologists that studied localization, liked that the headphone environment provided excellent control, yet worried that the results of these lateralization experiments were not representative of human hearing in general [15, 16]. In order to find filters that would more closely resemble human localization, they measured the transfer function of a subject's head. By using these measurements as a linear filter, they were able to mimic a source at a specified location when audio that was convolved with the measurements at each ear was played back over headphones.

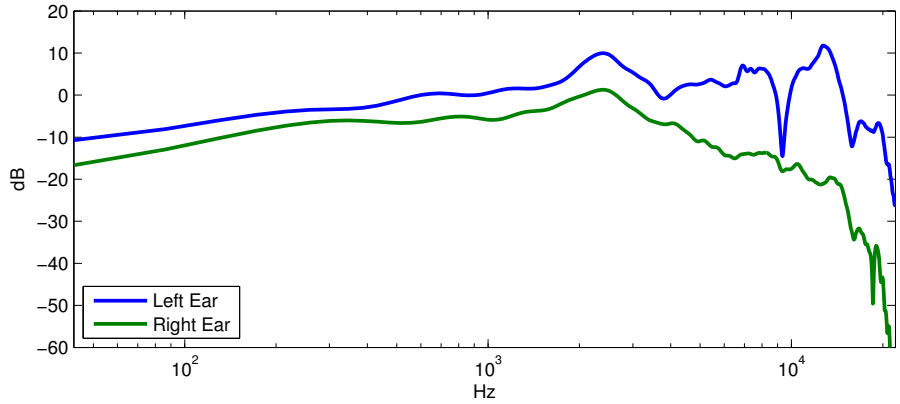
These measurements yield what is known as the head-related impulse response (HRIR) in the time-domain, and its corresponding frequency-domain representation is called the head-related transfer function (HRTF). An example of an HRIR and HRTF can be found in Figure 2.3. The HRTF encodes the IPD, ITD, and ILD of a subject, as well as the spectral cues from the pinnae in a pair of filters (one for each ear) for each source location. Wightman and Kistler went on to publish several papers on the HRTF and its capability for simulating a real environment. They found evidence that greater localization accuracy is obtained through the use of one's own HRTF, and they are largely responsible for the idea that HRTF personalization is necessary.

Use of the HRTF for spatialization is common not only in psychoacoustic work, but also in the creation of virtual environments that employ headphones [2]. In an entertainment setting, localization of the highest accuracy is not necessarily as important as creating a sense of immersion. One of the major drawbacks of strictly using panning and delays is that it results in an

inside-the-head sensation, where sources are localized inside the head rather than occurring naturally in the external environment. While HRTFs have alleviated this problem somewhat, there is not a consensus that this problem has been solved, or is even well defined.



(a) HRIR: note the timing and level cues between the two ears.



(b) HRTF: note the similar level at low-frequency, the greater level difference at high frequency and the spectral notch near 9 kHz at the left ear.

Figure 2.3: Example of an HRIR and HRTF. The angle of incidence is directly at the left ear. All example plots are taken from the MIT set for easy comparison with previous work [17].

2.3 Externalization

One of the most careful studies of externalization was performed by Hartmann et al. [18]. In this paper the authors tested the conditions for externalization by using headphone playback. A synthesized vowel, consisting

of a fundamental and 36 overtones, was presented over a loudspeaker and recorded at the two ears. They then matched the phases and amplitudes of the individual overtones for headphone playback. When this signal, called the baseline synthesis, was properly created, listeners could not distinguish between the real and virtual source. The presentation angle was made at an azimuth of 37 degrees. When properly matched to the speaker playback, the synthesized source could not be distinguished from the real-world source. Since the real-world source was perceived as externalized, it follows that the synthetic source was also externalized. This gives a strong argument that externalization is possible over headphones without large amounts of reverberation. The authors did note, however, that the experiments were not performed in the median plane because they produced mixed results. In my personal experience, sounds to the side externalize much better than sounds in the center. Reverberation does not necessarily solve this problem.

It will be helpful to examine some of the results of Hartmann's paper in order to get a sense of how the cues discussed in the section above relate to perception of audio rendered over headphones. In the paper, a variety of cues and classical acoustical ideas are investigated and careful experimentation allowed the manipulation of cues in a precise manner.

Note that the author's method did not rely on HRTFs, but rather a direct manipulation of the amplitudes and phases of the harmonics at each ear. Perfecting externalization for more complex sound may not be as straightforward.

2.3.1 Experiments

Constant interaural phase difference

In the first experiment, phases above a boundary frequency were altered to be a constant value ϕ_0. It was discovered that listeners could begin to distinguish the altered version from the external speaker at 1 kHz. This gives support to the idea that humans are insensitive to phase at high frequency. When the phase was altered below 1 kHz, subjects reported the sound as coming from within the head. For robust externalization, the region between 400 Hz and 600 Hz was deemed to be most critical for establishing an externalized

image.

Constant interaural time difference

In this experiment, the time difference between the two ears was set to be constant (i.e. linear IPD). According to diffraction around a sphere, ITD can be estimated as

$$ITD = \frac{3a}{c} sin(\theta) \tag{2.1}$$

where a is the radius of the sphere, c the speed of sound, and θ the azimuth relative to the front of the subject. It was found that for the 37^o azimuth tested, a constant time delay was equally as effective as the baseline synthesis. The subjects could not distinguish the real and virtual sources when an optimal constant time delay was used. Additionally, subjects were not overly sensitive to ITDs that were too large, but found small ITDs to be perceived as inside the head. ITDs that were too large moved laterally towards the side and were perceived as more distant.

Level experiments

The level experiments were similar to phase experiments except that ILD was set to zero *below* a target frequency. The authors observed that externalization was not a function of frequency, but rather the number of harmonics zeroed out. Thus if the sound was synthesized at a higher fundamental frequency, the transition took place at a higher frequency as well.

Inside-out experiment

Starting with the highest harmonic, the IPD of one harmonic at a time was set to zero. By doing so, the authors were able to show that a source could be continuously moved from outside to inside the head. Further experimentation showed that the results were difficult to obtain with the source directly in front of or behind the listener. These findings match my own personal

experience. Sounds that are presented directly to the left or right of the listener can be externalized by presenting a level cue with no interaural delay. As the sound approaches the front or back, however, distance becomes more ambiguous and less clearly externalized.

A limitation of this study is that it did not investigate the use of reverberation, which would have complicated the experimental setup. Also, as Hartmann notes, externalization is not necessarily a clearly defined concept, and so he chose to examine externalization in a carefully controlled context.

2.4 A Unified Study of Virtual Cues

Begault et al. [19] sought to test the three conditions of individualized HRTFs, head tracking and reverberation on the factors of localization accuracy, front/back errors and externalization. Their experiments focused on speech rather than wide-band noise, which was used in most other well known experiments. The use of speech is important because it represents a stimulus that is much more familiar to the end user than noise bursts and one that has greater implications for the user experience.

It was found that, contrary to other studies, individualized HRTFs did not outperform generic HRTFs in terms of localization accuracy. Additionally, individualized HRTFs had no impact on front/back reversals or externalization. It appears that for speech, the use of individualized HRTFs is not necessary.

The use of head tracking to reduce azimuth error varied with the individual subject. While it was found that head tracking did not reduce azimuth error in general, the individual that exhibited the greatest amount of head movement found the greatest benefit from the inclusion of head tracking. This individual was also the best localizer of the group. Head tracking was found to consistently reduce front/back error. The subjects in the study were only given a 3 second stimulus. In my experience, a trained subject with unlimited time will essentially achieve zero front/back reversals when head tracking is used.

Reverberation was found to be strongly linked to externalization. Perhaps counter-intuitively, reverberation was also shown to provide a small improvement to azimuth error. One hypothesis for this occurrence is that sources

that are perceived further in the distance have a lower sensitivity to small error. When a sound is very close, however, a small deviation in position results in a large error in azimuth. Pulling the sound outside the head may result in decreasing the sensitivity for virtual sources as well.

2.5 Chapter Summary

In the present chapter, cues for localization were discussed. Of particular relevance to our task is the minimum audible angle, which will influence the angular resolution with which we need to sample the HRTF for our virtualization scheme. Additionally, we can exploit the precedence effect in order to maintain accurate localization, while sacrificing some of the accuracy of reverberant directions.

CHAPTER 3

AMBISONICS

There are numerous methods for recording and playing back spatial audio, the three most common being binaural recording [2], Wave Field Synthesis (WFS) [20], and Ambisonics [21, 22]. As mentioned in the Introduction, binaural recording is limited from the perspective that it requires placing an array on a human head and that it "hardcodes" the HRTFs of the individual wearing the array onto the recorded audio. WFS, on the other hand, seeks to recreate a soundfield using large arrays of speakers. Of the three, Ambisonics is the most flexible because it enables both headphone and loudspeaker playback, and the symmetric response of the Ambisonic array allows the reorientation of the user's perspective in real time, enabling the use of head tracking.

It is important to examine Ambisonics in detail, as it represents the state of the art for spatial audio recording and playback. As such, it provides a benchmark for our mobile designs and informs our design decisions as well. While our work is not necessarily Ambisonic in nature, it will be shown that our playback solutions are equivalent when an Ambisonic array is used. Our goal is to find the best reconstruction of an audio scene possible from an arbitrary array. Our reconstruction technique, therefore, is a generalization of Ambisonics that allows for arbitrary array design and performs correction of non-ideal arrays, including Ambisonic ones. Ambisonics features many desirable characteristics that are useful to emulate in both array design and reconstruction capabilities. A closer look into Ambisonics will help motivate our solution to the mobile spatial audio problem.

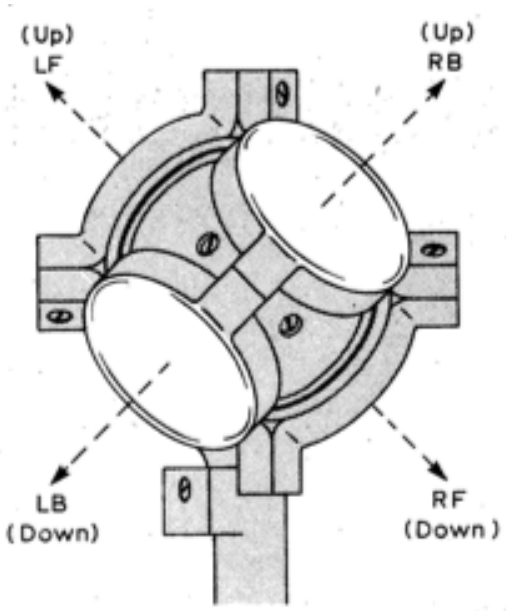


Figure 3.1: The original soundfield microphone design by Ken Farrar at Calrec [23].

3.1 The Soundfield Microphone

The heart of Ambisonic technology is the Soundfield microphone, which is designed with the goal of getting an identical polar response in all look directions [23]. Ambisonic signals are recorded in what is known as A-format, which is comprised of the signals from four cardioid microphones arranged in a tetrahedron (Figure 3.1). The idea behind this arrangement of cardioids is twofold:

1. A linear combination of the four microphones allows a response from the cardioid family[1] to be steered in any direction in three dimensions.

2. The tetrahedral arrangement allows the microphones to be placed as closely to coincident as possible.

One can think of the soundfield microphone as supplying an infinite amount of directional channels from just four microphones.[2] If one takes the outputs from a large number of steered directions in parallel, the audio from each channel can be convolved with a spatial filter corresponding to that direction,

[1] i.e. cardioid, hypercardioid, supercardioid, etc.

[2] These channels will, however, have a limited amount of spatial acuity.

thus making approximation of a 3D audio scene possible given a reproduction scheme capable of playing back all of the channels simultaneously.

The coincidence of the microphones is important because it allows a true cardioid to be formed for a wider frequency range. As the microphones become separated, the ability of the array to form perfect cardioids becomes compromised, mostly at high frequency. The arrangement of cardioids in the soundfield microphone built by Farrar allowed for effective co-incidence up to 10 kHz [23]. With the current ability to build significantly smaller microphone capsules, coincidence can be achieved to an even greater degree.

While A-format can be preferable for recording, Ambisonic signals are generally converted to B-format before being reconstructed. The B-format channels are commonly referred to as W, X, Y, and Z, consisting of an omnidirectional and three orthogonal figure-8 responses respectively. The equations to obtain these channels according to the labeling in Figure 3.1 are:

$$W = LFU + RFD + LBD + RBU \text{ (Omnidirectional Channel)} \tag{3.1}$$

$$X = LFU + RFD - LBD - RBU \text{ (Fig-8 in x, horizontal)} \tag{3.2}$$

$$Y = LFU - RFD + LBD - RBU \text{ (Fig-8 in y, horizontal)} \tag{3.3}$$

$$Z = LFU - RFD - LBD + RBU \text{ (Fig-8 in z, vertical)} \tag{3.4}$$

B-format is preferred for reconstruction because forming a response from the cardioid family in any direction features more straightforward mixing equations:

$$E(\theta, \phi) = \frac{\sqrt{2}}{2}W + cos(\theta)cos(\phi)X + sin(\theta)\cos(\phi)Y + sin(\phi)Z \tag{3.5}$$

This is possible because a linear combination of the three figure-8 channels can form a new figure-8 in any direction. For example, mixing the X and Y channel at equal gain with a gain of zero for the W and Z channels will result in a figure-8 at 45 degrees in the XY plane. Since a figure-8, or gradient response, has a positive and a negative lobe, the sum of a figure-8 and the omnidirectional channel form a cardioid.[3] Additionally, any other response

[3]Actually the weighting of the omnidirectional microphone is $\sqrt{2}/2$ for Ambisonics, but since this does not correspond to a commonly used weighting (e.g. supercardioid), we will

from the cardioid family can be formed by changing the weighting of the omnidirectional channel. For example, a weight of zero on W will result in no rear rejection (a figure-8), while a weight equal to the figure-8 constructed from the X, Y, and Z channels will result in a cardioid.

3.2 Ambisonic Soundfield Reconstruction

The original Ambisonic reconstruction equations were derived by solving for a least-squares solution of a soundfield using spherical harmonics. When a symmetric speaker array is used for playback, this can more intuitively be understood as assigning to each speaker the audio recovered by forming a cardioid in its direction. For example, a speaker directly to the right of the listener will play back audio recovered by steering the array to form a cardioid directly to the right. Each speaker plays back the audio recovered from its respective cardioid simultaneously, thus recreating the auditory scene [21, 22].

If an ideal array is used, meaning the four microphones are perfectly coincident and their polar patterns are perfect cardioids for all frequencies of interest, then the B-format channels X, Y, and Z are perfect figure-8's, and W will be a perfect omnidirectional response, all of which will be perfectly coincident.[4] In this case, the mixing weights w to form a cardioid from the B-format signals are simply:

$$w_W(\theta, \phi) = \frac{\sqrt{2}}{2} \tag{3.6}$$

$$w_X(\theta, \phi) = cos(\theta)cos(\phi) \tag{3.7}$$

$$w_Y(\theta, \phi) = sin(\theta)cos(\phi) \tag{3.8}$$

$$w_Z(\theta, \phi) = sin(\phi) \tag{3.9}$$

where θ is the azimuthal angle, and ϕ is the elevation.

use the term cardioid for brevity.

[4]None of these assumptions will be true in practice. In particular, the polar pattern of the channels will not be constant with frequency.

3.3 The Virtual Speaker Technique

While Ambisonics was originally intended as a system for playback over speakers, it is possible to reproduce Ambisonic audio over headphones using what is known as "virtual speakers." In this method, audio recovered from a given direction is placed in space virtually by convolving it with the HRIR pair corresponding to that direction, rather than playing it back through a loudspeaker [24]. For a completely synthetic audio scene, it may be beneficial to convolve the audio instead with an HRIR recorded in a reverberant space.[5] For audio that is recorded in real rooms, however, the direct sound, as well as the reflections, will each be recovered according to their respective angles of incidence and therefore will be convolved through different HRTFs. Therefore, the original spatial characteristics of the room are preserved without the need to add additional reverberation.

The principle advantage of Ambisonic headphone playback is the possibility of using a larger number of virtual speakers than would be practical with real speakers. Since the number of virtual speakers is unlimited, we can place sources according to the minimum audible angle described in the preceding chapter. This prevents audio "sticking to" and "jumping between" speakers when a large angle separates them [24]. Another advantage of headphones over speakers is that the acoustic space is entirely controlled. With speakers in a real room, there will be reflections that are not intended by the reconstruction scheme. The result is additional reverberations will be added upon playback, distorting the intended audio experience.

3.3.1 An Efficient Implementation of the Virtual Speaker Technique

One concern when using a large amount of physical speakers is that the number of filters needed to implement the system scales linearly with the number of speakers. Due to linearity, an efficient implementation is possible using virtual speakers for reconstructing the entire 3D scene using only $M \times 2$ filters, where M is the number of microphones, and 2 is the number of headphone speakers [25]. This greatly reduces the computational load of the algorithm, making it run much faster in real-time.

[5] This is known as a binaural room impulse response (BRIR).

Generalizing the above equations for an Ambisonic array of arbitrary order, if we denote $y[n, \theta]$ as the audio recovered from a set of microphone signals $x_m[n]$, when the array is steered in look direction θ (ignoring elevation for notational simplicity), then:

$$y[n, \theta] = \sum_{m=1}^{M} w_m(\theta)\ x_m[n] \tag{3.10}$$

where $w_m(\theta)$ are the weights for microphone m that form the desired beampattern in direction θ. For a single headphone speaker we can then write:

$$output[n] = \sum_{l=1}^{L} (hrir[n, \theta_l] \star y[n, \theta_l]) \tag{3.11}$$

$$output[n] = \sum_{m=1}^{M} \left[\left(\sum_{l=1}^{L} w_m(\theta_l) hrir[n, \theta_l] \right) \star x_m[n] \right] \tag{3.12}$$

Therefore, the filtering operation collapses down to a single filter for each microphone per headphone speaker:

$$h_m[n] = \sum_{l=1}^{L} w_m(\theta_l)\ hrir[n, \theta_l] \tag{3.13}$$

We will revisit this efficient implementation in the next section on least-squares reconstruction.

3.3.2 Error in the Ambisonic Reconstruction

Due to the fact that the Soundfield microphone forms responses from the cardioid family, neighboring beampatterns for each look direction will have substantial overlap.[6] A simplified example can be found in Figure 3.2. The result of this overlap is imperfect reconstruction. Sound incident from a given direction is processed by the HRTFs from all look directions according to the gain of their respective cardioids, meaning that the actual filter that gets applied to sound incident from a given direction is a linear combination of

[6]The goal of the current research, Higher Order Ambisonic systems, is to minimize this overlap.

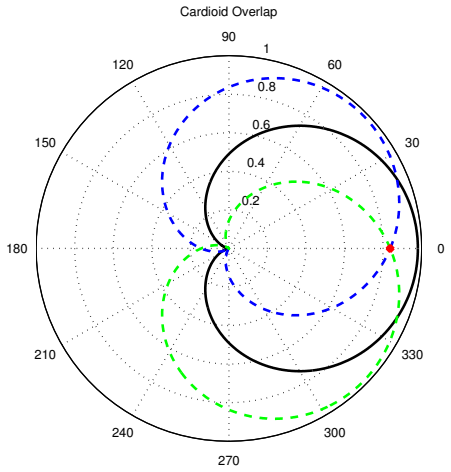


Figure 3.2: Depiction of overlapping cardioids for the simplified case of recovering from 0° with interfering HRTFs at ±45°. The red dot is the gain at which the HRTFs from the non-look directions will be mixed with the intended filter.

every HRTF. If $g(\theta_l, \theta)$ is the gain of the cardioid in direction θ when steered in direction θ_l[7], then the HRIR fit is

$$hrir_{fit}[n, \theta] = \sum_{l=1}^{L} g(\theta_l, \theta) hrir[n, \theta_l] \tag{3.14}$$

The resulting HRTF fit is a distortion of the intended HRTF, which is more evident in the contralateral ear than the ipsilateral ear (see Figure 3.3), because the large response at the ipsilateral ear is less sensitive to the additive error. The result is a modification of the localization cues, which brings about localization errors as well as other undesirable effects, such as the sensation that the sound is originating from within the head.

3.4 Chapter Summary

In this chapter we have examined the basics of Ambisonic recording and reconstruction. We have shown that the reconstruction error is a result of the limitations of the spatial acuity of the Ambisonic array. Since Ambisonics

[7] $g(\theta_l, \theta)$ is scalar for an ideal array.

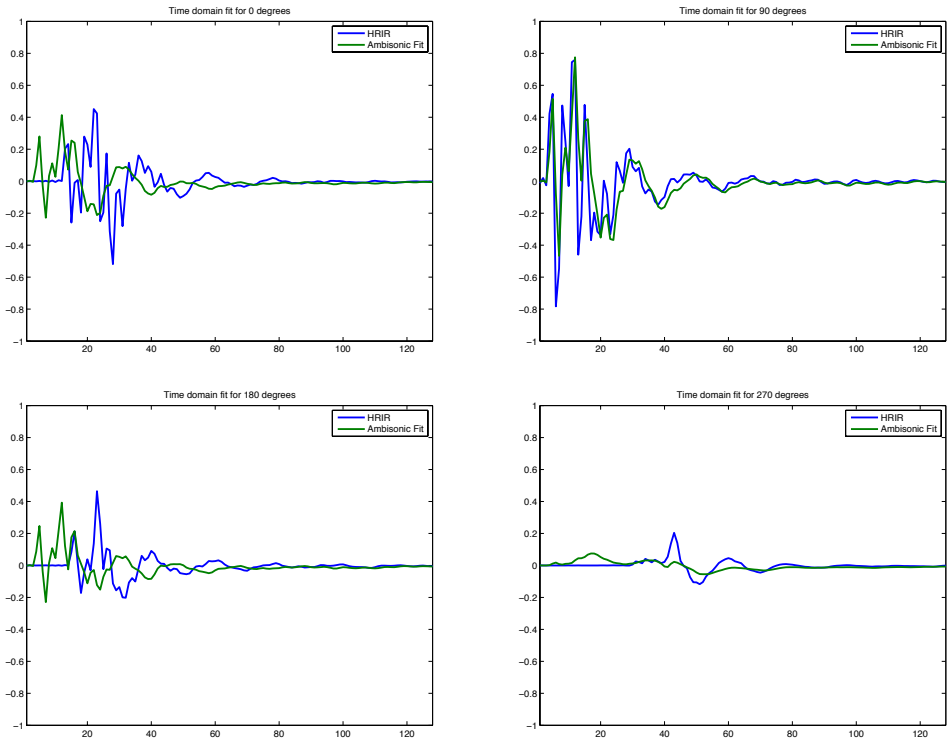


Figure 3.3: These plots show the Ambisonic fit to the HRIR for the left ear at four angles of incidence. The plots on the left represent a source directly in front of and behind the listener. At the top right, 90 degrees represents a source directly into the left ear. At the bottom right, 270 degrees represents the ipsilateral ear, or a source directly into opposite the ear. It is evident that the fit is best for the contralateral ear. As the source moves away from the ear, timing errors become problematic. The above fits were made for the case of horizontal "virtual speakers" spaced every 5 degrees.

simply involves a linear combination of microphones, it can also be thought of as a delay-and-sum beamformer. It is possible that alternative arrays or beamforming techniques may improve upon the Ambisonic results. In the next chapter we will examine using standard beamforming techniques for acquiring spatial audio from an arbitrary array.

CHAPTER 4

A NAIVE METHOD FOR RECONSTRUCTING A SPATIAL AUDIO SCENE

In the previous chapter we discussed reconstruction from an Ambisonic array. Unfortunately in mobile applications, specialized arrays of this type are not available due to cost and geometric considerations. An alternative option, which is more affordable, is an array of omnidirectional microphones, typically in some type of planar geometry.

In this chapter we will take inspiration from the Ambisonics virtual speaker technique. The general idea of the technique is to steer beams in the desired virtual speaker directions, then play back the audio recovered from those directions spatialized by the HRTF. This turned out to be an effective strategy when using a symmetric speaker array and a soundfield microphone. In this chapter, we will explore the efficacy of this idea when other arrays and beamforming techniques, not specifically designed for spatial audio, are used. It is important to note that the beamforming method should form fixed beams, as we are trying to capture audio from a static direction as opposed to adaptively cancel noise.

Multiple beamforming schemes were explored, including delay-and-sum and superdirective beamforming. As the results show, the increased directivity of the superdirectional algorithm greatly enhances spatial discrimination. In addition, the superdirective beampattern features a more consistent shape across frequency bands.

4.1 Spatial Sampling

Before getting into specific beamforming techniques, we will first cover some basic array processing background. The two factors that fundamentally limit the ability of an array to perform spatial discrimination are aperture size and sensor spacing. These two factors have important analogies in digital signal

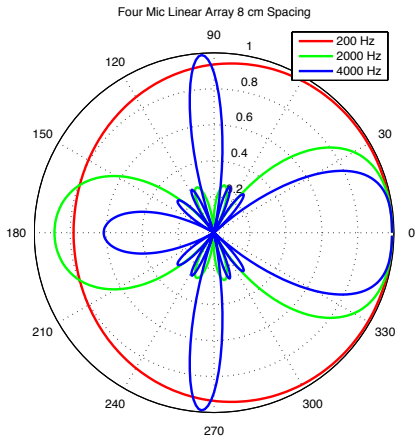


Figure 4.1: An example of spatial aliasing. At 200 Hz, there is no aliasing, but little directionality. At 2 kHz, sampling is slightly denser than Nyquist and directivity is high. At 4 kHz, severe aliasing occurs causing the directionality of the beampattern to be poorly defined. In addition to the added side lobes, the lobes of high energy are called *grating lobes.*

processing (DSP) sampling theory to sampling rate (sensor spacing) and sequence length (aperture size). The density of sensors in a given direction is essentially the spatial sampling rate, while the aperture size determines how long the signal is observed over space.

Sampling theory tells us that in order to sample a signal properly, we must have more than two samples per period at a desired maximum frequency f_{max}. If our sampling rate is too low, aliasing will occur. In spatial sampling, we also must have two samples per wavelength, or spatial aliasing will occur.[1] The wavelength of a given frequency can be calculated by

$$\lambda = c/f \tag{4.1}$$

where λ is the wavelength in m, c is the speed of sound in m/s, and f is the frequency of interest in Hz. Therefore, according to a desired maximum frequency, the first design criterion is to place our sensors at intervals of $\lambda_{max}/2$.

While having sensors spaced at $\lambda_{max}/2$ will prevent aliasing, as is evident from knowledge of sampling, denser sampling at a frequency that is already satisfying the Nyquist rate will not greatly improve the array's performance,

[1] See Figure 4.1.

just as more densely sampling in time will not enhance frequency resolution. If we want to increase our abilities for spatial discrimination, we must increase the size of the aperture. Since the spatial aperture of the array must be increased while continuing to satisfy the Nyquist criterion, the simplest solution is to increase the number of microphones with each spaced at a maximum of $\lambda_{max}/2$ from the nearest microphone.

Another possibility is to use different microphone spacings for different frequency ranges. A large spacing could be used for low frequencies and a small spacing for high frequencies. This technique allows the array to have more spatial acuity at low frequencies, while reducing the number of microphones. Therefore, a more uniform performance can be obtained across frequency by not completely filling in the "grid" of microphones that would form a uniform spacing.

The basic idea of array aperture design is that the longer interval of space over which we observe the signal, the better estimate we will have of its spatial characteristics. This is akin to how increasing the number of samples observed (at a fixed sampling rate) of a waveform gives more resolution in the frequency-domain for a signal that is relatively stationary in time (space). In time-domain sampling, we must observe low frequencies for a longer time interval than we observe high frequencies in order to get a similar estimate of their content. The same is also true in spatial sampling. The aperture must increase by a large margin in order to provide a nominal amount of discrimination, as the wavelengths become much larger at low-frequency.

Consult Table 4.1 for some example frequencies and their wavelengths. A typical mobile device has a short-side dimension of 5 cm to 18 cm and a long-side dimension of 11 cm to 25 cm. From this table it is clear that it will be difficult to achieve good spatial discrimination at low-frequency. Even at 1 kHz, a wavelength of 34.3 cm means we will only be able to place *at most* two sensors at the maximum resolution spacing of $\lambda/2$. More sensors could be placed in-between these two sensors, but as discussed above, by the sampling theorem, they will give little added benefit.

Table 4.1: Frequencies and their wavelengths

Frequency (Hz)	Wavelength (m)
20	17.15
100	3.43
200	1.72
500	0.686
1000	0.343
1500	0.229
3000	0.114
5000	0.069
8000	0.043
15000	0.023

4.2 Delay-and-Sum Beamforming for Spatial Audio

Real-world delay-and-sum (or filter-and-sum) beamformers can be implemented on a digital computer where audio is acquired at a sampling rate, f_s, sufficient to account for non-integer sample delays. Since a non-integer delay is sinc-like in nature, the filters which are designed in order to perform the beamforming operation must be truncated to some length N. For a source from the far field, for each frequency bin $k = fN/f_s$, look direction θ, and microphone m, we define a frequency-domain steering vector $\mathbf{d}(k, \theta)$, with elements $d_m[k, \theta]$ that have magnitudes A, and phases ϕ, at each of M microphones as

$$\mathbf{d}(k,\theta) = \begin{bmatrix} A_1(k,\theta)e^{j\phi_1(k,\theta)} & A_2(k,\theta)e^{j\phi_2(k,\theta)} & \dots & A_M(k,\theta)e^{j\phi_M(k,\theta)} \end{bmatrix}^T \quad (4.2)$$

where A is angle-dependent because the microphones generally do not have omnidirectional responses. The corresponding filter will simply use the conjugate of the steering vector

$$\mathbf{h}(k,\theta) = \mathbf{d}(k,\theta)^* \quad (4.3)$$

Since the delays are non-integer in general, we cannot simply perform a standard inverse DFT if we want to form a real-valued filter in the time-domain. Instead, we need to choose our frequency bins such that they correspond to

the interval $-\pi \leq \omega \leq \pi$ of the DTFT, or

$$\begin{cases} -\frac{N}{2} \leq k < \frac{N}{2}, & N \text{ is even} \\ -\frac{N-1}{2} \leq k \leq \frac{N-1}{2}, & N \text{ is odd} \end{cases} \tag{4.4}$$

For an even-length filter the inverse DFT then becomes

$$h_m[n,\theta] = \frac{1}{N} \sum_{k=-N/2}^{N/2-1} H_m[k,\theta] e^{j\frac{2\pi k}{N}n} \tag{4.5}$$

with a similar form for odd-length sequences.

If it is desired to use common FFT algorithms, the sequence simply needs to be inverse FFT shifted before the inverse transform is performed.

4.2.1 Delay-and-Sum Beampatterns Formed by Mobile-Sized Planar Arrays

This section examines the implications of limited aperture size and sensor numbers on spatial discrimination by looking at a few common array designs. The case studies, or example arrays, used throughout the rest of this thesis will be an Ambisonic array, a four-element box, a 25-element square grid, and an eight-element ring. Example beampatterns are given in Figure 4.2 through Figure 4.5. These arrays have been chosen since they are standard array types that may fit on a mobile device. While the 25-element grid (and even the 8-element ring) may use too many microphones to be practical, they help demonstrate how much improvement can be made as the order of the array increases.

Note that the beampatterns are not identical in all directions for the arrays of omnis. Their performance, however, is close enough that examining one representative direction is good enough to get an idea of each array's spatial discrimination abilities. For delay-and-sum beamforming, it will become evident that array shape changes dramatically over frequency for the omnidirectional arrays.[2] This is a major weakness of delay-and-sum beamforming when applied to audio. In addition, the low-frequency performance of each array is poor due to the limited aperture size. Therefore, increasing

[2]See Figure 4.3.

the array order will primarily be beneficial at high frequency where adding elements increases the array aperture to multiple wavelengths.

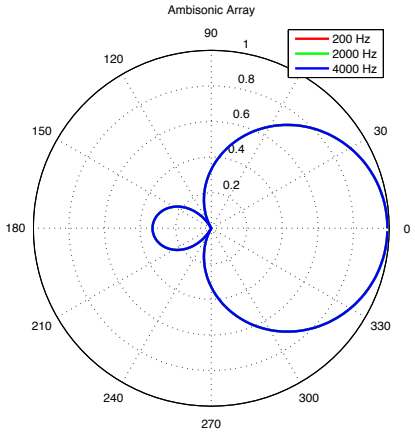


Figure 4.2: Beam patterns for an ideal Ambisonic array in the plane. Note that this array is only capable of forming beam-shapes from the cardioid family. Also note that beam-shape is constant for all look directions (not shown) and across frequency.

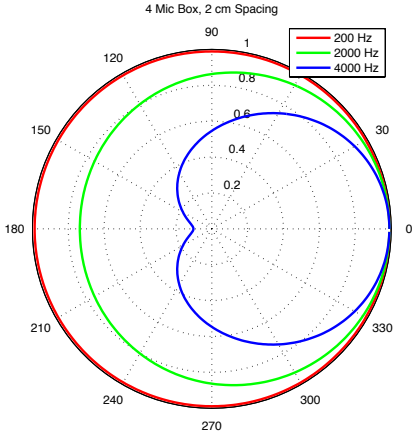


Figure 4.3: Beam patterns for four elements in a square with the sides of the square having a length of 2 cm. Note that at low-frequency the array is basically omnidirectional.

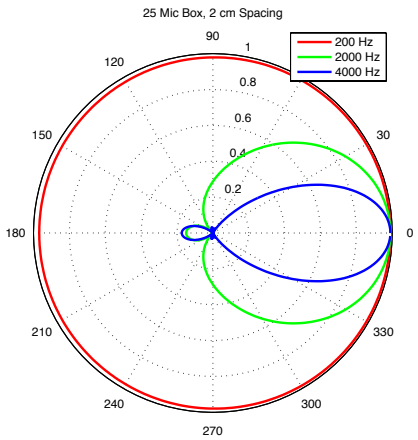


Figure 4.4: Beampatterns for a 25-element grid. Note that beam shape at 200 Hz is again primarily omnidirectional since the aperture of 8 cm on a side is still small compared to the 1.72 m wavelength at 200 Hz.

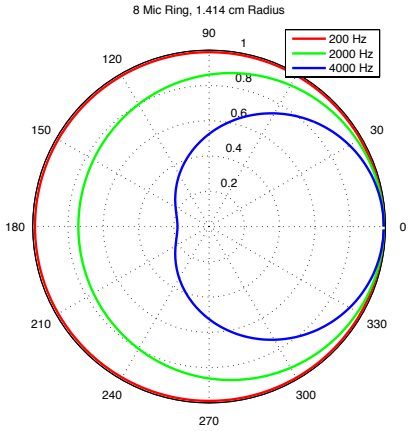


Figure 4.5: Beampatterns for an 8-mic ring with 1.414 cm radius. Note that this does not achieve significantly better performance than the four-element box, since the array aperture is identical.

4.3 Superdirective Beamforming for Spatial Audio

As we observed in the previous section, standard additive beamforming does not give us enough spatial resolution for spatial audio capture from a mobile-sized array. An alternative is superdirective beamforming, which is defined as beamforming that obtains greater directivity than an equally weighted summation of delayed channels [26]. One method of superdirective beamforming is by forming gradients.[3] In this chapter we will examine obtaining superdirectional beamformers by performing standard minimum-variance distortion-

[3]Covered in detail in the Appendix A.

less response (MVDR) beamforming in the presence of isotropic noise.

4.3.1 Overview of Superdirective Beamforming

MVDR is a popular method for designing adaptive beamformers, which minimizes the influence of noise from a dynamically changing environment. For spatial audio capture, however, we want to design *fixed* beamformers for capturing a scene accurately. We can obtain fixed beamformers from MVDR by specifying a stationary noise field.

The main idea behind MVDR beamforming is to solve for the lowest energy output (minimum variance) under the constraint of passing the desired look direction unchanged (distortionless response). This is accomplished by reducing the gain, or if possible by placing a null at angles where interfering sources appear, while setting an arbitrary gain at angles where there is no interference. In the presence of isotropic noise, however, it is not desirable to steer nulls or arbitrary lobes. Instead MVDR must reduce the influence of all angles with equal weight, without altering audio coming from the desired look direction. In an isotropic noise field, the superdirective algorithm finds the highest directivity beampattern possible in order to satisfy the minimum variance requirement [26].

The MVDR Algorithm

The MVDR algorithm is a type of statistically optimum beamformer [27]. Mathematically it can be expressed as

$$\min_{\mathbf{h}}\{\mathbf{h}(k,\theta)^H \mathbf{\Phi_{XX}}(k,\theta)\mathbf{h}(k,\theta)\} \text{ subject to } \mathbf{h}(k,\theta)^H \mathbf{d}(k,\theta) = 1 \tag{4.6}$$

where $\mathbf{\Phi_{XX}}(k,\theta)$ is a cross power spectral density matrix for discrete frequency k and look direction θ. The statistically optimal solution can be found by the method of Lagrange multipliers as

$$\mathbf{h}(k,\theta) = \frac{\mathbf{\Phi_{XX}}(k,\theta)^{-1}\mathbf{d}(k,\theta)}{\mathbf{d}(k,\theta)^H \mathbf{\Phi_{XX}}(k,\theta)^{-1}\mathbf{d}(k,\theta)} \tag{4.7}$$

Practical Issues

In three dimensions, spherically isotropic noise would be used to calculate the spectral correlations. Since we have been restricting ourselves to planar geometries, however, we will employ cylindrically isotropic noise. In order to simulate cylindrically isotropic noise for arbitrary array designs, we have used a ring of discrete sources. Adequately modeling isotropic noise requires the spacing to be such that the beamformer cannot steer a null at the individual sources or place a sidelobe with large gain between sources.

In order to perform MVDR, we must analytically find or estimate the power spectral density matrix Φ_{XX}. Since we are designing fixed beamformers for arbitrary arrays, it is more convenient to simulate the noise to approximate the covariance matrix numerically. The results of performing the simulation directly depend on the number of test sequences used. It is important to average the result over many possible noise-sequence realizations in order to design the best beamformer.

A standard method for forming the power spectral density matrix is taking the Fourier transform of the signal at each microphone

$$X_m(k) = DFT\{x_m[n]\} \tag{4.8}$$

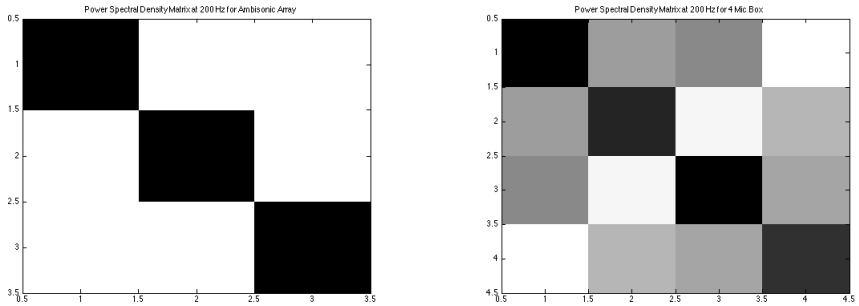
then estimating the ij^{th} element of $\mathbf{\Phi_{XX}}(k, \theta)$ as

$$\mathrm{E}[X_i(k)X_j(k)^H] \tag{4.9}$$

If, however, we wish to average over a large number of sample sequences, we can make a matrix in which the rows correspond to the different microphones and the columns correspond to the individual realizations. Thus taking the outer product still results in an $M \times M$ matrix, but averaged over the power spectral densities of the individual realizations. Since we are performing fixed beamforming and the noise is stationary, it is possible to use a large number of realizations in order to get a highly accurate simulation of an isotropic noise field.

4.3.2 Brief Analysis of Superdirective Cross Power Spectral Density Matricies

The cross power spectral density matrices in isotropic noise give some sense of the characteristics of each array.[4] Since this matrix is inverted in the MVDR algorithm, properties of this matrix also give some information about the stability of the superdirective solution for a given array. Examining two cases, we notice that for an Ambisonic array, the matrix is diagonal. The four-element box has significant off-diagonal energy. For extremely low frequencies, this matrix becomes close to singular. Therefore, despite the fact the Ambisonic array cannot achieve the performance of the four omnis (details below), the orthogonality of the channels imparts well-conditioned solutions.



(a) The cross power spectral density matrix of an Ambisonic array at 200 Hz. Note that for an ideal Ambisonic array, this matrix is constant across frequency and always diagonal.

(b) The cross power spectral density matrix of a four-mic box at 200 Hz. Note the off-diagonal terms.

Figure 4.6: Comparison of cross spectral density matrices for an Ambisonic array and a four-mic box.

[4] See Figure 4.6.

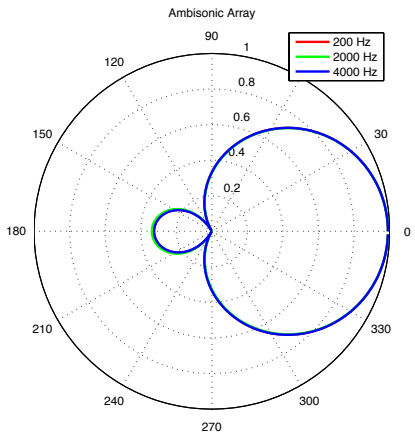


Figure 4.7: Superdirective beampatterns for the Ambisonic array. This is identical to the delay-and-sum solution with the omnidirectional channel given a gain of $\sqrt{2}/2$, confirming that this is the optimum setting.

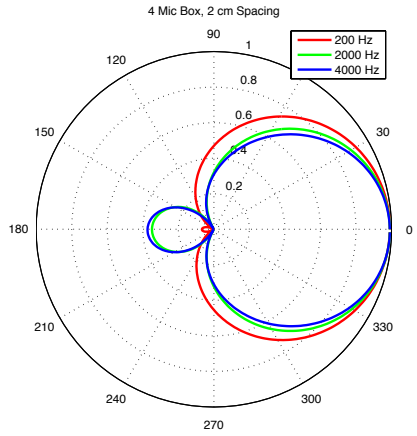


Figure 4.8: Superdirective beampatterns for a four-mic box. Note that the beam-shape is highly consistent across frequency.

4.3.3 Superdirective Beampatterns Formed by Mobile-Sized Planar Arrays

In the Figure 4.7 through 4.10 we will examine the beampatterns formed by superdirective arrays. We find that the superdirective algorithm does a much better job of spatial discrimination at low-frequency than the delay-and-sum algorithm examined above. Additionally, although not constant over frequency, the superdirective results are more consistent, which results in perceptually superior performance.

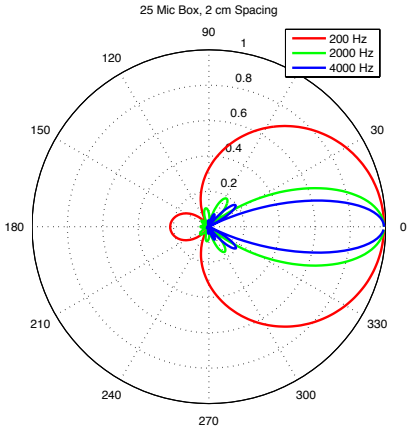


Figure 4.9: Superdirective beampatterns for a 25-mic grid. Results are similar to the four-element box at low frequencies, but much higher directivity is achieved at high frequencies.

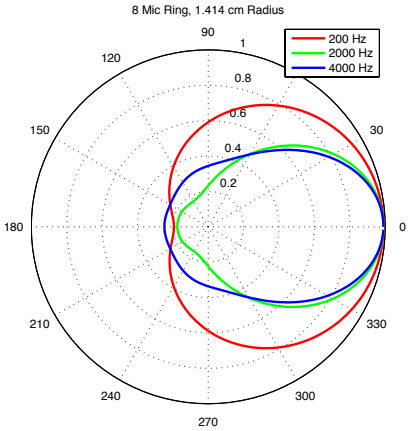


Figure 4.10: Superdirective beampatterns for an eight-mic ring. Note that contrary to the delay-and-sum case, when superdirective beamforming is used, the results dramatically change from the four-element box.

4.4 Chapter Conclusion

In this chapter we have analyzed methods for acquiring the directional audio from an array before applying spatial filters. It can be observed from looking at the polar plots that superdirective beamforming vastly outperforms delay-and-sum beamforming for small apertures and is thus preferred for capturing spatial audio from small devices.

The methods in this chapter mirrored the virtual-speaker technique of acquiring directional audio, then spatializing. In the previous chapter, it was observed that reconstruction error results from the overlap between beampatterns of different look directions. In the next chapter, we will look at an algorithm that aims at directly minimizing this reconstruction error, rather than performing traditional beamforming.

CHAPTER 5

LEAST-SQUARES FILTER DESIGN FOR SPATIAL AUDIO

As discussed earlier in the chapter on Ambisonics, the efficient implementation of Ambisonic reconstruction filters for an array with M microphones results in $M \times 2$ filters, despite the fact there are L ($L \gg M$) look directions. This is not a special case for Ambisonics. In fact, no matter what spatial audio method we use, an efficient implementation will result in the design of just $M \times 2$ filters due to linearity. The efficient implementation hints at the fact that the responses of the beampatterns for each look angle overlap and interfere with one another. Even with large arrays such as the 25-mic grid, this overlap is still a concern. The demonstration of the effects of beampattern overlap on the fitting of HRTF filters was shown in the chapter on Ambisonics. The current methods for minimizing this error have focused on increasing the spatial acuity of the array. An alternative approach is to simply find the best possible $M \times 2$ filters in order to reconstruct the spatial audio scene.

In this chapter we will describe a new technique, one that is capable of directly minimizing reconstruction error. The essential idea behind the method is that if we only have M filters to design the best spatial audio system for a given ear, then rather than designing a large number of filters independently (one per look direction), we should simply design the M (one per microphone) filters directly in order to minimize the spatial filter fitting error across all desired look directions *simultaneously*. By taking into account the complete picture, we should be able to enhance the performance of our system.

5.1 The Least-Squares Filter-Design Technique for Spatial Audio

In order to minimize spatialization error, we must first establish a suitable metric for error. We have previously described the HRTF and displayed one example of the HRTF fit that results from Ambisonic reconstruction. Because the HRTF is a measure of how sound arrives at the two ears, it is also a natural choice for defining error. If we were able to match the HRTF of an individual *exactly*, then in theory, they would also experience the scene exactly as if they had been there. [1]

Our method is to design our filters to directly minimize the squared error of the HRTF fitting in the frequency-domain. As it turns out, the Ambisonic and least-squares solutions are identical for an ideal B-format encoding with $w_W(\theta) = \frac{\sqrt{2}}{2}$ for all θ.[2] Our purpose, however, is not to design filters for existing specialized arrays, but rather to design filters for arbitrary arrays in order to capture spatial audio from any device. The capability of a particular array to emulate characteristics of these specialized arrays, such as forming identical beampatterns in any direction, will also increase its capability of successfully capturing spatial audio.

The following algorithm makes use of standard least-squares optimization in a full column rank (overdetermined) scenario. If we perform this optimization in the frequency-domain, we can make use of orthogonality of the discrete Fourier transform (DFT) at each frequency bin in order to design the filters in a binwise manner. Thus we will arrive at our time-domain filters simply by performing the inverse discrete Fourier transform (IDFT) on the binwise defined frequency-domain filter.

For a particular ear and frequency bin k, we can write the error as

$$\mathcal{E}(k) = \sum_{l=1}^{L} |\text{HRTF}[k, \theta_l] - \mathbf{d}(k, \theta_l)^{\text{T}} \mathbf{h}(k)|^2 \tag{5.1}$$

where θ_l is the direction of a virtual speaker, $\text{HRTF}[k, \theta]$ is a complex scalar element of the HRTF matrix and $\mathbf{d}(k, \theta)^{\text{T}} \mathbf{h}(k)$ is also scalar, being the dot product of the steering vector in direction θ and $\mathbf{h}(k)$, the filter coefficients

[1] This assumes that the HRTFs used are a perfect method for delivering spatialized audio.

[2] Derivation provided in the Appendix B.

to be designed.

Translating this equation into matrix notation, we define:

$$\mathbf{D}(k) = \begin{bmatrix} d_1[k,\theta_1] & \dots & d_M[k,\theta_1] \\ d_1[k,\theta_2] & & d_M[k,\theta_2] \\ \vdots & \ddots & \vdots \\ d_1[k,\theta_L] & \dots & d_M[k,\theta_L] \end{bmatrix} \tag{5.2}$$

$$\mathbf{h}(k) = \begin{bmatrix} H_1[k] & H_2[k] & \dots & H_M[k] \end{bmatrix}^T \tag{5.3}$$

$$\mathbf{b}(k) = \begin{bmatrix} \text{HRTF}[k,\theta_1] & \dots & \text{HRTF}[k,\theta_L] \end{bmatrix}^T \tag{5.4}$$

In the above equations, $\mathbf{h}(k)$ and $\mathbf{b}(k)$ have been written as transposed column vectors for convenience (the complex conjugate is not intended). Our goal then, is to find the filter coefficients $H_m[k]$ to satisfy:

$$\min_{\mathbf{h}} ||\mathbf{b}(k) - \mathbf{D}(k)^T\mathbf{h}(k)||_2^2 \tag{5.5}$$

Since this is an overdetermined system with $\mathbf{D}$ having full column rank, we can use the standard least-squares solution:

$$\mathbf{h}(k)_{LS} = (\mathbf{D}(k)^H\mathbf{D}(k))^{-1}\mathbf{D}(k)^H\mathbf{b}(k) \tag{5.6}$$

Transforming the coefficients via

$$h_m[n] = \frac{1}{N}\sum_{k=-N/2}^{N/2-1} H_m[k]e^{j\frac{2\pi k}{N}n} \tag{5.7}$$

produces the filters $h_m[n]$.

This method generalizes the Ambisonic solution to fit arbitrary array designs when using HRTFs as opposed to minimizing the squared error at a soundfield microphone in the original work.

5.2 Case Studies of Mobile Sized Arrays and the Least-Squares Method

In this section, we will revisit the arrays explored in the previous chapter in order to examine the relative performance of the least-squares method compared to the delay-and-sum and superdirective strategies. It is important to note that simply attaining a better squared error does not guarantee better performance. Psychophysical studies would need to be performed in order to determine the degree to which minimizing squared error increases perceptual accuracy. Because of this, we will not engage in a detailed comparison of the squared error between methods. Furthermore, a comparison of squared error has limited usefulness since, by design, the least-squares method will always obtain the lowest squared error of any method.

Instead, we will take a qualitative approach of examining the fits obtained from the least-squares method in order to verify their usefulness. It is possible (though unlikely) that a solution that minimizes squared error does not form a filter that is perceptually similar to the HRTF compared to some other method with a higher squared error. We must verify that this is not the case. A discussion of the author's perceptual experiences will also be included as additional qualitative analysis.

In the author's perceptual experience using the arrays presented in this thesis, the least-squares method equaled or outperformed the delay-and-sum and superdirectional methods in every case. Of additional interest is the comparison of an ideal Ambisonic array with an array of four omnidirectional microphones. This comparison is interesting because an array of four omnidirectional microphones can be used to simulate a 2D Ambisonic array.[3] Despite the fact the Ambisonic array was designed specifically for spatial audio, to the author's ears, the least-squares method implemented directly on the omnidirectional microphones performs better. Using the least-squares technique, the squared error of four omnidirectional microphones is indeed less than that for an array of two figure-8's and an omni, giving support to the notion that squared error is a suitable metric for obtaining a perceptually accurate spatial audio system.

[3]Directional microphones are essentially combinations of omnidirectional elements and can thus be thought of as an omnidirectional array [26]. See the Appendix A on gradients.

5.2.1 2D Ambisonic Array

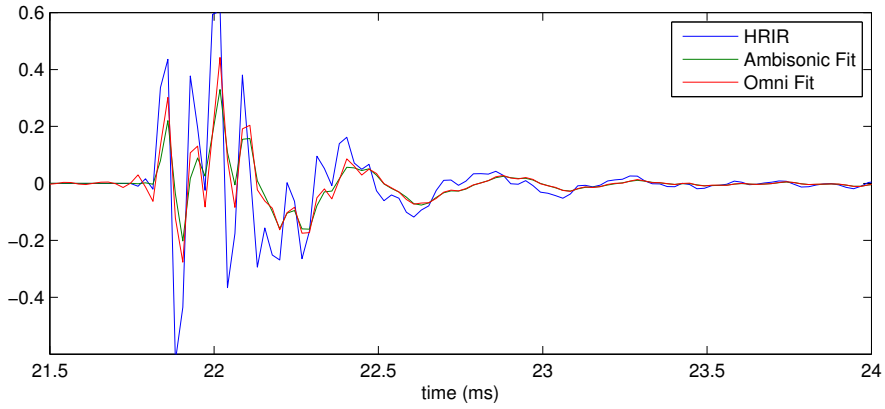
One interesting aspect of our least-squares filter-design technique is that when used with an Ambisonic array, they obtain the same solution. Thus our technique is a generalization of the Ambisonic method for arbitrary arrays.[4] In the previous chapter, it was also shown that the superdirective method also obtained a solution equivalent to Ambisonics. This is because in an Ambisonic array, all microphones are theoretically perfectly co-located and thus forming a beampattern in a given look direction is simply a matter of picking a scalar gain for each channel. The Ambisonic method essentially chooses the gains that result in the highest directivity in a given direction. In this case, finding the highest directivity is equivalent to minimizing the squared error of the HRTF fitting.

In practice, however, microphones will not be perfectly co-located and in addition will not have a flat response across frequency. In particular, gradient microphones have a 20 dB/dec roll-off at low-frequency that must be compensated before reconstruction takes place. If, however, accurate measurements of the microphones used in an Ambisonic system exist, the least-squares method can automatically design these low-frequency compensations into the filters being designed, as well as account for imperfections in co-location. This makes use of the least-squares method in an Ambisonic context superior to the simple method of changing a scalar gain between channels.

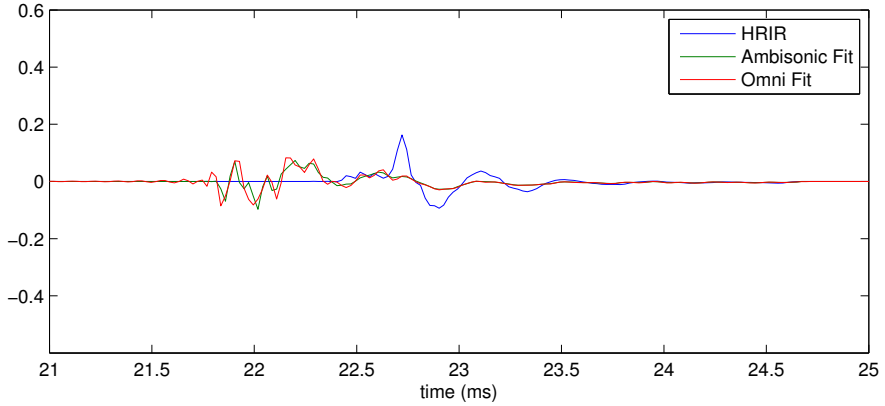
While this is encouraging, the main contribution of the least-squares method is that it enables the use of arrays that are not Ambisonic. In the next section we will show that we can exceed Ambisonic performance with the same number of microphones using a different array.

[4] A derivation of this fact can be found in the Appendix B.

5.2.2 Comparison of an Ambisonic array with Four Omnidirectional Microphones



(a) Comparison of ipsilateral fits for sound incident directly at the ear. Note that the four-mic box is able to achieve a slightly better fit than the Ambisonic array.



(b) Comparison of contralateral fits for sound incident directly at the opposite ear. Note that neither reconstruction forms a close fit.

Figure 5.1: A comparison of HRIR fits for the Ambisonic array and four-mic box for sound at the ipsilateral and contralateral ears.

A 2D Ambisonic array employs three microphones, namely two figure-8s and an omni, while our four-mic box is made up of exclusively omnidirectional microphones. What is the purpose of comparing the two arrays then? Since we have restricted ourselves to using only omnidirectional mics for mobile devices, one way we could simulate an Ambisonic array is by differencing omnis to obtain figure-8s. In order to construct a 2D Ambisonic array from omnidirectional microphones, we must use at least four sensors.

A useful comparison then is to examine the fits between the two arrays with

essentially the same number of elements. One array is specifically designed for spatial audio, while the other is a simple arrangement capable of being built into a mobile device. As it turns out, the array of omnis performs very favorably when compared to the Ambisonic array. This is especially noticeable at the edges, or when sound is incident directly at one of the ears. The image for the four-mic case is capable of being noticeably wider. This is a great benefit to a spatial audio system, since sounds at the edges tend to be more aurally compelling because of the strength of their cues.

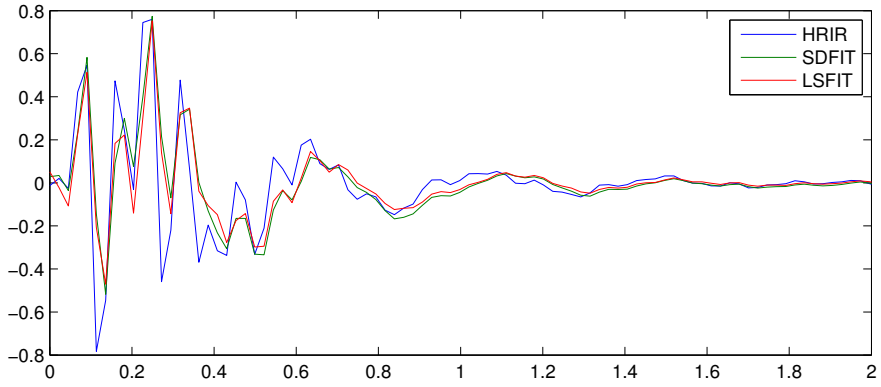
In Figure 5.1 we can see that for the ipsilateral and contralateral fits, the four-mic box performs at least as well as the Ambisonic array. For the ipsilateral ear, it is clear that the four-mic box obtains a superior fitting.

5.2.3 Comparison of the Superdirective and Least-Squares Techniques

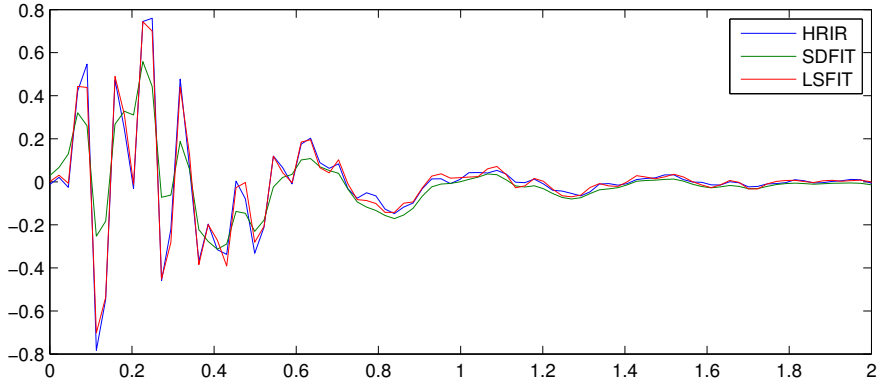
In the previous section we compared an Ambisonic array with the four-mic box. There is not as much use in a comparison when increasing the number of elements in the array, since the four-mic box already obtains superior performance. In the previous chapter, however, we described a technique for using superdirective beamforming to obtain spatial audio from an arbitrary array. How much better does the least-squares method perform, if at all?

The answer is that it depends on the number of microphones used in the array. For a small number of microphones the performance of the two methods is nearly identical. The least-squares method is able to exploit some of the information obtained from designing the filters in concert, but not enough to make a significant difference. As the number of elements increases, however, the least-squares method starts outpacing the superdirective one. The additional degrees of freedom enable the algorithm to take advantage of structure in the HRTFs to make trade-offs to achieve better performance.

In Figure 5.2 we can see that there is little difference between the superdirective and least-squares fits. For the 25-mic case, however, the least-squares method is able to form a nearly perfect fit, while the superdirective method still contains noticeable error.



(a) Comparison of the least-squares and superdirective fits for the four-mic box. While the least-squares fit has a slight advantage in squared error, the perceptual quality is essentially the same.



(b) Comparison of the least-squares and superdirective fits for the 25-mic grid. Note that the extra degrees of freedom begin to allow the least-squares algorithm to outperform the superdirective algorithm.

Figure 5.2: A comparison of the least-squares and superdirective HRIR fits for the ipsilateral ear with the sound incident directly at the ear.

5.2.4 An Analysis of Standard Arrays for Spatial Audio using the Least-Squares Reconstruction Method

In the previous chapter, we analyzed the capabilities for directional discrimination of four different arrays. In this section, we will ignore the ideal Ambisonic array, since the filter-design simply results in placing a scalar gain on each channel. We will, however, examine the four-mic box, eight-mic ring, and the 25-mic grid in more detail. Figures 5.3 through 5.8 show examples of the filters designed as well as the fits obtained for the ipsilateral and contralateral ear for a sound incident directly at one of the ears.

Though not explicitly discussed in the problem formulation, performing the filter-design optimization in bins in the frequency-domain results in finding the best filter under circular convolution rather than linear convolution. If only a few omnidirectional microphones are used or the correlation between channels is small, such as with orthogonal gradient microphones, the filters that are designed will be compact in the time-domain and circularity will not be an issue.

As shown in the figures, however, the filters that are designed may not decay at the edges, and thus significant circular effects will result. Since these filters no longer are able to achieve meaningful results via linear convolution, another approach must be taken. The solution presented in this chapter is to regularize the solution via adding an identity matrix with a small scalar weighting to $\mathbf{D}(k)^H\mathbf{D}(k)$ before inverting. In the next chapter we will present a method for obtaining the filters in the time-domain and thus avoiding the circularity issue altogether.

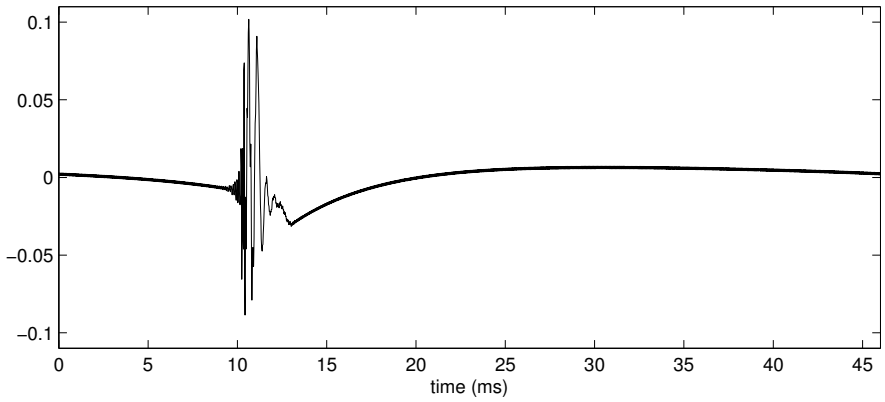
5.3 Weighted-Least-Squares Filter-Design

One extension to the least-squares technique discussed above is weighted least-squares (WLS). While the actual practical implementation of WLS requires further development (principally, methods for choosing the weighting matrix to arrive at desired results), the idea will be described briefly here.

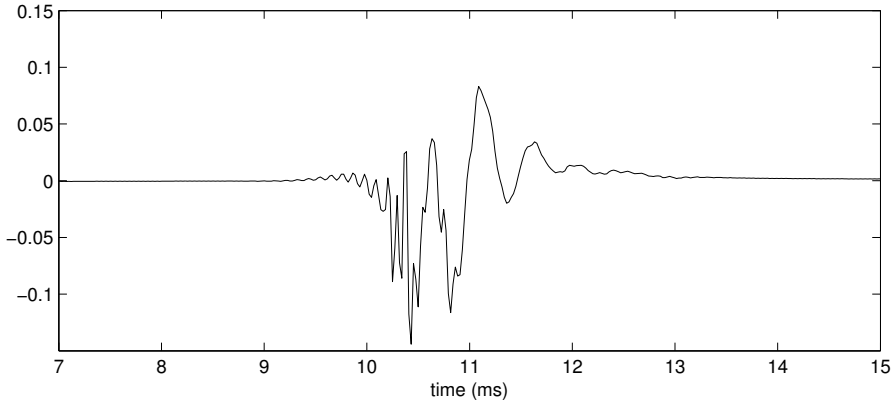
Spatial audio is a perceptual phenomenon, and like most perceptual phenomena, practitioners desire a "knob to tweak." The WLS scheme described below allows a practitioner to decide which angles are of primary importance. For example, if a scene is shot on camera, one might expect that accurate localization of the action on-screen is more important than ambient sounds off-screen. A WLS framework allows the practitioner to decide which errors are important to them, and then weight those errors accordingly.

Another example of how WLS could be used is if a reliable direction-finding system is employed, angles corresponding to direct sound could be weighted higher, while angles from which there are only reflections would receive less weight. If a dynamic direction-finding system is used, the WLS filters could be recalculated in real time in order to track the source.

Other ideas for weighting schemes include weighting based on deficiencies



(a) Example filter for the four-mic box. Note the slow decay to a near zero value at the start and end of the filter. While this filter will work for linear convolution, circularity issues are present.

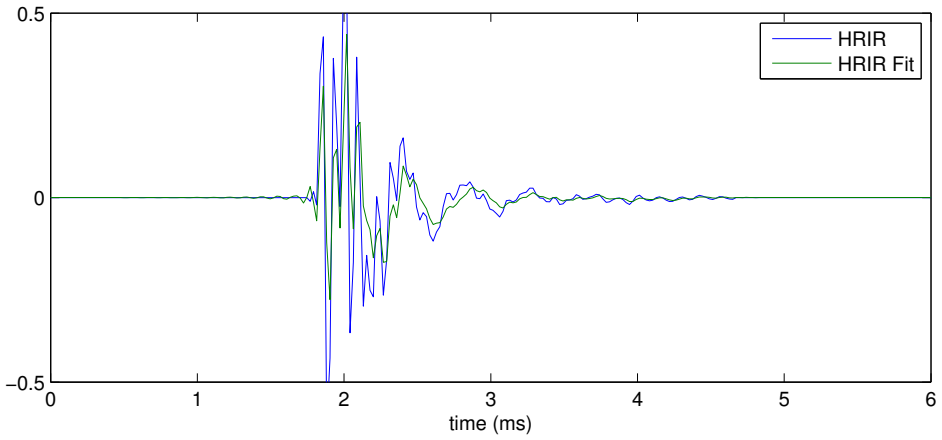


(b) Example filter for the four-mic box with regularization. Note that regularization helps control circularity. This picture has been zoomed in to show filter detail. The main energy of the filter has decayed adequately before the edges of the filter to prevent circular issues.

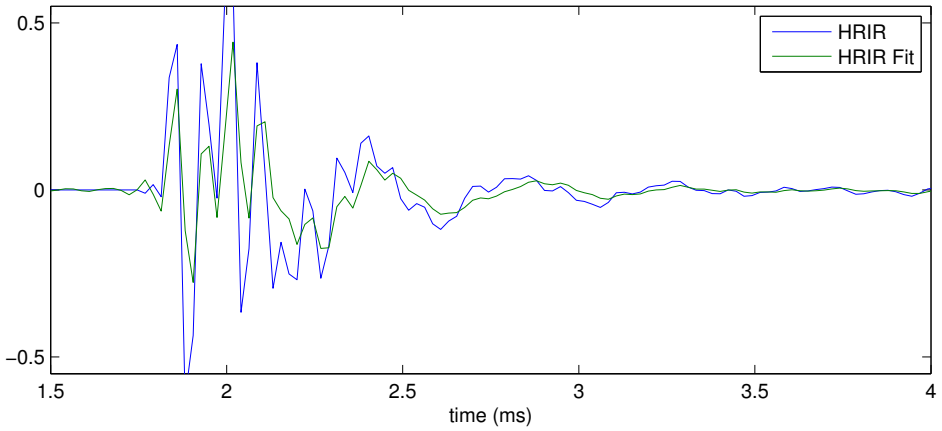
Figure 5.3: Comparison of regularized and non-regularized filters for the four-mic box.

in array geometry. It is common for arrays of omnis to achieve better fittings in the ipsilateral than contralateral ear because the relative energy of that filter is higher. Weighting by the log of the energy in a given direction helps give more weight to the contralateral ear.

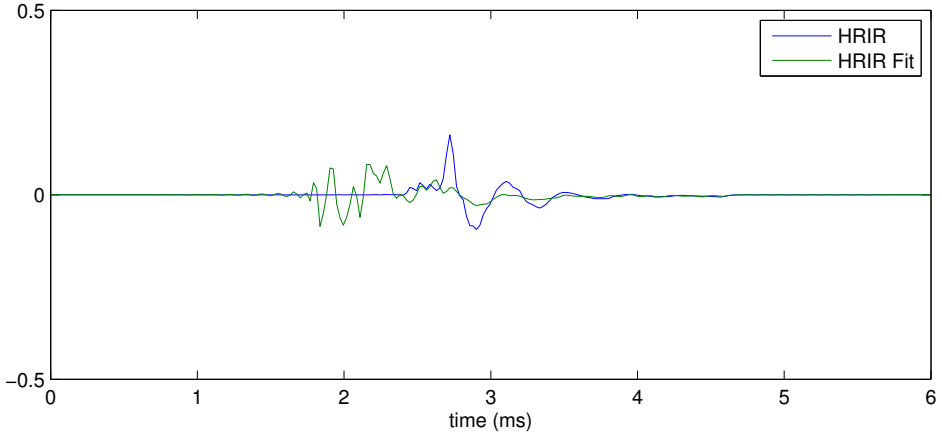
The weighting matrix takes the form of a square matrix with the number of rows and columns equal to the number of virtual speakers. The weight given to each direction is placed on the diagonal. While only diagonal weighting matrices were explored, it may be possible to use off-diagonal entries to improve the results. Because the WLS problem is solved in frequency bands, it would also be possible to make this weighting matrix frequency-dependent.



(a) HRIR fit for the ipsilateral ear with sound directly at the ear using the regularized filters.

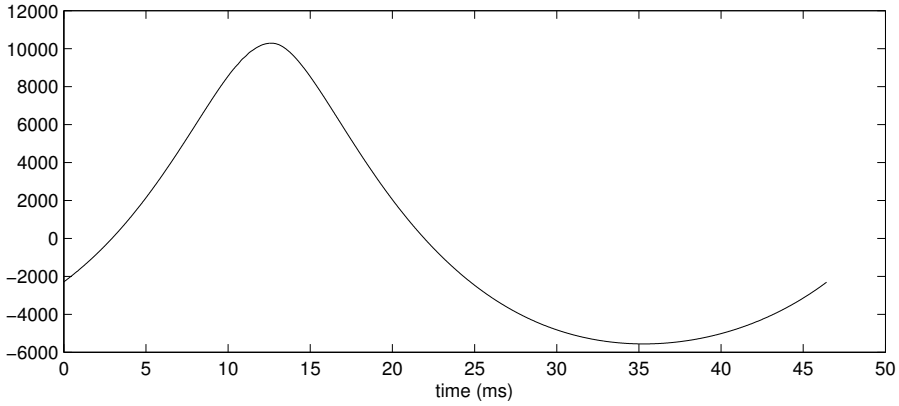


(b) Same plot as above with zoom to show detail.

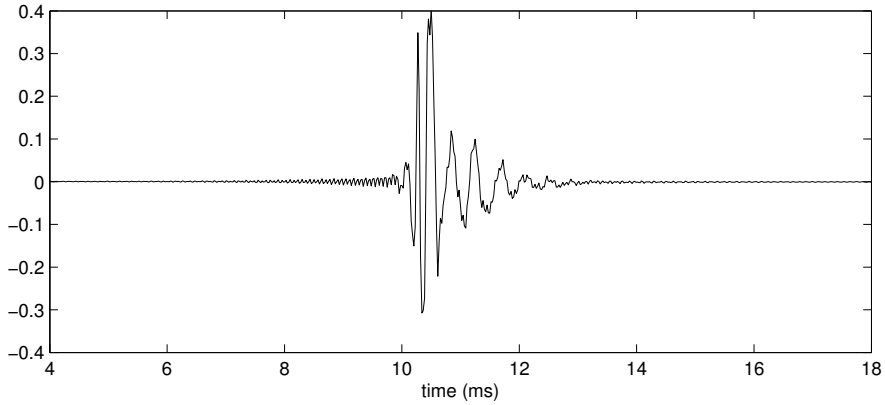


(c) HRIR fit for the contralateral fit with sound incident directly at the opposite ear.

Figure 5.4: Regularized HRIR fits for the four-mic box.



(a) Example filter for the eight-mic ring. Note that circularity issues are present and that the filter energy is large.



(b) Example of a regularized filter for the eight-mic ring. Note that circularity issues have been tamed. Some ringing occurs.

Figure 5.5: Comparison of regularized and non-regularized filters for the eight-mic ring.

Only weightings that were constant with frequency were tested.

A final note about WLS: if the number of desired directions is less than or equal to the number of microphones, the HRTFs can be solved for exactly. This may not be desirable perceptually in practice. Giving all other directions a weight of a much smaller relative value than those given to the source directions would result in highly accurate filtering from those desired angles.

5.4 WLS Formulation

Revisiting the error equation from the LS formulation, yields:

$$\mathcal{E}(k) = \sum_{l=1}^{L} \left| \mathbf{W}(k,\theta_l) \left[HRTF[k,\theta_l] - \mathbf{d}(k,\theta_l)^{\mathrm{T}} \mathbf{h}(k) \right] \right|^2 \tag{5.8}$$

$$\min_{\mathbf{h}} ||\mathbf{W}(k)[\mathbf{b}(k) - \mathbf{D}(k)^T \mathbf{h}(k)]||_2^2 \tag{5.9}$$

$$(\mathbf{W}(k)\mathbf{D}(k))^H \mathbf{W}(k)\mathbf{D}(k)\mathbf{h}(k) = (\mathbf{W}(k)\mathbf{D}(k))^H \mathbf{W}(k)\mathbf{b}(k) \tag{5.10}$$

$$\mathbf{h}(k)_{WLS} = [(\mathbf{W}(k)\mathbf{D}(k))^H \mathbf{W}(k)\mathbf{D}(k)]^{-1} (\mathbf{W}(k)\mathbf{D}(k))^H \mathbf{W}(k)\mathbf{b}(k) \tag{5.11}$$

$$\mathbf{h}(k)_{WLS} = (\mathbf{D}(k)^H \mathbf{W}(k)^H \mathbf{W}(k)\mathbf{D}(k))^{-1} \mathbf{D}(\mathbf{k})^H \mathbf{W}(k)^H \mathbf{W}(k)\mathbf{b} \tag{5.12}$$

where $\mathbf{W}(k)$ is a diagonal matrix with elements w_{ij}:

$$w_{ii} = w(k,\theta_l) \tag{5.13}$$

$$w_{ij} = 0, \text{ for } i \neq j \tag{5.14}$$

where $w(k,\theta_l)$ is the chosen weight for frequency bin k in look direction θ_l.

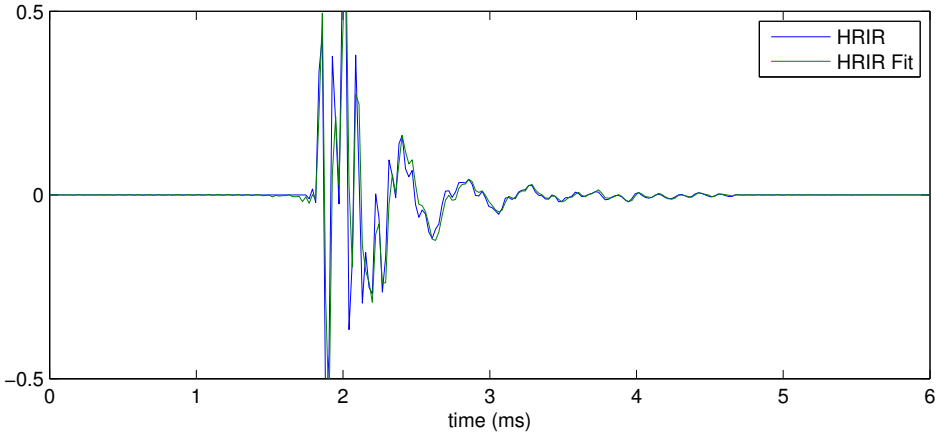
5.5 Conclusion

In this chapter we have presented a new technique for directly minimizing the reconstruction error relative to the HRTF filters in a spatial audio system. Unlike previous methods which have focused on attaining higher directivity from specialized arrays, this technique was designed to provide the maximum performance from an arbitrary array. The generality of this method allows it to work with specialized arrays, such as an Ambisonic array. In this case, the standard Ambisonic and least-squares solutions are equivalent.

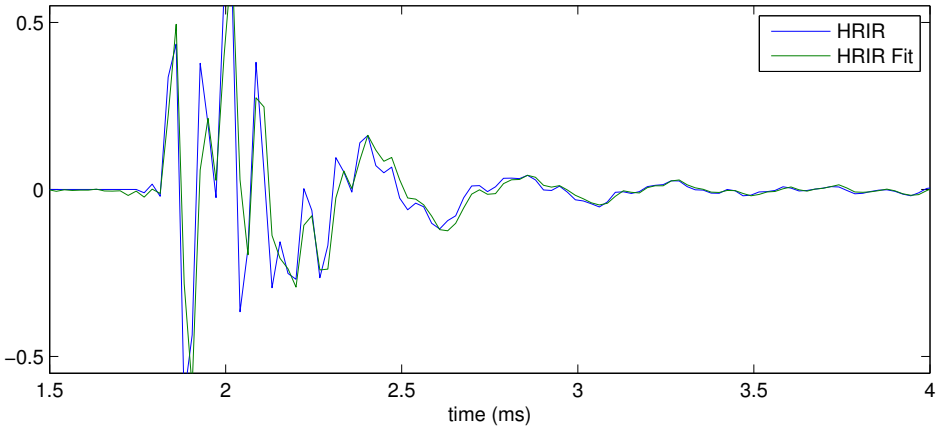
In addition to permitting the use of a broader category of arrays for spatial audio, the least-squares technique also enables using weighted least-squares to specify an angular preference for fitting accuracy. Developing methods for automatically choosing weighting matrices given an auditory scene or incorporation with video merits further research.

Arrays of omnidirectional microphones will result in needing to invert a matrix that is nearly singular. This inversion results in circularity problems. Regularization can help avoid circularity. In the next chapter, we will exam-

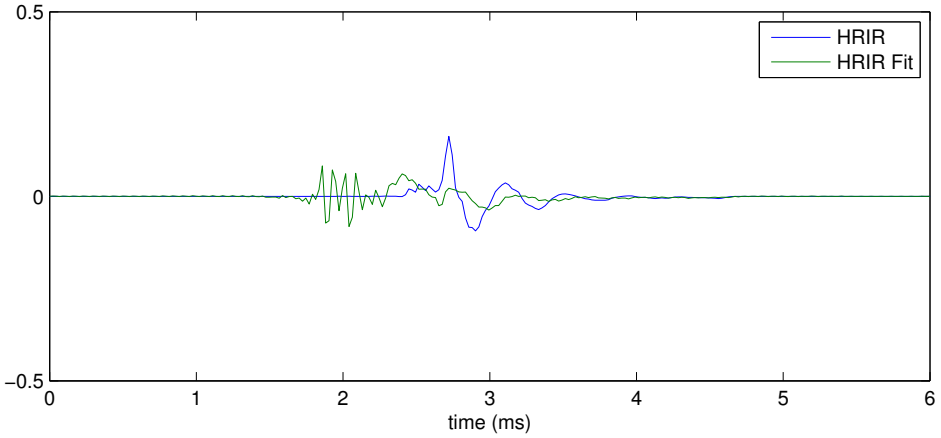
ine implementation of the algorithm in the time-domain, which avoids the circularity issue altogether and does not require regularization.



(a) HRIR fit for the ipsilateral ear with sound directly at the ear using the regularized filters. Note that the ipsilateral fit is noticeably improved over the four-mic box.

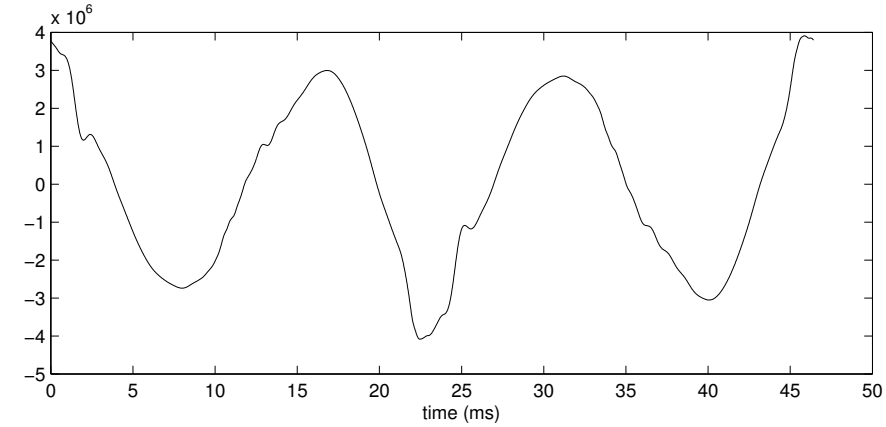


(b) Zoomed in version of the above plot to show detail.

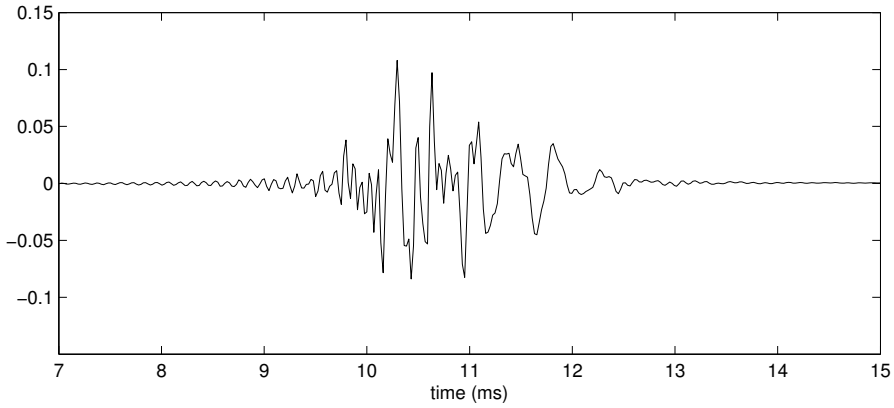


(c) Contralateral fit for the same angle of incidence. Note that the contralateral fit has not been significantly improved over the four-mic case.

Figure 5.6: Regularized HRIR fits for the eight-mic ring.

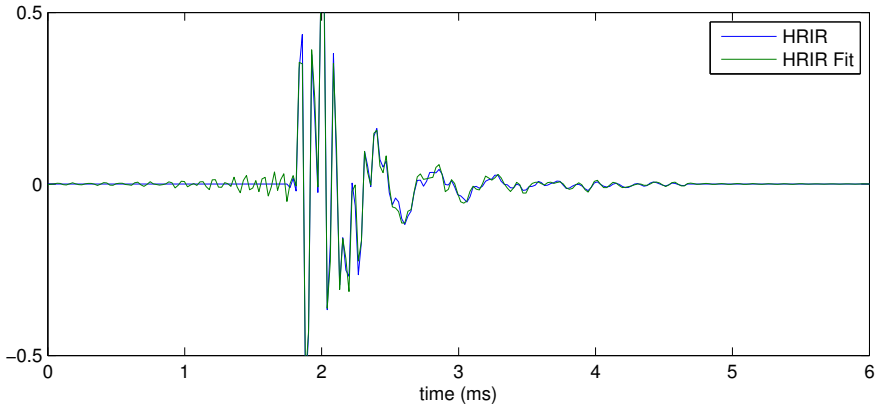


(a) Example filter for the 25-mic box. Again, for large arrays of omnis, the filters that are designed are unusable without regularization.

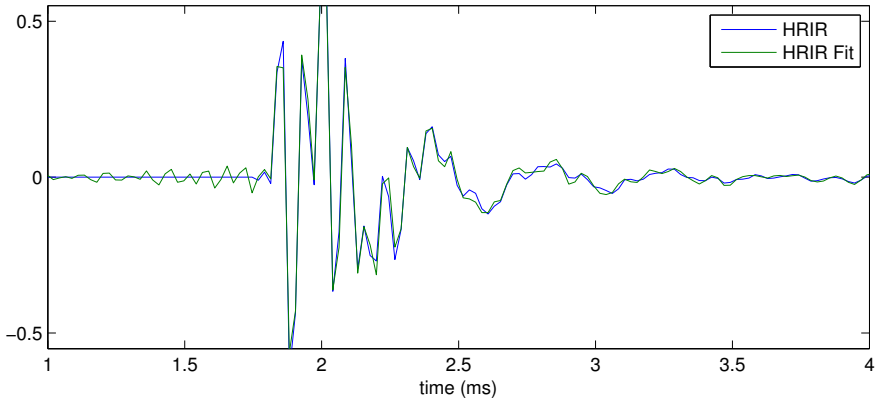


(b) Example of a regularized filter for the 25-mic box. Circularity has been tamed; however, ringing exists and the filter has less resemblance to an HRIR.

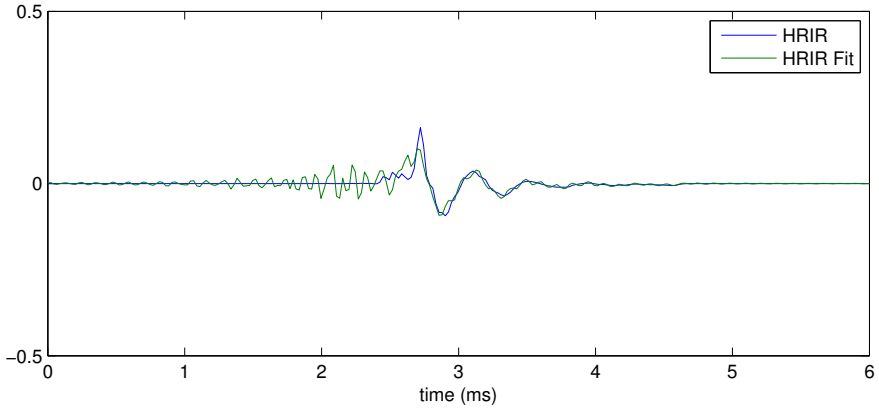
Figure 5.7: Comparison of regularized and non-regularized filters for the 25-mic grid.



(a) HRIR fit for the ipsilateral ear with sound directly at the ear using the regularized filters. Note that the ipsilateral fit is nearly perfect.



(b) Zoomed-in picture of the above filter to show detail.



(c) Contralateral fit for the 25-mic box. Note that this array with the least-squares method is capable of achieving a much closer fit for the contralateral ear than any previous combination.

Figure 5.8: HRIR fits for the 25-mic grid.

CHAPTER 6

PERFORMING LEAST-SQUARES FILTER-DESIGN IN THE TIME-DOMAIN

Given that the least-squares filter-design of the last chapter relied on multiplication in the frequency-domain, there was a possibility for the filter designed via circular convolution to poorly approximate linear convolution. It was demonstrated that linear convolution with the derived filters for certain arrays resulted in poor HRTF fitting, despite the fact that a good fit was obtained using circular convolution. One way of avoiding this circularity problem is to solve for the filters in the time-domain. Since the time-domain solution is derived from linear convolution, the issue of circularity is avoided altogether.

The main advantage of working in the frequency-domain is its simplicity, such as the method of solving for the solution in individual frequency bins described in the previous chapter. Because the time-domain setup cannot be easily decoupled into simpler problems, implementation can be problematic due to memory usage when computing the correspondingly larger matrices, since we must compute every sample of the filter at the same time. Therefore, after presenting the time-domain theory, practical implementation details will be discussed at the end of the chapter.

6.1 Convolution by Matrix Multiplication

In order to convolve a length N vector $\mathbf{a}$, with a length P vector $\mathbf{x}$, the vector $\mathbf{a}$ can be formed into a matrix $\mathbf{A}$ to perform the convolution via matrix multiplication. Since the solution of a convolution between length N and length P vectors is length $N+P-1$, the matrix $\mathbf{A}$ will have $N+P-1$ rows, and P columns. The form of the matrix is

$$\mathbf{A} = \begin{bmatrix} a[1] & 0 & \dots & 0 & 0 \\ a[2] & a[1] & 0 & \dots & 0 \\ a[3] & a[2] & a[1] & 0 & \dots \\ \vdots & & \ddots & \ddots & \\ a[n] & a[n-1] & \dots & a[2] & a[1] \\ 0 & a[n] & a[n-1] & \dots & a[2] \\ \vdots & & \ddots & \ddots & \\ 0 & \dots & 0 & a[n] & a[n-1] \\ 0 & 0 & \dots & 0 & a[n] \end{bmatrix} \tag{6.1}$$

Using matrix $\mathbf{A}$ we can express convolution as

$$\mathbf{a} \star \mathbf{x} = \mathbf{A}\mathbf{x} \tag{6.2}$$

where $\star$ represents linear convolution.

6.2 Least-Squares Filter-Design in the Time-Domain

Just like in the frequency-domain, the problem takes the form

$$\min_{\mathbf{h}} ||\mathbf{G}\mathbf{h} - \mathbf{b}||_2^2 \tag{6.3}$$

The matrices will look somewhat different, however, since we are working in the time-domain.

In the time-domain formulation, $\mathbf{h}$ is a concatenated vector of three FIR filters, each of length P, and thus $\mathbf{h}$ will have length $M \times P$, where M is the number of microphones in the array.

$\mathbf{G}$ is a matrix representing the convolution of the "time-domain steering vectors" with the vector $\mathbf{h}$. In Chapter 4 we defined a steering vector as a vector of complex numbers giving the phase of an incoming wave at each microphone according to its angle of incidence and frequency. Essentially, a steering vector describes the propagation of the wave and its arrival at each microphone.

To accomplish a similar task in the time-domain, the propagation at a given angle of incidence must be described for all frequencies simultaneously. One

way of doing so is to write the steering vectors as a linear filter $\mathbf{a}$ that, when convolved with the incident audio, delays the signal appropriately according to the length of the path. We can find this time-domain filter by taking the inverse DFT of all the frequency bins of our frequency-domain steering vector for a given angle and microphone. Since the delay will, in general, be non-integer, this filter will take the form of a truncated sinc function. It is beneficial therefore to make this filter fairly long and to window such that the edges do not cause artifacts when inverse filtering. It is also important to ensure that the peaks of these sinc functions lie near the middle of the filters for every microphone and angle of incidence to achieve accurate results.

To construct a convolution matrix $\mathbf{A}(m, \theta)$ for each microphone and angle of incidence, these sub-matrices are tiled inside a larger matrix $\mathbf{G}$

$$\mathbf{G} = \left[\begin{array}{c|c|c|c} \mathbf{A}(m_1, \theta_1) & \mathbf{A}(m_2, \theta_1) & \dots & \mathbf{A}(m_M, \theta_1) \\ \hline \mathbf{A}(m_1, \theta_2) & \mathbf{A}(m_2, \theta_2) & \dots & \mathbf{A}(m_M, \theta_2) \\ \hline \mathbf{A}(m_1, \theta_3) & \mathbf{A}(m_2, \theta_3) & \dots & \mathbf{A}(m_M, \theta_3) \\ \hline \vdots & \vdots & & \vdots \\ \hline \mathbf{A}(m_1, \theta_L) & \mathbf{A}(m_2, \theta_L) & \dots & \mathbf{A}(m_M, \theta_L) \end{array}\right] \tag{6.4}$$

The filter to be solved for has the form

$$\mathbf{h} = \begin{bmatrix} \mathbf{h}_{m_1} & \vdots & \mathbf{h}_{m_2} & \vdots & \dots & \vdots & \mathbf{h}_{m_M} \end{bmatrix}^T \tag{6.5}$$

and the HRIR vector

$$\mathbf{b} = \left[\begin{array}{c} \mathbf{hrir}(\theta_1) \\ \hline \mathbf{hrir}(\theta_2) \\ \hline \mathbf{hrir}(\theta_3) \\ \hline \vdots \\ \hline \mathbf{hrir}(\theta_L) \end{array}\right] \tag{6.6}$$

Thus we can solve equation (6.3) using the standard least-squares solution from the previous chapter

$$\mathbf{h}_{opt} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{b} \tag{6.7}$$

The vector $\mathbf{h}$ must be put into a form where a solution to the least-squares problem returns a useful result. Since a causal filter is being designed, the HRIR must be zero-padded to a position in the vector where the designed filter can manipulate the time-domain steering vectors into a reasonable solution.

6.3 Solution via Gradient Descent

Since the objective function we are trying to minimize is quadratic in nature, gradient descent is a good choice for numerically finding a solution.

Taking the gradient of

$$F(\mathbf{x}) = ||\mathbf{G}\mathbf{h} - \mathbf{b}||^2 \tag{6.8}$$

we obtain

$$\nabla F(\mathbf{x}) = 2(\mathbf{G}^T\mathbf{G}\mathbf{h} - \mathbf{A}^T\mathbf{b}) \tag{6.9}$$

and equating with zero, we can solve for the minimizing solution. The gradient descent algorithm is

$$\mathbf{h}_{n+1} = \mathbf{h}_n - \mu\nabla F \tag{6.10}$$

$$\mathbf{h}_{n+1} = \mathbf{h}_n - 2\mu(\mathbf{G}^T\mathbf{G}\mathbf{h} - \mathbf{G}^T\mathbf{b}) \tag{6.11}$$

where μ is the step size at each iteration. As noted above, calculating the full $\mathbf{G}$ matrix requires a large amount of memory if we desire large filters or a large number of microphones.

$\mathbf{G}$ has $(N+P-1)\times L$ rows and $P\times M$ columns, where N is the length of the time-domain steering vector, P is the length of the filter being solved for, M is the number of microphones in the array, and L is the number of virtual speaker angles. In general, L will be much larger than M. Therefore, $\mathbf{G}^T\mathbf{G}$

has reduced dimensions of $(P \times M)$ by $(P \times M)$, since $\mathbf{G}$ is a tall matrix. $\mathbf{G}^T\mathbf{G}$ can be computed by only storing one or two convolution matrices, $\mathbf{A}(m, \theta)$, at a time and iteratively summing their inner products. In a similar fashion, we can compute $\mathbf{G}^T\mathbf{b}$ by a summation of the inner products of individual convolution matrices and the concatenated HRIR vector. Since $\mathbf{G}^T\mathbf{G}$ and $\mathbf{G}^T\mathbf{b}$ do not change, we can precompute them before performing gradient descent for fast computation.

CHAPTER 7

QUICK AND DIRTY SPATIAL AUDIO FROM TWO MICROPHONES

Throughout this thesis we have examined various techniques for reconstructing a spatial audio scene from an arbitrary array of microphones. Despite acknowledging the constraints placed on the arrays in real-world devices, having access to even four-microphones is uncommon. Currently, it is still the norm for most mobile devices to have only one or two microphones. While there is little that can be accomplished with a single microphone, the two-microphone case still has merit and is worth examining due to the ubiquity of two-microphone devices.

Before delving into the possible reconstruction schemes for two element arrays, let us first examine their fundamental limits. The first limit is spatial aperture. Since only two microphones are used, it is not possible to design an array that is ideal for capturing both low and high frequencies. Therefore the microphone spacing must be chosen in order to compromise between the two extremes. This makes a gradient implementation[1] especially attractive with a two-microphone array, since constant spatial discrimination can be achieved across frequency.

In addition to the limitations on optimizing array aperture, it is also important to recognize that the use of two microphones will allow only left/right imaging and will be unable to place sounds at the front and back. While front/back localization represents a weak cue, the arrays described in the above chapter all had the capability of easily incorporating head tracking. While some type of vestibular stimulation may still be possible in the two-microphone case, it is not possible to attain the symmetric (or near symmetric) radial response that was so attractive in the previously studied arrays and thus the angular distance of the head turn will be somewhat restricted.

Despite these limitations, compelling spatial audio can be obtained from just two microphones. In personal listening tests, we have found that the

[1] See Appendix A.

majority of the static (not head tracked) spatial audio experience comes from the dominant left/right cues.

7.1 Method 1: Opposite-Facing Cardioids

In this method, we simply construct opposite-facing cardioids, then apply the extreme left pair of HRTFs to the left-facing cardioid and the extreme right pair of HRTFs to the right facing cardioid. Despite the use of just two spatial filters, intermediate azimuthal locations are still handled gracefully due to the linear combination of the time delays.

7.2 Method 2: Least-Square Fitting

The main disadvantage of the above method is that it uses just two spatial filters. We can improve the performance of our system by also solving for intermediate azimuths (albeit without front/back information) by performing the standard least-squares reconstruction algorithm described above for two microphones.

There are a few advantages to using the least-squares method. One advantage is that it does not require finding an inverse filter to correct the low-frequency response of the gradient array. A second advantage is that the overall scene will be imaged more accurately than the two cardioid method.

A disadvantage of the least-squares method is that it will not reproduce extreme left and right angles as well as the two cardioid case. It is possible, however, to control the width of the image using the weighted-least-squares technique described above. For example, giving all other angles except the extreme sides a negligible or zero weight approximates the opposite facing cardioids method above. Using a less aggressive weighting would allow solutions in between to be obtained as well.

CHAPTER 8

CONCLUSIONS

In his 2008 paper [28], Bruce Wiggins asked, "has Ambisonics come of age?" He concluded that, yes, technology had come far enough to make Ambisonics a reality. Despite the technological capability for Ambisonics, spatial audio has still been restricted to a hobbyist market. Maybe the limitation is not the existence of technology, but rather the convenience of use to the general public. Therefore, a more general question we might ask is, "has spatial audio come of age?" The answer to this question I believe is answered not via technology, but by platform delivery. Have we reached a stage where we can make spatial audio attractive to the general public due to ease of use, availability of content, and compelling experience?

Commercialization of spatial audio has failed up to this point because of it has been expensive and inconvenient to record and play back. Now more than ever, there is a platform capable of delivering personal content through cheap methods such as headphones or binaural cross-talk cancellation. Furthermore, the small screen size and public usage of mobile devices begs for the more immersive experience afforded by spatial audio. The ability to capture content as well only furthers their attractiveness as an immersive entertainment device.

The basic inspiration behind this thesis was developing a method to bring spatial audio to mobile devices within a realistic budget, geometry, manufacturing, and computational constraints. It is also important to keep in mind what potential capabilities we want our spatial audio systems to have. First and most importantly, we want the ability to acquire high-quality spatial audio. We have demonstrated that by using superdirectional techniques (either by employing gradients, MVDR, or the least-squares method developed specifically for spatial audio), it is possible to bring spatial audio to the mobile platform without compromise. In particular, in the chapters on least-squares solutions, we demonstrated that the less restricted arrays of om-

nidirectional microphones were able to achieve a better fit than the current Ambisonic arrays employing an equivalent number of omnidirectional microphones. This encouraging technique should be applied to other non-mobile arrays in the future for comparison with higher-order Ambisonic techniques currently in use.

Another important capability for spatial reconstruction techniques is the ability to enable head tracking. As discussed in Chapter 2, Wallach [12] demonstrated that vestibular cues, or cues from active head movements, always override pinna cues. Since pinna cues in spatial audio systems are degraded and in general not personalized, head tracking allows the user to maintain the ability to judge front from back as well as elevation. The efficient implementations described in this thesis are perfect for real-time implementation. As long as the array used is reasonably symmetric, head tracking can be effectively employed.

Spatial audio has been around for decades, capturing the imagination of audio enthusiasts worldwide. Implementing the technology on real-world devices with real audio capture and playback capability is possible. It is now simply a matter of convincing manufacturers that it is a feature that people want, and growing the technology is simply a matter of putting it into people's hands.

APPENDIX A

GRADIENTS AND DIFFERENTIAL MICROPHONE ARRAYS

In this appendix we will explore building differential microphone arrays from omnidirectional elements. In the context of this thesis, differential arrays have capabilities that are attractive for spatial audio. Since differential arrays are useful in other applications as well, we will present a fairly general treatment that can be adapted to specific applications as appropriate.

Traditional beamforming is performed by adding the phased outputs of a microphone array so that the waves from a direction of interest sum, while waves arriving from other angles are subject to cancellation and are attenuated. An alternative type of spatially selective array can be created with different properties (both good and bad) by differencing the microphones. A combination of such elements forms what is known as a differential microphone array.

In the previous chapters we looked at various methods for soundfield reconstruction over headphones, including Ambisonics, delay-and-sum beamforming, superdirective beamforming, as well as a least-squares fitt of the HRTF. We looked at how the results changed when an array of omnidirectional microphones was used, as opposed to an array of figure-8's. We found that figure-8's, or gradients, have many qualities that make them advantageous for beamforming on a small device, as well as other qualities that make them useful for wideband audio beamforming in general. In this chapter we will look deeper into how differential arrays work, as well as how to improve their directivity by adding more microphones.

A.1 Differential Arrays on Mobile Devices

Little has been written about gradients and their use in forming differential arrays since the original work by Harry Olson in the 1940s [29]. Recording

studios, which often embrace the old rather than the new, almost exclusively use cardioid-type microphones, which are single-capsule differential devices. However, one of the earliest directional microphone designs, the ribbon microphone, which has a figure-8, or gradient response, has been finding a renewed sense of popularity. In the context of a recording studio, this has more to do with their tonality rather than their ability to perform complex spatial filtering. One exception is the mid-side recording technique [30] that combines a figure-8 microphone coincident with a cardioid to create a stereo image. The advantage of this technique over the use of two directional microphones in an XY configuration, is that the signals from the figure-8 and cardioid can be blended in variable amounts in order to create a wider or a more centrally focused image. This flexibility gives the recording engineer more control during the mixdown phase of production.

In this appendix, we will demonstrate some alternative uses of gradients in the hope they might find favor with a new audience. In the chapter on Ambisonics we examined the figure-8, or first-order gradient. Olson also derived higher-order gradient responses, which have a higher directivity but maintain a uniform beampattern across frequency. In this appendix we will investigate the gradient response and its characteristics, as well as demonstrating how to use gradients to create other higher-order differential responses, such as second- and third-order cardioids, for capturing spatial audio. Of usefulness to the practicing engineer, we will examine in detail how to obtain these differential responses as a combination of omnidirectional microphones and how they can be arranged in order to steer a gradient or cardioid response of various orders in an arbitrary direction in the plane using the minimum number of omnis. This will benefit applications, such as mobile devices, where the type, spacing, and quantity of microphones are highly restricted.

A.1.1 Advantages of Differential Microphone Arrays

We will briefly describe the theory of differential microphone arrays, in particular creating various polar patterns. In our application, differential microphone arrays have three important advantages:

1. They are superdirectional.

2. They are physically small compared to the wavelengths of interest.

3. They have uniform beampatterns across frequency.

A superdirectional array is defined as an array that has a higher directivity than is possible by summing delayed versions of the individual channels with uniform gain (i.e., some version of delay-and-sum beamforming). This is of obvious interest to us, as the aperture size of our array is limited and higher directivity helps to minimize overlap between virtual speakers.[1] Given that we cannot simply increase the aperture size of our array, some form of superdirectional array is necessary.

The third advantage, uniform beam shape across frequency, is most important as it helps make the reconstructed experience perceptually superior. This benefit does not come without cost, however. The added directionality comes at the sacrifice of gain, especially at low frequency (where directional gains are the highest). While adding two channels gives a 3 dB boost over noise, subtracting channels necessarily reduces gain. The gain is reduced according to the difference in phase of the wavefront at the two microphones. Since the phase difference between the two microphones is very small at low frequency (where the wavelength is large), the loss in gain is high, which in turn greatly increases low-frequency self-noise.

A final advantage, which may be important in some applications, is that differential arrays do not require complex processing. The superdirective technique required knowledge of or an assumption about a noise field. The least-squares technique was limited to spatial audio. In addition, both techniques required accurate steering vectors in order to obtain solutions. A differential array on the other hand simply requires a fixed gain and a delay. This is an attractive property in applications where simple processing is required.

A.2 The Differential Microphone Array

The response of a first-order differential microphone array is described by the limaçon, parameterized by α and β

$$E(\theta) = \alpha + (1 - \alpha)cos(\theta - \beta) \tag{A.1}$$

[1] See the chapter on Ambisonics.

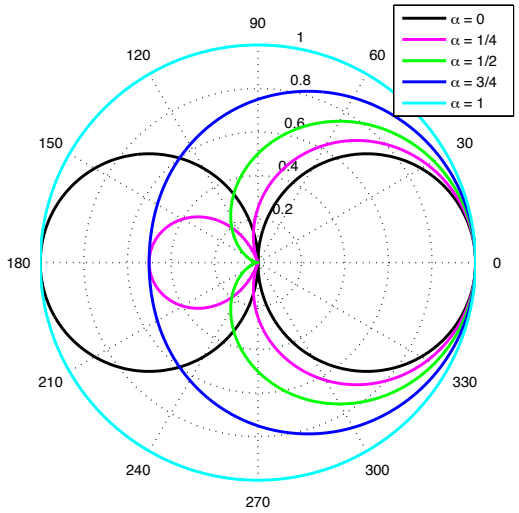


Figure A.1: Plots of the limaçon for various values of α. When $\alpha = 0$, the response is a figure-8; when $\alpha = 1/2$, the response is a cardioid; when $\alpha = 1$, the response is omnidirectional.

where $0 \leq \alpha \leq 1$ and $0 \leq \beta < 2\pi$. The α parameter makes a trade-off between an omnidirectional response ($\alpha = 1$) and a figure-8 ($\alpha = 0$), while β rotates the response in the xy-plane. Responses for various values of the parameter α are shown in Figure A.1. In order to realize these patterns given two omnidirectional microphones, the microphone in the intended steering direction must be delayed. Given an array of two elements, for a desired α, the time delay needed between the two microphones is

$$\tau = \left(\frac{d}{c}\right)\left(\frac{\alpha}{1-\alpha}\right) \tag{A.2}$$

For a thorough derivation of Eq. A.2 as well as a detailed presentation of higher order differential arrays, see [31].

A.3 The First-Order Gradient Microphone Array

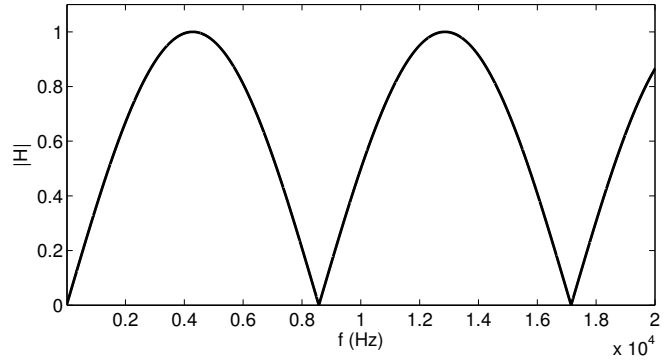
The gradient response is a special case of the differential array with $\alpha = 0$, and therefore there is no time delay between the channels (Eq. A.2). To obtain the figure-8, we simply take the difference of the two omnidirectional channels.

In the design of a gradient microphone array there is a fundamental trade-

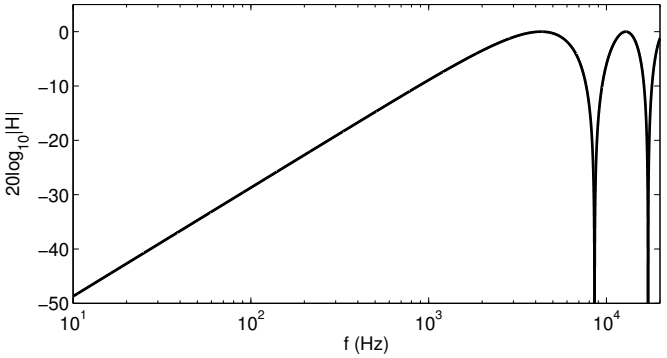
off between low-frequency roll-off and comb-filtering. This trade-off is caused by the relative phase between microphones.[2] Since we are differencing the two channels, nulls appear at integer multiples of the wavelength corresponding to the distance between the two microphones. Therefore, placing the microphones closer together pushes the first null to a higher frequency. The nulls in the on-axis frequency response occur at

$$f_{null}(n) = n \times \left(\frac{c}{\lambda_d}\right),\ n \in \mathbb{N} \tag{A.3}$$

with the first null occurring when $n = 1$.



(a) HRIR; note the timing and level cues between the two ears.



(b) Log-log plot of the same response showing the 20 dB/dec roll-off at low frequency.

Figure A.2: Linear and log plots of an on-axis gradient-microphone frequency response.

While it is important to avoid comb filtering, we do not want to place the

[2]See Figure A.2.

microphones as close to one another as possible, either. In addition to the nulling at high frequency there is a roll-off of 20 dB/dec (6 dB/oct) starting at the magnitude peak which occurs at

$$f_{peak} = \frac{c}{2 * \lambda_d} = \frac{f_{null}(1)}{2} \tag{A.4}$$

the frequency at which the microphones are 180° out of phase. Therefore, we want to place the microphones close enough together to push the first null outside the frequency range of interest, but no further in order to preserve as much low-frequency gain as possible.

Another way of viewing the on-axis behavior of a differential microphone array is that it applies an FIR differencing filter to the incoming wave. For clarity, assume that the distance between the two microphones corresponds exactly to an integer sample delay. Therefore, the FIR filtering operation would be:

$$\mathbf{h} = [1\ 0\ 0 \ldots 0\ -1] \tag{A.5}$$

where the number of zeros corresponds to the time it takes for the incoming wave to arrive at the second microphone relative to the first, or

$$\tau_d = \frac{d}{c} \tag{A.6}$$

$$sample\ delay = \tau_d * f_s \tag{A.7}$$

where f_s is the sampling rate. Therefore in the complex plane, this has the effect of placing a single zero at DC on the unit circle, causing a low-frequency roll-off for which there is no stable inverse. We can, however, compensate for this roll-off by placing a single pole at DC somewhere near the unit circle.

While this filter will reclaim some of the gain that is lost, it will also have the effect of boosting the noise at low frequency. This effect can be mitigated somewhat by placing the pole farther from the unit circle (sacrificing low frequency gain), or employing a noise reducing scheme, such as a Wiener filter.

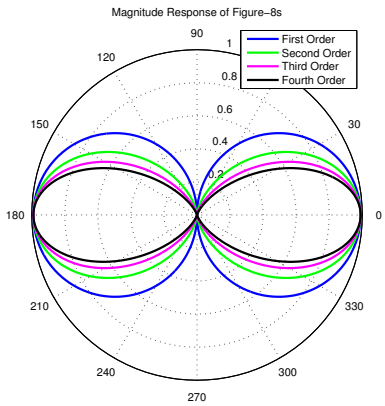


Figure A.3: A comparison of magnitude responses for first- through fourth-order gradients.

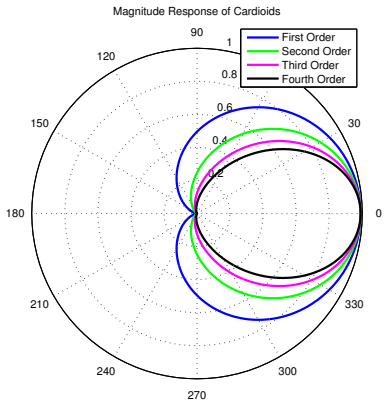


Figure A.4: A comparison of magnitude responses for first through fourth order cardioids.

A.3.1 Higher-order gradients

A first-order gradient has the form

$$E(\theta) = cos(\theta - \beta) \tag{A.8}$$

A gradient of order n is then

$$E(\theta) = cos^n(\theta - \beta) \tag{A.9}$$

For example, a second-order gradient on the x-axis is simply $cos^2(\theta)$. See Figure A.3 for a comparison of various gradient responses and Figure A.4.

In [29], Olson demonstrated how to obtain higher-order gradients by a linear combination of first-order elements. In order to create a second-order gradient, simply difference two first-order gradients chained end-to-end. Naively this task could be performed with four omnidirectional microphones, using two omnis to form each first-order gradient, and then differencing the first-order outputs. We can achieve the same effect with three microphones if the outer two elements share the center to form the individual gradients.

In the previous section, we showed that the differencing operation that forms the gradient results in a low-frequency roll-off. A second differencing operation will increase this roll-off, resulting in the need for further compensation. This added directivity comes at a price, then, and the decision to use higher-order patterns must be made carefully.

In this section we have established how to form the basic first- and second-order gradients $cos(\theta-\beta)$ and $cos^2(\theta-\beta)$. Later in the appendix we will show how to use the minimum number of these elements, or "building blocks" in order to steer a higher-order pattern in an arbitrary direction in the plane.

A.4 Differential Beamforming in an Arbitrary Direction

In the chapters beginning with Ambisonics we laid the groundwork for a general spatial audio system. The basic idea is that we want to form beampatterns in every direction of interest (with identical beampatterns if possible), then apply some type of HRTF or spatialization filter to each output.

We want to point our directional response in as many directions as possible, while using the fewest number of microphones. For an Ambisonic array, this means using four cardioids in a tetrahedron (A-format), or three figure-8s and an omni (B-format). In this section we will examine the design of differential arrays from omnidirectional microphones in order to form identical responses in any direction in the plane.[3]

[3]Performing this task in three dimensions is also possible, but beyond the scope of our mobile application.

A.4.1 The first-order differential array in the plane

From our study of Ambisonics, we know that we can form a response from the cardioid family in any direction in the plane using just two figure-8s and an omni. In general the response will be

$$E(\theta) = \alpha + cos(\theta - \beta) \tag{A.10}$$

which describes a member of the cardioid family pointing in the arbitrary direction β. Our goal is to decompose this response into the "building blocks" discussed above. Since we know how to create arbitrary orders of gradients by differencing omni's, if we can write the general response in terms of these building blocks, then we do not need to build a separate array for each direction of interest, making the design much more practical in terms of the number of sensors used. We can do this by decomposing the steerable equation into channels made up of a gain multiplied by a building block steered in a fixed direction. Using a basic trigonometric identity

$$\alpha + cos(\theta - \beta) = \alpha + cos(\beta)cos(\theta) + sin(\beta)sin(\theta) \tag{A.11}$$

which is a linear combination of an omni (α), with the gradient channels $cos(\theta)$ and $sin(\theta)$.[4] Therefore we have derived the Ambisonic equation from the previous chapter and we can use a linear combination of the channels

1. $W = \alpha$
2. $X = cos(\theta)$
3. $Y = sin(\theta)$

We can construct this differential array from just four or five omnis. An example layout is given in Figure A.5. By differencing two omnis on the x-axis, we obtain the X channel, two omnis on the y-axis gives us the Y channel, and we can obtain the W channel by summing all four omnis, or placing a fifth omni at the center of the array.

[4] $sin(\theta) = cos(\theta - \pi/2)$, or a gradient response in the $\pi/2$ direction.

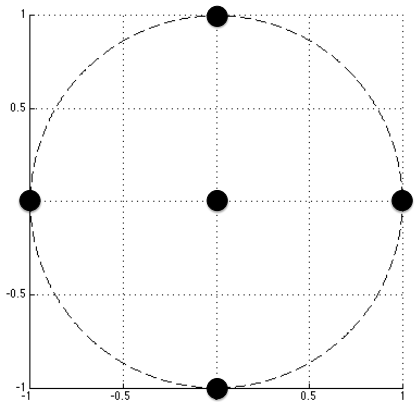


Figure A.5: Omnidirectional microphone layout for a first-order differential array. The center omni may be omitted.

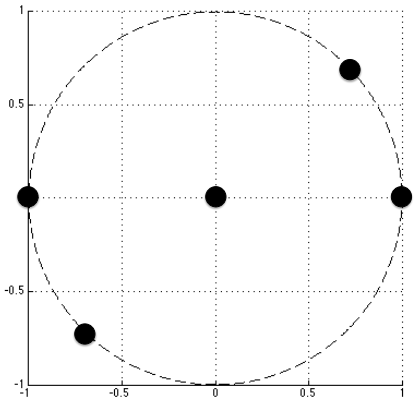


Figure A.6: Omnidirectional microphone layout for a second-order differential array.

A.4.2 The second-order differential array in the plane

In the previous section we examined how to create first-order responses in an arbitrary direction with just four or five omnidirectional microphones. Forming second-order responses in an arbitrary direction also requires just five sensors. An example layout is given in Figure A.6.

We are looking for responses of the form

$$[\alpha + cos(\theta - \beta)]^2 \tag{A.12}$$

In this case, we will be looking for the building blocks to be omnidirectional, first-order gradient, and second-order gradient. Therefore we want to expand

equation (A.12) to be in terms of the functions $\{1, cos(\theta-\beta)$, and $cos^2(\theta-\beta)\}$.

Expanding

$$[\alpha + cos(\theta - \beta)]^2 = \alpha^2 + 2\alpha\cos(\theta - \beta) + cos^2(\theta - \beta) \tag{A.13}$$

The terms α^2 and $2\alpha\cos(\theta - \beta)$ can be accounted for using the W, X, and Y channels from the previous section. The term $cos^2(\theta - \alpha)$ represents a second-order gradient pointing in direction β. In order to steer the second-order response in an arbitrary direction, we must find a way to write it as a linear combination of a finite number of second-order building blocks. Using the double-angle formula

$$cos^2(\theta - \beta) = \frac{1}{2} + \frac{1}{2}cos[2(\theta - \beta)] \tag{A.14}$$

$$cos[2(\theta - \beta)] = cos(2\beta)cos(2\theta) + sin(2\beta)sin(2\theta) \tag{A.15}$$

We can rewrite the terms $cos(2\theta)$ and $sin(2\theta)$ again using the double-angle formulas

$$cos(2\theta) = 2cos^2(\theta) - 1 \tag{A.16}$$

which is a combination of a second-order gradient on the x-axis and an omni. For the $sin(2\theta)$ term we have

$$sin(2\theta) = cos(2\theta - \pi/2) \tag{A.17}$$

$$= cos[2(\theta - \pi/4)] \tag{A.18}$$

$$= 2cos^2(\theta - \pi/4) - 1 \tag{A.19}$$

which is a combination of a second-order gradient response 45^o between the x-axis and y-axis. Therefore

$$cos^2(\theta - \beta) = \frac{1}{2} + \frac{1}{2}\left[cos(2\beta)(2cos^2(\theta) - 1) + sin(2\beta)(2cos^2(\theta - \pi/4) - 1)\right] \tag{A.20}$$

which can be expanded to

$$cos^2(\theta - \beta) = \frac{1}{2}\left[1 - cos(2\beta) - sin(2\beta)\right] + cos(2\beta)cos^2(\theta) + sin(2\beta)cos^2(\theta - \pi/4) \tag{A.21}$$

Before plugging into equation (A.13), we can simplify by defining channels

$$W = 1 \tag{A.22}$$
$$X = cos(\theta) \tag{A.23}$$
$$Y = sin(\theta) \tag{A.24}$$
$$Q = cos^2(\theta) \tag{A.25}$$
$$R = cos^2(\theta - \pi/4) \tag{A.26}$$

Therefore

$$[\alpha + cos(\theta - \beta)]^2 = \tag{A.27}$$
$$\frac{2\alpha^2 + 1 - cos(2\beta) - sin(2\beta)}{2} W + 2\alpha\cos(\beta)X + 2\alpha\sin(\beta)Y + cos(2\beta)Q + sin(2\beta)R \tag{A.28}$$

which is a linear combination of:

- An omnidirectional microphone.
- Two gradient microphone arrays, one along the x-axis, one along the y-axis.
- Two second-order gradient microphone arrays, one along the x-axis, and one 45^o between the x-axis and y-axis.

Note: these are not the same equations as outlined in [32], which are based on spherical harmonics and which are *not* symmetric in the xy-plane. Instead we have derived the equations for second-order gradients in the plane, which are preferable when restricted to two dimensions.

If the channels above are implemented directly, it would require 7 microphones. If, however, we form the Y channel as a linear combination of $cos(\theta)$ and $cos(\theta - \pi/4)$, we can obtain the second order array from just five microphones.

Using the standard trigonometric identities

$$cos(a - b) = cos(a)cos(b) + sin(a)sin(b) \tag{A.29}$$

we can write

$$cos(\theta - \pi/4) = cos(\pi/4)cos(\theta) + \sin(\pi/4)sin(\theta) \tag{A.30}$$

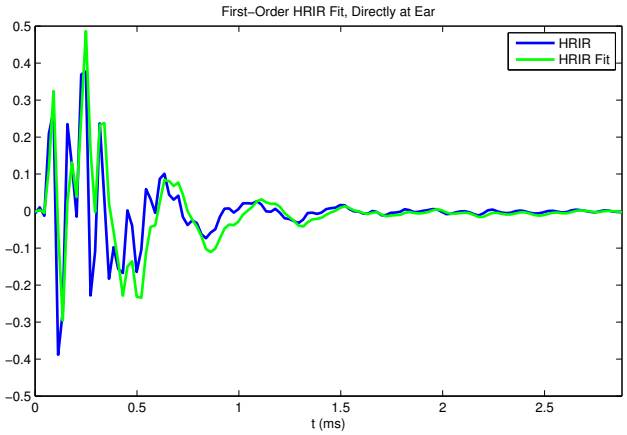
Rearranging

$$sin(\theta) = cos(\theta) + \sqrt{2}\cos(\theta - \pi/4) \tag{A.31}$$

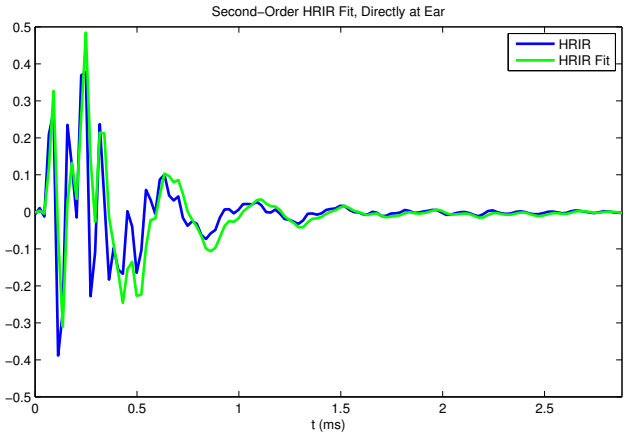
and we can therefore eliminate the two microphones on the y-axis.

A.5 HRTF Fits Using Higher-Order Differential Arrays

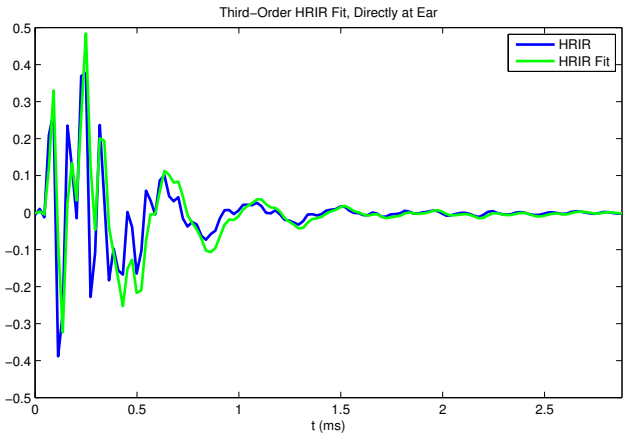
Now that we know how to implement higher-order gradient arrays, a reasonable question to ask is, "how much benefit can we expect to obtain from using them in a spatial audio system?" While the polar plots above demonstrate their higher directivity, we can get a better sense of their influence by looking at how they affect the reconstruction error in the HRIR fits. Figures A.7 and A.8 look at the cases when sound is incident directly at the ear and directly opposite the ear respectively. It is evident that increasing the order of the array does not significantly enhance the ipsilateral ear, but does have an effect on the contralateral ear. Although this effect is audible, the amount of audibility is subject to diminishing returns. From my personal experience, the third-order array does not dramatically outperform the first-order array. Therefore, with the expense of increasing orders in mind, the decision to use an increased order must be made carefully.



(a) Ipsilateral fit for first-order gradient array.

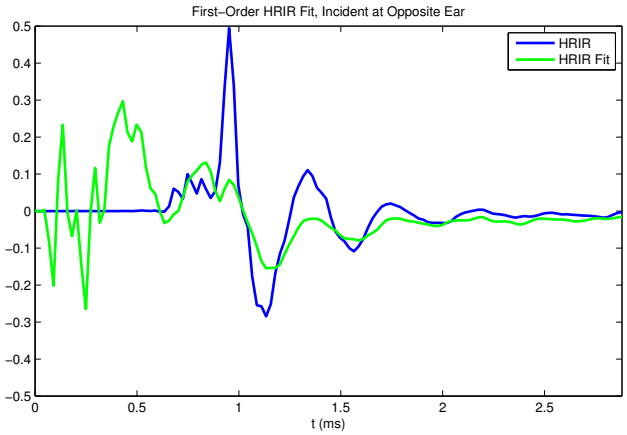


(b) Ipsilateral fit for second-order gradient array.

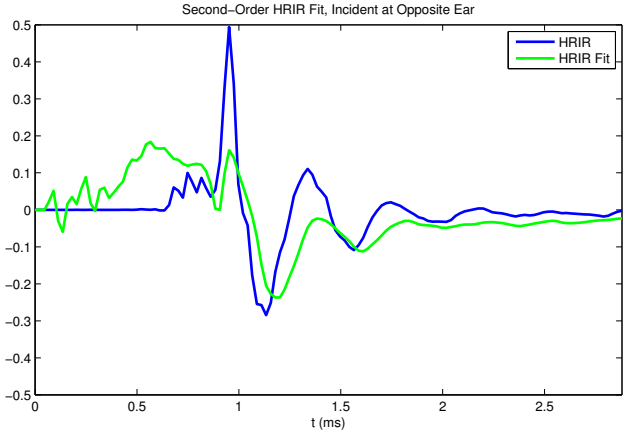


(c) Ipsilateral fit for third-order gradient array.

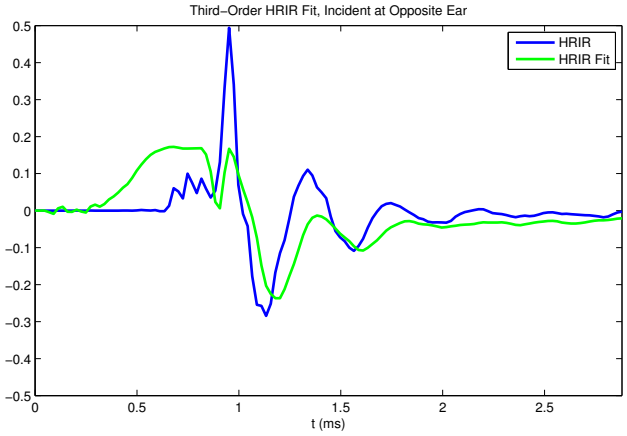
Figure A.7: Ipsilateral fits for sound incident at the opposite ear (ipsilateral fit). Little benefit is found since the magnitude of the incident HRIR is larger than other angles of incidence.



(a) Contralateral fit for first-order gradient array.



(b) Contralateral fit for second-order gradient array.



(c) Contralateral fit for third-order gradient array.

Figure A.8: HRIR fits for sound incident at the opposite ear (contralateral fit). The higher-order array significantly reduces preringing which leads to more accurate ITD cues.

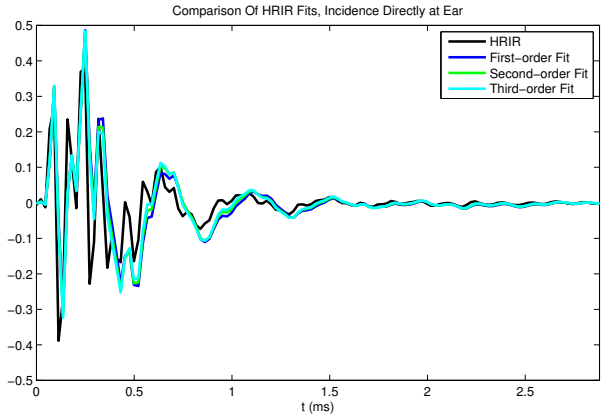


Figure A.9: Comparison of first-, second-, and third-order fits for the ipsilateral ear.

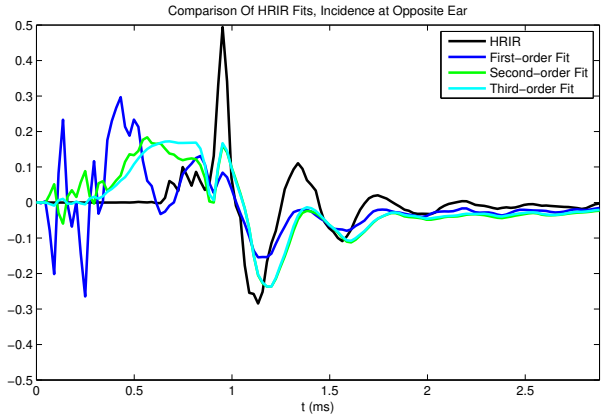


Figure A.10: Comparison of first-, second-, and third-order fits for the contralateral ear.

APPENDIX B

LEAST-SQUARES AND AMBISONIC EQUIVALENCE WHEN USING A SOUNDFIELD MICROPHONE

As was discussed in the chapter on Ambisonics, the output of a soundfield microphone can be placed in B-format. If we examine this for two-dimensional reconstruction, the channels have responses:

$$W = \frac{\sqrt{2}}{2} \tag{B.1}$$

$$X = cos(\theta) \tag{B.2}$$

$$Y = sin(\theta) \tag{B.3}$$

Placed in a matrix we have three columns, $cos(\theta))$, $sin(\theta))$, $\frac{\sqrt{2}}{2}$, where θ ranges from 0 to 2π in discrete steps according to the virtual speaker sampling.

In the least-squares solution,

$$\mathbf{h}_{LS} = (\mathbf{D}^H\mathbf{D})^{-1}\mathbf{D}^H\mathbf{b} \tag{B.4}$$

the matrix $\mathbf{D}^H\mathbf{D}$ is diagonal because the three columns are orthogonal to one another. Therefore the inverse matrix $(\mathbf{D}^H\mathbf{D})^{-1}$ is also diagonal. The result is that $(\mathbf{D}^H\mathbf{D})^{-1}$ merely acts as a scaling matrix to the term $\mathbf{D}^H\mathbf{b}$. For the case when the gains on channels W, X, and Y above are $\frac{\sqrt{2}}{2}$, 1, and 1 respectively, the $(\mathbf{D}^H\mathbf{D})^{-1}$ will be the identity matrix.

REFERENCES

[1] Personal communication, Ian Lewis, Senior Acoustics Engineer Blackberry, 2012.

[2] H. Moller, "Fundamentals of binaural technology," *Applied Acoustics*, vol. 36, 1992.

[3] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[4] J. Daniel, R. Nicol, and S. Moreau, "Further investigations of High Order Ambisonics and wavefield synthesis for holophonic sound imaging," in *AES 114TH Convention, Amsterdam*, 2003.

[5] S. Sakamoto, J. Kodama, S. Hongo, T. Okamoto, Y. Iwaya, and Y. Suzuki, "A 3D sound-space recording system using spherical microphone array with 252ch microphones," in *Proceedings of 20th International Congress on Acoustics*, 2010.

[6] E. Choueiri, "Optimal crosstalk cancellation for binaural audio with two loudspeakers," *Self Published*.

[7] J. Blauert, *Spatial Hearing*. MIT Press, 1983.

[8] L. Rayleigh, "Our perception of the direction of a source of sound," *Proceedings of the Musical Association*, pp. 75–84, 1876.

[9] L. Rayleigh, "On our perception of sound direction," *Philosophical Magazine Magazine*, vol. 12, pp. 214–232, 1907.

[10] A. Mills, "On the minimum audible angle," *Journal of the Acoustical Society of America*, vol. 30, pp. 237–246, 1958.

[11] A. Mills, "Lateralization of high-frequency tones," *Journal of the Acoustical Society of America*, vol. 32, pp. 132–134, 1960.

[12] H. Wallach, "On sound localization," *Journal of the Acoustical Society of America*, vol. 10, pp. 270–274, 1939.

[13] H. Wallach, "The role of head movements and the vestibular and visual cues in sound localization," *Journal of Experimental Psychology*, 1940.

[14] H. Wallach, E. Newman, and M. Rosenzweig, "The precedence effect in sound localization," *The American Journal of Psychology*, vol. 62, pp. 315–336, 1949.

[15] F. Wightman and D. Kistler, "Headphone simulation of free-field listening. I: Stimulus synthesis," *Journal of the Acoustical Society of America*, vol. 85, pp. 858–867, 1989.

[16] F. Wightman and D. Kistler, "Headphone simulation of free-field listening. II: Psychophysical validation," *Journal of the Acoustical Society of America*, vol. 85, pp. 868–878, 1989.

[17] B. Gardner and K. Martin, "HRTF measurements of a KEMAR dummy," 1994.

[18] W. Hartmann and A. Wittenberg, "On the externalization of sound images," *Journal of the Acoustical Society of America*, vol. 99, pp. 3678–3688, 1996.

[19] D. Begault, E. Wenzel, and M. Anderson, "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *J Audio Eng. Soc.*, vol. 49, pp. 904–916, 2001.

[20] K. Brandenburg, S. Brix, and T. Sporer, "Wave field synthesis," in *IEEE 3DTV Conference*, 2009.

[21] P. Fellgett, "Ambisonics. Part one: General system description," *Studio Sound*, vol. 17, 1975.

[22] M. Gerzon, "Ambisonics. Part two: Studio techniques," *Studio Sound*, vol. 17, 1975.

[23] K. Farrar, "Soundfield microphone," *Wireless World*, 1979.

[24] D. Zotkin, R. Duraiswami, and L. Davis, "Rendering localized spatial audio in a virtual auditory space," *IEEE Transactions on Multimedia*, vol. 6, pp. 553–564, 2004.

[25] S. Zhao, R. Rogowski, R. Johnson, and D. L. Jones, "3D binaural audio capture and reproduction using a miniature microphone array," in *Int. Conference on Digital Audio Effects*, 2012.

[26] M. Brandstein and D. Ward, Eds., *Microphone Arrays*. Springer, 2001.

[27] B. V. Veen and K. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE ASSP Magazine*, pp. 4–24, 1988.

[28] B. Wiggins, "Has Ambisonics come of age?" *Proceedings of the Institute of Acoustics*, vol. 30, 2008.

[29] H. Olson, "Gradient microphones," *Journal of the Acoustical Society of America*, vol. 17, pp. 192–198, 1945.

[30] J. Eargle, *The Microphone Book*. Focal Press, 2004.

[31] G. Elko, *Acoustic Signal Processing For Telecommunication*, S. Gay and J. Benesty, Eds. Kluwer, 2000.

[32] J. Daniel, "Acoustic field representation, application to the transmission and the reproduction of complex sound environments in a multimedia context," Ph.D. dissertation, l'Universite Paris, 2000.