Binaural Hearing Effects of Mapping Microphone Array’s Responses to a Listener’s Head-Related Transfer Functions

James A. Hughet

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science
In
Mechanical Engineering

Michael J. Roan
Ricardo A. Burdisso
Cory M. Papenfuss

September 29, 2011
Blacksburg, VA

KEYWORDS: Binaural Hearing, HRTF Matching, Array Processing, Listener Test, Head-Movement

COPYRIGHT © SEPTEMBER 2011, James A. Hughet
Binaural Hearing Effects of Mapping Microphone Array’s Responses to a Listener’s Head-Related Transfer Functions

James Hughet

Abstract

This thesis focuses on the mapping of the microphone array’s response to match the characteristics of a human subject’s Head-Related Transfer Function (HRTF). The mapping of the response is first explored with a ‘monaural HRTF matching’ that filters the response independent of the arrival angles. For arbitrary array geometry with the listener external to the acoustic, the monaural HRTF matching did not provide listeners with enough spatial information to precisely localize sound sources. To correct this, a preprocessor control algorithm was added to the HRTF matching, a ‘binaural HRTF matching’ process. The binaural HRTF matching increased the listeners’ performance in perceiving the location of a sound source. With the addition of simulated head movement, the listeners’ perception increased by 20%. An issue with this approach is the use of HRTFs other than the listeners’ measured HRTF, creating a psychoacoustic based error in localization, i.e., front/back confusion.
Acknowledgment

I would like to express my gratitude to each and every one who contributed to the completion of this thesis, including: Michael Roan for his advice, patience, and guidance throughout my course of work in the Vibration and Acoustics Laboratory; Phil Chin for being an attentive colleague and friend; the subjects who participated in the numerous listener tests; and my family for their support.
Table of Contents

Title Page ................................................................. i
Abstract ........................................................................ ii
Acknowledgement .......................................................... iii
Table of Contents ........................................................... iv
List of Figures ............................................................... vi
Chapter 1. Introduction ...................................................... 1
  1.1 Thesis Motivation ..................................................... 1
  1.2 Problem Investigation ............................................... 1
  1.2.1 Thesis Investigation .............................................. 3
  1.3 Thesis Objectives and Overview .................................. 3
Chapter 2. Background ...................................................... 5
  2.1 Microphone Array Fundamentals ................................. 5
  2.2 Human Hearing Attributes for Localization of Sound Sources ................................................. 8
  2.2.1 Localization of Sound Sources ................................ 8
  2.2.2 Limitations of the Auditory System ....................... 11
Chapter 3. HRTF Filter Design and Implementation .................. 15
  3.1 Introduction into HRTFs ............................................. 16
  3.1.1 Why HRTFs are used ........................................... 17
  3.2 Introduction into HRTF Filter Design ......................... 19
  3.2.1 Generalization of the HRTFs ................................. 20
  3.2.2 Interpolation of HRTF Datasets ............................... 22
  3.2.3 Head Movement Implementation Consideration ........... 25
  3.2.4 Inherent Limitations of HRTFs .............................. 27
Chapter 4. HRTF Matching .................................................. 29
  4.1 Classic Approach of Beamforming ............................... 30
  4.2 HRTF Matching ....................................................... 32
  4.2.1 Limitations of Monaural HRTF Matching .................. 35
  4.2.2 Reformulation of HRTF Matching ............................ 39
  4.2.3 HRTF Match with Head Movement ......................... 41
Chapter 5. Listener Test of Binaural HRTF Matching ............... 43
  5.1 First Listener Test of the Binaural HRTF Matching with Fixed Head Position ......................... 45
  5.2 Second Listener Test of the Binaural HRTF Matching with Head Movement ..................... 51
  5.3 Third Listener Test: Validation of Front/Back Confusion ...... 54
  5.4 Additional comment on the Second Listener Test ............. 55
Chapter 6. Discussion of Process Limitations and Future Work ....... 57
  6.1 Limitation of preprocessing control algorithm ................. 57
  6.1.1 Source Localization (Find the Number and Location of Sources) ........................................... 58
  6.1.2 Source Separating ............................................... 61
  6.2 Conclusion of Limitation ........................................... 62
  6.3 Future Work ......................................................... 63
Chapter 7. Conclusion .......................................................... 64
Appendix A. Characteristics of Microphone Array ..................... 66
  Appendix A.1 Construction of the Headphone Array and Headrest Array ................................. 66
  Appendix A.2 Performance of the Headphone and Headrest Array .............................................. 67
  Appendix A.3 Gathering Array Data ................................... 69
Appendix B. Measuring the HRTF ...........................................................................................................72
Appendix B.2 Equalization of HRTFs....................................................................................................73
References
List of Figures:

Figure 2.1) An illustration of the head-mounted, Headphone Array outputting the microphone signal $X_t$ for a sound source with an arrival angle of $\theta$. .......................................................... 7

Figure 2.2) The impact of changing the ITD and ILD on a listener in locating the azimuth of a sound source. The original sound waves are shown in blue. .......................................................... 10

Figure 2.3) Cone of confusion (i.e., the psychoacoustic region where multiple sounds have the same ITD and ILD). .......................................................... 12

Figure 2.4) Cognitive reasoning used to find a sound source with head movement. .......................................................... 13

Figure 2.5) MAA as a function of frequency of successive tone and angle of source relative to an elevation of $0^\circ$. .......................................................... 14

[14]

Figure 3.1) The natural filtering of a sound source by the HRTF on its way to the inner ear (Left). The restoring of the HRTFs to the listener with a head-mounted system (Right). .......................................................... 15

Figure 3.2) HRTF of the KEMAR manikin [19] in the time domain and frequency domain for a sound source originating at 45° Left and 20° above the manikin with the ITD and ILD emphasized. .......................................................... 19

Figure 3.3) The phase components of the original HRTF and the minimum phase HRTF at relatively low frequencies. .......................................................... 21

Figure 3.4) The phase components of the original HRTF and the minimum phase HRTF at relatively high frequencies. .......................................................... 22

Figure 3.5) The MSE between the interpolated HRIR and the measured HRIR for the KEMAR manikin [19]. .......................................................... 24

Figure 3.6) Overlap-add method of a repeating window utilizing a Hann window function with 50% overlap. .......................................................... 27

Figure 4.1) Schematic of a head-mounted system giving the microphone array the characteristics of the HRTFs. This real-time process is shown as the green path. The blue path represents the storing of the information before relaying it to the listener. .......................................................... 30

Figure 4.2) Comparison of ITD for microphone 4, 6 of the head-mounted Headphone Array to the KEMAR manikin ITD at an elevation of $0^\circ$. .......................................................... 36

Figure 4.3) Array manifold matrix of KEMAR Manikin left ear with a normal size pinna. .......................................................... 37

Figure 4.4) Array manifold matrix of the head-mounted Headphone Array. .......................................................... 37

Figure 4.5) Mapped array manifold matrix of the head-mounted Headphone Array to the KEMAR manikin left ear with a normal size pinna. .......................................................... 38

Figure 4.6) Schematic of the binaural HRTF matching that map the microphone array’s response to match the characteristics contained in the HRTFs. .......................................................... 41

Figure 5.1) Schematic of the binaural HRTF matching that maps the microphone array’s response to match the characteristics contained in the HRTFs. .......................................................... 44

Figure 5.2) Excerpt from the listener test sheet; listener indicated the perceived azimuth location by circling one of the angles. .......................................................... 46

Figure 5.3) Listeners’ responses of the first listener test for the Headphone Array. The colors represent a particular sound; red is music, blue is voice, and green is noise. Each dot corresponds to a different test subject’s response. .......................................................... 47

Figure 5.4) The percentage of front/back confusion in the first listener test. .......................................................... 49

Figure 5.5) Spatial response of the KEMAR with a normal size pinna (left) and the KEMAR with a large size pinna (right) [19]. .......................................................... 50

Figure 5.6) Listeners’ responses of the second listener test shown as degrees off from the actual response. Although each dot corresponds to a different test subject’s response, there are many points that overlay each other...... 53

Figure 6.1) PSD output of (6.5) for the head-mounted Headphone Array of a single speaker at an arrival angle of 50°. .......................................................... 60

Figure 6.2) PSD output of (6.5) for the head-mounted Headphone Array of a single speaker at an arrival angle of 150°. .......................................................... 61

Figure A.1) Left to Right, Headphone Array and Sensor Construction, where the microphones receive their power from the DAQ. .......................................................... 66

Figure A.2) Left to Right, Headrest Array (Front, Right, Back, Left) and Sensor Construction. .......................................................... 67

Figure A.3) Beampattern of Headphone Array steered to an azimuth of $\theta_s = 90^\circ$ with a DSB, for the frequency range of human hearing, 0 Hz to 20 kHz. .......................................................... 69

Figure A.4) Array test setup for Headphone Array. .......................................................... 70
THIS PAGE INTENTIONALLY LEFT BLANK
Chapter 1. Introduction

1.1 Thesis Motivation

Humans have the ability to focus on a single speaker among a sea of conversations and background noise. The intelligibility of speech has been found to be significantly improved when the target and competing sentences are spatially separated [1]. Though, when listening to a response of a microphone array, the perceptual awareness of the listener is lost. This degradation of the listener’s perception to localize a sound source is viewed as a loss of the human subject’s spatial cues. The returning of the lost spatial cues is performed with a mapping process (HRTF matching) that maps the microphone array’s response to match the spatial characteristics of a human subject, captured in the Head-Related Transfer Function (HRTF).

This thesis focuses on the mapping of the microphone array’s response to match the characteristics of a human subject. The mapping of the response will contain enough spatial information for the listener to perceive the location of a sound source. The validation of this mapping process is then determined by the listener’s ability to locate a sound source via a series of listener tests.

1.2 Problem Investigation

The need to have a microphone array with the characteristics of a human subject is driven by the listener’s need to perceive the location of a sound source. These types of systems are found often in systems with the listener immersed within the environment and rely on the acoustic information being relayed to the listener. The information that is relayed to the listener could cause a loss in the listener’s ability to locate a sound source by not contain the correct HRTF, or could cause the listener to incorrectly identify the sound’s direction.
A common microphone array system that gives the listener spatial perception is a hearing aid. The spatial perception given to the listener by the hearing aid design is more accurately stated as spatial preserving [2]. The placement of the hearing aid microphones naturally captures the binaural spatial cues, the Interaural Time Difference (ITD) and the Interaural Level Difference (ILD), all of which are required to locate a sound source. The design of the hearing aid system enhances the level of speech by suppressing background noise levels and preserving binaural spatial cues. Whether the hearing aid systems uses one or multiple microphones with fixed beamforming [3], adaptive beamforming [3], or Multichannel Wiener Filtering [4], there is no outlet to modify the spatially-preserved characteristics. To incorporate the spatial characteristic of the HRTF, a filter is added to the microphone signals.

There have been several attempts to give the microphone array response the characteristics of the human subject’s HRTF. These attempts include:

— Bronkhorst and Verhave, 2005: This approach of giving the listener spatial perception was achieved by mapping the transfer functions of the microphone array to the HRTF [5].

— Gillett, 2009: This approach of giving the listener spatial perception advanced the ‘Bronkhorst and Verhave’ method by incorporating the array geometry. Gillett’s method incorporates the array characteristics by minimizing the mean square error between the desired spatial pattern of the human subject and the spatial pattern of the microphone array. The benefit of this method is that the microphone array’s response is matched to the human subject for all scenarios [6].

Gillett’s method would be an ideal solution since it maps the microphone array’s response independently of the array’s data. Gillett, though, only investigated head-mount arrays, leaving a gap of whether his method can be used on other array geometry. This thesis uses Gillett’s
method as a starting point to investigate the mapping of a microphone array’s response to match the characteristics of a human subject.

### 1.2.1 Thesis Investigation

The mapping of the microphone array’s response to match the characteristics of a human subject is first explored with the limitation of the fomented head-mounted system with a monaural HRTF matching, (Gillett’s method). The reader should take note that our investigation differs from this method with the audience being external to the acoustic environment, where the information is stored before it is played to the listener. Secondly, this thesis focuses on using the HRTF matching for any microphone array geometry, not just simple head mound arrays. During the listener tests, it was noted that the monaural HRTF matching did not provide listeners with enough spatial information to precisely localize a sound source. To provide listeners with the desired perceptual awareness, the data-independencies were removed for the HRTF matching. To decrease the high levels of listener confusion regarding whether the direction of the sound was in front or behind the listener, the addition of head-movement was also investigated.

### 1.3 Thesis Objectives and Overview

This thesis focuses on the mapping of the microphone array’s response to match the characteristics of a human subject’s HRTF. The mapping of the response will contain enough spatial information for the listener to perceive the location of a sound source. The validation of this ‘HRTF matching’ process is determined by the listeners’ ability to locate a sound source via a series of listener tests. This thesis will explore HRTF matching for any microphone array geometry with the listener external to the acoustic environment.
Chapter 2 will first present a generalized background into the two different components of the HRTF matching: 1) The model of a microphone array and 2) The characteristics the human auditory system uses to spatially determine the location of a given sound source.

Chapter 3 focuses on the mathematical model of the HRTF filter design, along with the challenges that are required for the implementation of the HRTF filters with head-movement.

Chapter 4 explores the HRTF matching required to map the microphone array’s response to match the characteristics contained in the HRTFs.

Chapter 5 evaluates the HRTF matching process listener tests by reviewing the listeners’ ability to locate a sound source.

Chapter 6 presents a brief introduction into a realistic case. This is by no means a detailed exploration; the Chapter 6 investigation is intended to highlight the limitations of our investigation and where future investigation is required.

Chapter 7 provides an overview of this thesis as a whole.
Chapter 2. Background

The goal of this work is to produce a virtual 3D sound representation from measurements taken by a microphone array. Before discussing the mapping of microphone array’s response to match the characteristics of a human subject, this Chapter presents an introduction into the two different components of this transformation: 1) The model of a microphone array and 2) The characteristics the human auditory system uses to spatially determine the location of a given sound source.

2.1 Microphone Array Fundamentals

A microphone array captures $M$ filtered versions of the sound source, $s(t)$. The filtering of the sound source by the microphone array can be modeled a number of ways. The free-field, far-field, or near-field models are the most commonly used models [7]. In the free-field model, the captured signals of the sound source behave as if in free space (open air with nothing to interfere with its propagation); the sound level falls off as a function of distance from the source [8]. The free-field model is problematic for a realistic environment because of the over-simplification of broadband signals (frequency range of interest is relatively large). For that reason, the analysis in this thesis uses the far-field approach that assumes planer waves [9].

For a single broadband sound source, $s(t)$, the filtering of $s(t)$ is characterized by an Impulse Response (IR), denoted by $h_m$ of length $L$ [10]. The captured signal at the $m^{th}$ microphone, $x_m(t)$, at time $t$ is then:

$$x_m(t) = h_m(L, \theta) * s(t) \tag{2.1}$$

where $*$ represents a convolution. Stacking all $M$ microphone signals of (2.1) together yields the microphone array’s response:
\[ \mathbf{X(\omega)} = \mathbf{a(\omega, \theta)s(\omega)}, \quad \omega = [\omega_0, ..., \omega_{\text{FFT}}] \tag{2.2} \]

where the sound source \( \mathbf{s(\omega)} \) is the Fourier transform of \( s(t) \) and \( \mathbf{a(\omega, \theta)} \) is the array steering vector:

\[ \mathbf{a(\omega, \theta) = [H_1(\omega, \theta), ..., H_M(\omega, \theta)]^T} \tag{2.3} \]

The array steering vector contains all \( M \) transfer functions \( \mathbf{H}_m \), which are the Fourier transform of the microphones’ IR, \( h_m \). With a frequency dependency, \( \omega \), that goes from 0 Hz to half of the sampling frequency rate. The microphone array’s response of (2.2) is expressed in the frequency domain since all of the processing is conducted in the frequency domain.

The filtering of the sound source in (2.2) is completed by the array steering vectors that contain all \( M \) transfer functions. With the microphone array data being captured in the free-field environment, the filtering of a sound source by the transfer function can be assumed to contain only the modifications generated by the physical diffractive nature of the microphone array’s geometry. These spatial characteristics are dependent on the sound source arrival angle \( \theta \), which is measured counterclockwise from the array axis at the origin’s reference point. This reference point is taken to be the center of the microphone array. The reference point for one of the arrays used in this thesis, the head-mounted, Headphone Array is shown in Figure 2.1. The array in Figure 2.1 is shown outputting the microphone array’s response \( \mathbf{X(t)} \) for a sound source with an arrival angle of \( \theta \).
Figure 2.1) An illustration of the head-mounted, Headphone Array outputting the microphone signal $X(t)$ for a sound source with an arrival angle of $\theta$.

The transfer functions of the microphone array and the transfer functions or Head Related Transfer Function (HRTF) of the human subjects are the spatial characteristics used in Chapter 4 to generate the set of filters that map the microphone array characteristics to those of a human subject. The filtering combines the microphone signals of the microphone array with complex gains that either enhance or reject the received signals according to a specific attribute via a set of filters, $\mathbf{w}$, to generate the array output, $Y(\omega)$ [9].

$$Y(\omega) = \mathbf{w}^H X(\omega)$$  \hspace{1cm} (2.4)
The superscript $(\cdot)^H$ denotes the conjugate transpose. The derivation of $\mathbf{w}$ that maps the microphone array’s response to match the characteristics of the human subject is discussed in Chapter 4.

2.2 Human Hearing Attributes for Localization of Sound Sources

Humans have a remarkable ability to precisely localize sound sources with only two sensors (ears). This ability to localize sound sources is due to the diffractive nature of the head, torso, and pinna. The auditory system localizes sound sources with two sub-systems: 1) the biological/mechanical apparatus, or how the ears translate sound into biological signals and 2) the neurological processing of these biological signals, or how the brain interprets sound [11]. Each of these systems plays a role in locating a sound source. This thesis does not dwell on how the human auditory system works; rather it focuses on what the auditory system requires to locate a sound source.

2.2.1 Localization of Sound Sources

Humans have a remarkable ability to localize sound sources accurately and rapidly from all directions. Locating a sound source can be placed into three subcategories: locating the sound source’s azimuth, elevation, and range. Localization of the sound source’s azimuth is the ability to locate the sound source in the horizontal plane. Two (binaural) spatial cues are used to determine the azimuth of a sound source. These spatial cues are the Interaural Time Differences (ITD) and the Interaural Level Differences (ILD) [11]. The ITD is the difference in time the incoming signal takes to reach the ears. The ILD is the difference in the level of the signal between the ears. The ILD, caused by the head acting as a barrier to high frequency sound waves, is termed head-shadowing.
The auditory system uses both the ITD and ILD to determine the azimuth of the sound source. How the ITD and ILD are used to find the azimuth location of a sound source is referred to as the Duplex theory, named by Lord Rayleigh [12]. Lord Rayleigh’s Duplex theory shows that the ITD are more important at frequencies below 800 Hz. At higher frequencies, multiple cycles can occur between the ears, making the ITD an inaccurate measurement. For frequencies above 1600 Hz, the Duplex theory states that the ILD is a more effective spatial cue. At lower frequencies, the wave length compared to the length of the head is long, so there is no head shadowing [11].

While the auditory system uses both the ITD and ILD to determine the azimuth of the sound source, it places a greater emphasis on one or the other depending on the frequency [11]. Additionally, this means that if the incorrect ITD or ILD is fed to the listener, it will have a direct impact on the listener’s ability to locate a sound source. An example of this is shown in Figure 2.2. The left example in Figure 2.2 shows that if the incoming sound in the right ear is delayed, the azimuth of the sound source will artificially shift to the left. Whereas, the right example in Figure 2.2 shows that if the incoming sound to the left ear is attenuated, the azimuth of the sound source will artificially shift back to the front of the listener.
Figure 2.2) The impact of changing the ITD and ILD on a listener in locating the azimuth of a sound source. The original sound waves are shown in blue.

Localizaion of the sound source’s elevation is the ability to locate the sound source in the vertical plane (i.e., whether or not a sound source originated above or below the ears). The auditory system is far less accurate in locating the elevation than in locating the azimuth. In locating the sound source’s elevation, the auditory system relies on the spectral diffraction caused by the listener’s body. The reflection caused by the torso and head plays a role in locating the elevation, though the main factor is the reflection of the pinna. The reflection of sound off a person’s torso in determining the elevations of a sound source are most notable when the sound source originates at a low elevation. At an elevation well below the listener’s head, the spatial diffraction contained within the HRTF is sensitive to the posture of the listener (i.e., the listener standing upright verses being slumped over) [13].

Localization of the sound source’s distance is the ability to determine the sound source range. In determining the range of a sound source, the auditory system primarily uses the level of the source sound, but also uses other factors such as the ILD for close sound sources and reverberation off nearby objects. For determining the distance of the sound source, the auditory
system alone is not a strong device. The perception of the sound level of a person shouting in the distance and a person whispering nearby can be equivalent. In this case, head movement will show larger azimuth changes for nearby objects than objects in the distance. However, the 1972 Plenge experiment revealed that listeners used ‘familiarity’ over the auditory cues. In the Plenge experiment, listeners always assumed that a sound source of a person whispering was nearby regardless of how faraway the sound source actually was [11].

The auditory system can also perceive spatial attributes such as reverberation. These spatial attributes for the perception of space are determined by the Interaural Coherence (IC). The IC is a metric of the auditory system that indicates the similarity or lack of similarity between the received signals of each ear [11].

2.2.2 Limitations of the Auditory System

Locating the azimuth of a sound source with only the ITD and ILD causes psychoacoustic based errors in localization, known as the ‘cone of confusion’. The cone of confusion is a psychoacoustic region where the auditory system is confused as to whether a sound is above/below and/or in front/behind the listener [11]. The cone of confusion is caused by different points in space having the same ITD and ILD. Ignoring the pinna, the head can be approximated as a spherical shape, then the points that have the same ITD and ILD form a cone (hyperboloid) centered on an Interaural axis, shown in Figure 2.3. Since the scope of this thesis is on locating the azimuth of a sound source, the cone of confusion is reduced to a front to back confusion. That is, the listener is confused as to whether a sound is in front of or behind the listener.
Currently, there is no solution that eliminates the cone of confusion. Under normal listening conditions, this confusion is naturally solved by head movements and spatial modifications of the pinna [11]. When creating the virtual 3D representation discussed in the next Chapters, the human characteristics are not those of the specific listeners. The characteristics generated by the body’s head, torso, and pinna are very person specific. There is a level of inaccuracy created by using another person’s specific spatial information. The creating of the virtual 3D representation often does not take the listener’s current head position into account; the mapping is performed with the listener removed. This artificial constraint means that the natural small instinctual head movements that allow for cognitive reasoning to pinpoint the source location, illustrated in Figure 2.4, are absent.
Figure 2.4) Cognitive reasoning used to find a sound source with head movement.

Having two ears also limits the human perception to determine the location of a sound source. The minimum audible angle (MAA) is the threshold of the listener’s ability to spatially locate a sound source. Listeners can detect changes in angles of 1° when the source is straight ahead. As the sound source moves to the side of the head, the MAA accuracy drops off. The MAA illustrated in Figure 2.5 are the results of Mills’ study. This illustration shows the smallest angular separation that can be detected between the sources of two successive tone pulses (MAA) [14]. Additionally, Figure 2.5 shows the ability to localize a sound source azimuth is frequency and arrival angle dependent.
Figure 2.5) MAA as a function of frequency of successive tone and angle of source relative to an elevation of 0° [14].
Chapter 3. HRTF Filter Design and Implementation

The Head-Related Transfer Function (HRTF) represents the natural filtering of a sound source as it interacts with our bodies on its way to the inner ear. This normal listening process is shown in the left illustration of Figure 3.1. The right illustration of Figure 3.1 shows the restoring of the HRTFs to the listener with a head-mounted system. These systems have the listener immersed within the environment and relay on the acoustic information being relayed to them. The relayed information of the microphone array’s response does not contain the correct HRTFs required to spatial perceive the location of a sound source. In preparation for the restoring of the missing HRTFs with the mapping process of Chapter 4, this Chapter explores how the HRTFs are modeled and interpolated.

Figure 3.1) The natural filtering of a sound source by the HRTF on its way to the inner ear (Left). The restoring of the HRTFs to the listener with a head-mounted system (Right).
3.1 Introduction into HRTFs

The HRTFs describe how a given sound wave is transformed by the diffraction and reflection properties of the head, torso, and pinna (outer ear). This time invariant process is illustrated in Figure 3.1 and defined by:

\[
\begin{align*}
X_L(\omega, \theta) &= \text{HRTF}_L(\omega, \theta)s(\omega) \\
X_R(\omega, \theta) &= \text{HRTF}_R(\omega, \theta)s(\omega)
\end{align*}
\] (3.1)

for an incident arrival angle of \( \theta \). \( X_L \) and \( X_R \) are the respective left and right ear responses. In this thesis, the HRTF will be discussed in regards of containing both the left and right HRTF:

\[
\text{HRTF}(\omega, \theta) = \begin{bmatrix} \text{HRTF}_L(\omega, \theta) \\ \text{HRTF}_R(\omega, \theta) \end{bmatrix}, \quad \omega = [\omega_0, ..., \omega_{\text{LFFT}}] \] (3.2)

The HRTFs directional dependences are set relative to the spherical head-related coordinate system, where the origin is located halfway between the entrances of the two ear canals. The elevation angle has been omitted from the HRTFs because the scope of this work only pertains to locating of a sound source azimuth direction. The elevation throughout this thesis will pertain to an elevation of \( \phi = 0^\circ \) unless otherwise stated.

Obtaining the HRTF can be determined using either a direct or an indirect approach. The direct approach is simply the actual measurement of the subject’s HRTF. The indirect approach uses some form of 3D imaging processing, which processes the image with finite element analysis of the head [15].

In (3.1), direct measurements of the HRTFs are obtained by noting the ear responses while playing a known broadband noise sound at specific points of azimuth and elevation. The measured input sound \( s(\omega, \theta) \) used in (3.1) varies depending on the approach. The most common approach is a free-field transfer function approach [16], [17]. With this approach, the HRTFs are obtained by relating a subject’s previous ear responses to the measured signal using the same
source to a point corresponding to the center of the head \( s_o(\omega) \) while the subject is not present \[18\]:

\[
\text{HRTF}_L(\omega, \theta) = \frac{X_L(\omega, \theta)}{s_o(\omega)}, \quad \text{HRTF}_R(\omega, \theta) = \frac{X_L(\omega, \theta)}{s_o(\omega)}
\] (3.3)

This method is used to obtain the KEMAR HRTFs of \[19\], the HRTFs of the ‘Listen Group’ subjects \[20\] and those of Subject 3 in the listener test discussed in Chapter 5. The free-field transfer function assumes a free-field approach. This approach requires the measurement to be conducted in an anechoic environment and at a considerable distance (within the far-field region) from the sound source. Detailed schematics showing the procedure for obtaining the HRTFs of the ‘Listen Group’ are outlined in \[20\] and discussed in Appendix B for Subject 3.

### 3.1.1 Why HRTFs are used

The mapping process of Chapter 4 uses the HRTFs to map the microphone array’s response to match the characteristics of a human subject. HRTFs are used since they contain the ITD, ILD and the spatial modifications of the head, torso, and pinna. These spatial cues are needed in estimating the azimuth, elevation and range of a sound source.

The ITD is the main spatial cue used in locating the azimuth of a sound source at frequencies below 800 Hz. The ITD is the measured difference in the time it takes for an incoming signal to reach one ear verses the other, seen in the time domain of the HRTF in Figure 3.2 as the delay between the start of the impulse response of the individual left and right HRTF. Theoretically, the ITD can be found by approximating the head as a sphere, for an infinitely distant source the ITD is found by the Woodworth simple formula \[21\]:

\[
\text{ITD} = \frac{a}{c} (\theta + \sin \theta)
\] (3.4)
where $\theta$ is the azimuth angle, $a$ is the radius of the sphere, and $c$ is the speed of sound.

Woodworth’s formula (3.4) assumes that the wave-length is much smaller than the diameter of the scatter. This approximation proved a remarkably close solution to the exact solution.

The spherical solution of (3.4) does not take the elevation characteristics of the ITD into account. The ITD change with elevation is attributed to the non-spherical shape of the head and to the ears being displaced behind and below the center of the head [21]. The ellipsoidal model in [21] shows that when the ears are offset, the ITD model yields the correct ITD for both azimuth and elevation. It should be noted that the Extended Woodworth equation provides an additional scaling factor of $\cos \phi$ for elevation.

The ILD is the main spatial cue in locating the azimuth of a sound source at frequencies above 1.6 kHz. The ILD is the attenuation of frequency caused by the diffraction of sound around a person’s head. The ILD can be seen in both the time domain and frequency domain of the HRTF in Figure 3.2 as the difference in level between the individual left and right HRTF.

There are two common methods that attempt to theoretically match the ILD to a human subject. These methods are the spherical head model [22] and the ‘snowman’ model [13], which adds a torso to the spherical head model. Both of these models can explain the major features of a pinna-less system.

The spatial modifications of the pinna are seen in the frequency domain of the HRTF in Figure 3.2 as peeks and dips. Since these spatial modifications are substantially more complex than those of the ILD and ITD, a good theoretical model has yet to be found. This, coupled with the fact that a pinna-less system yields poor performance in the spatial perception of the listener [23], the filters developed to give the microphone array the missing spatial cues will be discussed in terms of measured HRTFs instead of using a mathematical model.
3.2 Introduction into HRTF Filter Design

There is currently no closed form solution that compares to the measured HRTFs of the listener. For this reason, a large dataset of measured HRTFs dominates the addition or returning of the spatial cues to a given sound. The filters based on the measured HRTF are either the HRTFs themselves or a simplification of the HRTFs. The simplification of the subject’s HRTF can be placed into two categories: an interpolation of the HRTFs and a generalization of the HRTFs. Both the generalization process and the interpolation process are seen in regards to reducing the large computational load that the HRTFs place on a system’s memory. This thesis will not discuss this process in terms of lowering the complexity of the HRTFs, but instead in the context of adding a dynamic process that incorporates head movement.

Figure 3.2) HRTF of the KEMAR manikin [19] in the time domain and frequency domain for a sound source originating at 45° Left and 20° above the manikin with the ITD and ILD emphasized.
### 3.2.1 Generalization of the HRTFs

There are numerous ways to generalize the HRTFs. This thesis focuses on the minimum phase generalization of the HRTFs. The minimum phase generalization of the HRTF comes from the realization that the excess phase resulting from subtracting the original phase response from the minimum phase counterpart has been found to be approximately linear \(^{11}\). The HRTF can then be modeled with a minimum phase filter along with a pure delay. The realization of minimum phase HRTFs has shown that they present no perceptual consequences \(^{24}\).

The minimum phase generalization of the HRTF, starts with a decomposition of the HRTF into three separate parts: a minimum phase counterpart function \(H_{\text{min}}\), an all-pass function \(H_{\text{ap}}\), and a linear phase or pure delay \(H_{pd}\):

\[
HRTF(\omega) = H_{\text{min}}(\omega)H_{\text{ap}}(\omega)H_{pd}(\omega)
\]  

(3.5)

where \(H\) is the original HRTF \(^{25}\). The minimum phase counterpart function, \(H_{\text{min}}\) is determined by the log magnitude spectrum of \(H\) through the Hilbert transform.

\[
H_{\text{min}}(\omega) = |H(\omega)|e^{i\omega\psi_{\text{min}}}
\]  

(3.6)

where \(\psi_{\text{min}}\) is the phase of the a minimum phase counterpart obtained with the Hilbert transform of:

\[
\psi_{\text{min}}(\omega) = \text{Im}(\text{Hilbert}[−\log(|H(\omega)|)])
\]  

(3.7)

The Hilbert transform relates the real and imaginary parts of a complex signal. It was developed to solve a special case of Hilbert problems. The Hilbert transformation is used since the minimum phase approximate properties of the log magnitude spectrum and phase spectrum are Hilbert transform of each other \(^{26},^{27}\).

The all-pass function, defined as \(H_{\text{ap}}(\omega) = e^{i\omega\psi_{\text{exc}}},\) where \(\psi_{\text{exc}}\), is the excess phase component. The minimum phase generalizations assume that the excess phase component does
not have a bearing on the spatial awareness of the listener and hence can be neglected [28]. Then
the minimum phase HRTF is given as:

\[
\text{HRTF}(\omega, \theta) = H_{\text{min}}(\omega, \theta)H_{\text{pd}}(\omega, \theta)
\] (3.8)

where the pure delay function is \(H_{\text{pd}}(\omega) = e^{\pm j\omega \tau/2}\) and \(\tau\) is the ITD.

The minimum phase HRTF does not reproduce all of the phase components of the HRTF, thus the higher frequency phase components are lost. The minimum phase HRTFs presenting no perceptual consequences would indicate that the loss of the higher frequency phase components are not of great perceptual importance [24] [29]. Having no perceptual consequences may be contributed to the accuracy of the low frequency phase component of the HRTF being accurately represented by the minimum-phase HRTF [29]. The phase components of the original HRTF and the minimum phase HRTF is shown in Figure 3.3 for low frequencies and Figure 3.4 for the high frequencies to illustrate these points.

![Figure 3.3](image.png) The phase components of the original HRTF and the minimum phase HRTF at relatively low frequencies
Figure 3.4) The phase components of the original HRTF and the minimum phase HRTF at relatively high frequencies

3.2.2 Interpolation of HRTF Datasets

The measured HRTFs are typically measured at specific increments of azimuth and elevation points on a sphere around a subject. The resulting spatial dataset may lack the desired spatial location, requiring an interpolation of the HRTF dataset. These interpolation methods also lend themselves to a dynamic process, where head movements have been applied to the HRTFs to provide a realistic feel of moving sound with no noticeable discontinuity.

There are a number of HRTF interpolation methods that can be implemented in either the time or frequency domain. The most simple and straightforward method applies linear interpolation using the nearest neighboring HRTFs in the time-domain [30]. The linear interpolated HRTF in the time domain or Head-Related Impulse Response (HRIR) for the arrival angle \( \theta \) is found by:
\[ \text{HRTF}(n, \theta) = \alpha \text{HRIR}(n, \theta_a) + (1-\alpha)\text{HRIR}(n, \theta_b) \]  

(3.9)

where

\[ \alpha = \frac{(\theta_b - \theta)}{(\theta_b - \theta_a)} \]  

(3.10)

and \( \theta_a \) and \( \theta_b \) represents the arrival angle of the nearest neighboring HRTFs [31].

Another commonly implemented HRTF interpolation is the cubic interpolation method, which is not limited to the two nearest available HRTFs. The cubic method uses a cubic interpolation of the four, ideally nearest neighboring HRTFs. The interpolated HRTF does not have to fall perfectly in between the bounds \((\theta_1 < \theta_2 < \theta < \theta_3 < \theta_4)\) the only constraint is that it falls between the outer bounds\((\theta_1 < \theta < \theta_4)\). The cubic interpolated HRIR for the arrival angle of is \( \theta \) found by:

\[
\text{HRTIR}(n, \theta) = \left( \frac{(\theta - \theta_2)(\theta - \theta_3)(\theta - \theta_4)}{(\theta_1 - \theta_2)(\theta_1 - \theta_3)(\theta_1 - \theta_4)} \right) \text{HRIR}(n, \theta_1) \\
+ \left( \frac{(\theta - \theta_1)(\theta - \theta_3)(\theta - \theta_4)}{(\theta_2 - \theta_1)(\theta_2 - \theta_3)(\theta_2 - \theta_4)} \right) \text{HRIR}(n, \theta_2) \\
+ \left( \frac{(\theta - \theta_1)(\theta - \theta_2)(\theta - \theta_4)}{(\theta_3 - \theta_1)(\theta_3 - \theta_2)(\theta_3 - \theta_4)} \right) \text{HRIR}(n, \theta_3) \\
+ \left( \frac{(\theta - \theta_1)(\theta - \theta_2)(\theta - \theta_3)}{(\theta_4 - \theta_1)(\theta_4 - \theta_2)(\theta_4 - \theta_3)} \right) \text{HRIR}(n, \theta_4) 
\]  

(3.11)

The performance of the interpolation techniques are evaluated against the Mean-Squared-Error (MSE) between the interpolated HRIR (\( \text{HRTIR} \)) and the measured HRIR (HRIR) for the arrival angle of \( \theta_i \):

\[
\text{MSE} = 10 \log_{10} \sum_{n=0}^{L} \frac{\| \text{HRIR}(n, \theta_i) - \text{HRTIR}(n, \theta_i) \|^2}{\| \text{HRIR}(n, \theta_i) \|^2} 
\]  

(3.12)

The performance of the time-domain interpolation methods can be improved by time-aligning them before interpolating the HRIR as shown in Figure 3.5, with the interpolated time-aligned
counterparts having a lower MSE. Figure 3.5 shows the interpolated MSE of the KEMAR manikin HRIRs for different interpolating techniques, with an angle of 0° corresponding to a sound directly in front of the KEMAR Manikin. The time-aligning of the HTRFs or HRIR is discussed in the next section.

Figure 3.5) The MSE between the interpolated HRIR and the measured HRIR for the KEMAR manikin [19].

The time aligning of the HRIR is done so all of the HRIR will occur at the same time, pertaining to the arrival angle of 0°. The difference in the delay of the HRIRs can be extracted from the HRIR data by finding where the maximum cross correlation occurs. The HRIRs then shift according to their relative delay. The left and right HRIR are aligned separately.

The interpolated HRIR are reconstructed by properly shifting the HRIR data by the appropriate delay (in the time domain) or by adding the proper phase delay (in the frequency domain) with an all pass filter. The appropriate delay is found using the interpolation technique with the HRIRs being replaced with their relative delay found prior.
3.2.3 Head Movement Implementation Consideration

The generalization and the interpolation of the HRTFs were discussed in the context of adding a dynamic process that incorporates head movement. Head movement can be incorporated into the creating of the virtual 3D representation by:

1) Find the current head-position of the listener
2) Extracting or interpolating the correct HRTF according to the head position of the listener
3) Transform the desired sound source with the extracted HRTF with (3.1)
4) Output the transformed (binaural) sound back to the listener
5) Repeat steps 1-4 as needed

These steps must be done in real-time to give the listener the desired head movement affects. By eliminating the artificial constraining of the listener head, the listener cognitive reasoning skills increase better equipping the listener in pinpointing the location of the sound source.

Besides the limitations of the HRTFs being measured at specific points of azimuths and elevations, the time-invariant nature of the HRTFs will hinder the returning of head movement, also. Specifically, the length of the impulse response of the time-invariant HRTF is required to be small compared to the time scale or quasi-stationary frame of head motion [32]. In other words, the length of the filter will limit the segment size of the microphone array data that can be processed at one time and/or the maximum allowable rate of head movement. The amount of information that can process at one time is governed by:

\[ L_{\text{FFT}} > \left( Q - \frac{(L - 1)}{4} \right) + L - 1 \]  \hspace{1cm} (3.13)

where \( L_{\text{FFT}} \), \( Q \), and \( L \) are the length of the Fourier transformation, array data, and filter. When (3.13) not hold true, the array data can be zero padded out to \( L_{\text{FFT}} \) [33]. This processing is done to avoid time domain aliasing.
The processing of the microphone array data is done in the frequency domain. The microphone array’s data is transformed using a Fourier Transformation. The Fourier Transformation defines the array data in terms of infinite duration, where a finite segment of data is treated as a periodic signal, with the period equal to the duration of the data segment. This means that if the signal is not long enough, the Fourier transformation can create differences between the starting and ending values of the acoustical signal. These differences create undesirable clicks and tones, referred to as musical noise.

The differences between the starting and ending values of the segmented data can be eliminated by placing less of an emphasis on these end points. By adding a windowing function \( w(\cdot) \) to the segmented microphone data the end points are forced to zero. The \( m^{th} \) microphone segmented data for the \( i^{th} \) frame is then given as:

\[
x_m(i) \triangleq w x_m(iR, ..., L_{\text{FFT}} + iR)
\]

where \( R \) is the frame or hop step size, which is the number of samples advanced each successive window. There is no constraint on the window function \( w(\cdot) \), only that the window overlap adds to a constraint for the length of \( R \) being used. The windowing must be performed in a way that the signal is not changed and perfect reconstruction is attained. This thesis uses the overlap-add method of (3.14) with a Hann windowing function [33].

A visual example of the overlap-add method is shown in Figure 3.6, where the original sine wave is cut up into multiple segments using (3.14). The segments are advance by \( R/2 \) and overlaid upon each other as shown in the center window of Figure 3.6. The results of adding all the windows together is shown in the right window of Figure 3.6
The returning of head movement to the listener is not just computationally complex, but also requires additional hardware consideration. Ultimately, this process is not explored in the work of this thesis. The elimination of the artificial constraining of the listener’s head movement is done to explore the cause of front to back confusion and completed with a predetermined head movement. Filtering with respect to the head movement is not impossible and can be seen in [34], [35], and [32].

3.2.4 Inherent Limitations of HRTFs

In the next Chapter, the restoring of the missing HRTFs with a mapping process does not use the subject’s own HRTF; a foreign HRTF is used. The HRTFs that are generated by our body’s head, torso, and pinna are very person specific. That is, if what one person is hearing is relayed to another, there is a level of inaccuracy created by not having the original person’s
specific spatial information. The main cause of front/back confusion put forth in Chapter 5 is shown to be the use of foreign HRTF.
Chapter 4. HRTF Matching

The aim of this thesis is the mapping of the microphone array’s response to match the characteristics of a human subject, HRTFs. This mapping is need because humans have a remarkable ability to locate sound sources with only two sensors, their ears. Though, when listening to a response of a microphone array this ability is lost. This inability to localize sound sources is caused by the array not containing the spatial characteristics of the human body, the HRTFs. This Chapter explores the mapping process needed to map the microphone array’s response to match the characteristics contained in the HRTFs.

This mapping process is first explored with the limitation of a head-mounted system. These systems have the listener immersed within the environment and relay on the acoustic information being relayed to them to hear. The relayed information of the microphone array’s response does not contain the correct HRTFs required to spatial perceive the location of a sound source, requiring a ‘HRTF matching’ to restore the HRTFs. The aim of this work is on microphone arrays, not just a head-mounted system. Thus after exploring the limitations of this system, the HRTF matching method is modified to be used for any microphone array geometry.

Before continuing it should be noted that the audience of this thesis is external to the acoustic environment and is listening to the environment via a recording through a pair of headphones. This situation is repeated for the head-mounted system. Normally for the head-mounted system depicted in Figure 4.1, the acoustic information that is captured by the individual microphones is relayed back to the listener after the HRTF matching. This path is shown in green in Figure 4.1. The intended audience of this thesis is external to the acoustic environment, wherefore the process information is stored before relaying it back to the listener. This is shown as the blue path in Figure 4.1.
Figure 4.1) Schematic of a head-mounted system giving the microphone array the characteristics of the HRTFs. This real-time process is shown as the green path. The blue path represents the storing of the information before relaying it to the listener.

4.1 Classic Approach of Beamforming

The head-mounted system relays the acoustic information to the listener in real-time, the addition of the HRTFs is required to be low in computational complexity. Therefore the HRTF matching is a data-independent beamformer that maps the microphone array’s response to match the characteristics of a human subject, the HRTFs. The data-independent beamformers are chosen to present a specified response for all incoming signal scenarios, independent of arrival angle. A data-independent beamformer is a preferred method over other conventional techniques.
for its lower computational complexity, ease of implementation, and absence of a required control algorithm to avoid signal distortion and/or signal cancellation [36].

The basic premise of beamforming consists of signals of interest arriving from specific directions and being captured by microphones located at different spatial positions. The captured signals are weighted so signals from a specific direction are constructively added and signals not pertaining to that direction are destructively added (attenuated/null). As a result, a specific spatial response of the array system is achieved with ‘beams’ pointing to the desired signals and ‘nulls’ toward the interfering ones [37].

The beamformer forces the array response to ‘look’ in one set direction by choosing the appropriate weights, $w$ of (2.4). The classic beamforming approach steers the response of the array by adding the appropriate delays for the different microphone outputs [37] that satisfy the following two constraint conditions [2]: 1) the signals pass the ‘look’ direction without distortion, and 2) minimize the total output power, where the power is minimized for all the signals not coming from the ‘look’ direction, $\theta$.

The weighting is performed by applying a complex weight to the output of each microphone and summing across the spatial aperture. In the frequency domain, the weights associated with the first optimization constraint are:

$$w^{hi}(\omega) a(\omega, \theta) = 1$$

where either the near-field steering vector $d(\omega, \theta)$ or the far-field array steering vector $a(\omega, \theta)$ can be used. The array steering vector contains all $M$ transfer functions associated to the incident direction of $\theta$:

$$a(\omega, \theta) = [H_1(\omega, \theta), ..., H_M(\omega, \theta)]^T$$

(4.2)
where $M$ is equal to the number of microphones in the array, i.e., $M = 2$ for the array steering vector of the human subject, which is the HRTF. The difference between the near-field array steering vector and the far-field array steering vector is that the near-field array steering vector contains both the gain and phase components, whereas the far-field array steering vector only contains the phase components. To be consistent with most literature, the far-field model is utilized in this derivation. The second optimization constraint to minimize power of the beamformer output is then [2]:

$$
p = E[|\tilde{y}|^2] = E[|w^H \tilde{x}|^2] = w^H Rw
$$

$$=
|\tilde{\sigma}|^2 + w^H E[|\tilde{\nu} \tilde{\nu}^H|]w
$$

$$=
|\tilde{s}|^2 + \sigma^2_w w^H w
$$

(4.3)

Then the optimized $w_o$ is obtained by the minimization of the noise component:

$$w_o = \min_w (w^H w)$$

(4.4)

The optimized beamformer is found from the optimization constraints of (4.1) and (4.4) [2]:

$$w_o(\omega, \theta) = \frac{a(\omega, \theta)}{a(\omega, \theta) a^H(\omega, \theta)} = a(\omega, \theta)$$

(4.5)

The $\frac{1}{M}$ scaling factor of the optimized weight has been omitted to be consistent with the majority of the literature [9]. The optimized weight is referred to as the Bartlett beamformer [9] or a simple Delay-and-Sum Beamformer (DSB).

### 4.2 HRTF Matching

The HRTF matching is a match filter ($w_M$) that is an extension of a DSB. Not only does the $w_M$ compensate for the relative delay between the individual microphones similar to the DSB, it compensates for the mismatches in the responses using a filter and sum approach. Formally, the weights in a data-independent $w_M$ are designed so the beamformer output approximates a desired response independent of the array data [38]. By choosing the appropriate
weight $w_{M,m}$, the $m^{th}$ microphone actual response $x_m(\omega, \theta)$ can match the desired response, $r(\omega, \theta)$

$$r(\omega, \theta) = w_{M,m}^H(\omega, \theta)x_m(\omega, \theta) \quad (4.6)$$

where, $\theta$ is the desired ‘looking or steering direction,’ ($\theta$ of the desired and actual response can be different, but for now it is assumed that they are the same).

The goal is for the desired response to be that of the listener, (4.6), which can be expressed in terms of the HRTF by:

$$w_{M,m}(\omega, \theta) = \left( \frac{r(\omega, \theta)}{x_m(\omega, \theta)} \right)^H = \left( \frac{\text{HRTF}(\omega, \theta)s(\omega)}{H_m(\omega, \theta)s(\omega)} \right)^H = \left( \frac{\text{HRTF}(\omega, \theta)}{H_m(\omega, \theta)} \right)^H \quad (4.7)$$

where $H_m$ is the acoustic transfer function of the $m^{th}$ microphone, and the HRTF($\omega, \theta$) containing both the left and right HRTF, defined by (3.2). The array HRTF matching is then:

$$w_M(\omega, \theta) = \left( \frac{\text{HRTF}(\omega, \theta)}{d(\omega, \theta)} \right)^H \quad (4.8)$$

where, $d(\omega, \theta)$ is the array steering vector. $d(\omega, \theta)$ contains all $M$ transfer functions associated to the incident direction of $\theta$.

The $w_M$ of (4.8) is shown for a given arrival angle. As humans we can locate sound in all directions, the complete HRTF matching then needs to match the response for this range as well. The HRTF matching is then chosen to match the complete spatial characteristics of the microphone array of the human subject’s HRTF for all arrival angles. This can be done by minimizing the MSE between the actual and desired spatial differences at $P$ points, where $P = nN$, $n$ being the set of all of azimuth points and $N$ representing the set of all elevation points [6]. The scope of this work only pertains to the azimuth direction, then $P$ only pertains to the set of all measured elevation points. The $(M \times P)$ spatial matrix that is generated is the array.
manifold matrix which contains all $M$, microphone for the array manifold matrix of the microphone array:

$$D(\omega, P) = [d(\omega, \theta_1), ..., d(\omega, \theta_n)]^H$$

and all $M$ ears for the array manifold matrix of the human subject:

$$R(\omega, P) = [\text{HRTF}(\omega, \theta_1), ..., \text{HRTF}(\omega, \theta_n)]^H$$

The minimization criterion is given as:

$$\min \left( R(\omega, P)^H w_{MF}(\omega) - D(\omega) \right)$$

The solution for (4.11) is given as:

$$w_M(\omega) = R(\omega, P) D^+(\omega, P)$$

where $+$ represents the pseudo-inverse. The pseudo-inverse is on $D(\omega, P)$ since the array manifold matrix will likely not be a square matrix. The pseudo-inverse is a generalized inverse that can be performed on 2-dimensional matrices of any size [6]:

$$D^+(\omega, P) = \left( D^H(\omega, P) D(\omega, P) \right)^{-1} D^H(\omega, P)$$

The design of the HRTF matching of (4.12) is for a head-mounted array shown in Figure 4.1. The choices of $P$ and $M$ can be changed to better optimize the design of the HRTF matching. The system shown in Figure 4.1 is for a monaural setup, an HRTF matching is required for each ear. The HRTF matching then for the left channel will use only the left ear HRTFs in the array manifold matrix of (4.10) and only the microphones that pertain to that side of the microphone array will be used in the array manifold matrix of (4.9). The HRTF matching for the right channel is carried out the same way. This HRTF matching of (4.12) is referred to as the ‘monaural HRTF matching’ method.
4.2.1 Limitations of Monaural HRTF Matching

The main benefit of data-independent beamformers is that the weights do not depend on the incoming microphone array data. The weights are designed so the beamformer response approximates the human subject independent of the array signals or statistics [36]. This is to the benefit, but more so the determent of the mapping process. The drawback of the monaural HRTF matching is that the data-independent beamformer has only one set of weights (one for each frequency) that must map the gain and phase of the microphone array to the HRTFs, for all directions. This limitation will be shown by mapping the head-mounted, Headphone Array phase and gain components to match the HRTFs of the KEMAR Manikin. The Headphone Array, discussed in detail in Appendix A, was designed for the placement of the microphones to be as close to the human ear as possible, as well as to maximize the number of microphones on the headphone.

From Chapter 3, the HRTFs can be approximated as a minimum phase filter, i.e., the phase component of the HRTFs are modeled as a pure delay of the ITD. Thus, the mapping of the phase component of the microphone array is a matching of the Time Difference (TD) to the KEMAR’s ITDs. The matching of the Headphone Array TD is shown for two different microphone placements in Figure 4.2. Figure 4.2 shows microphone 6 placement that is similar to the location of the ear and microphone placement 4 that is in front of the of the ear canal. One can see in Figure 4.2 that the weights of the monaural HRTF matching have difficulty mapping the microphone placement 4 ITD to match the KEMAR’s ITDs for all arrival angles.

In truth the matching of the phase components is not really seen in the overall delay, the non-HRTF matching of the TD is the same as those shown in Figure 4.2. The phase matching is seen more in correcting the difference (error) between the individual microphones. The monaural HRTF matching not correcting the delay is the nature of the system having to correct the delay
for all azimuth directions. What this entails is that the placement of the microphones is more crucial to the matching of the TD than the matching process itself.

![Image showing microphone placement and ITD comparison]

Figure 4.2) Comparison of ITD for microphone 4, 6 of the head-mounted Headphone Array match to the KEMAR manikin ITD at an elevation of 0°.

From Chapter 2, the gain components of the HRTF can be split into two different categories: 1) the ILD and 2) the spatial modification of the pinna. The nature of the Headphone Array being mounted on the head naturally captures the ILD (head-shadowing and torso reflection). The mapping of the microphone array gain components to match the KEMAR manikin deals mainly with matching the spatial modification caused by the pinna. Figure 4.3-4.5 show the spatial pattern of the KEMAR Manikin left ear with a normal size pinna, the spatial pattern of one of the left microphones of the head-mounted Headphone Array, and the mapped spatial pattern of the head-mounted Headphone Array, respectively. The similarities between Figure 4.5 and Figure 4.3 would indicated the monural matching does a good job matching the gain for all the arrival angles.
Figure 4.3) Array manifold matrix of KEMAR Manikin left ear with a normal size pinna

Figure 4.4) Array manifold matrix of the head-mounted Headphone Array
Figure 4.5) Mapped array manifold matrix of the head-mounted Headphone Array to the KEMAR manikin left ear with a normal size pinna

Depending on the microphone placement, the ILD and ITD approximate the KEMAR ILD and ITD relatively well. The question that arises is whether ‘approximate’ is close enough for the listener to have enough spatial information to perceive the location of a sound source. Analysis of the results of a series of listener tests revealed that the monaural HRTF matching did not perform well. The listener test of the monaural HRTF matching that mapped the microphones of the Headrest Array, which have spacing dissimilar to those of a human subject, revealed that the monaural HRTF matching did not provide adequate spatial cues to the listener. Listeners could not locate the signal regardless of angle. Listeners stated that sounds seemed to come from either inside their head or above them for all azimuths tested. The listener test of the monaural HRTF matching that had microphone placement similar to that of the human ear
(Headphone Array) demonstrated that listeners had a more accurate spatial perception of the sounds’ location.

The basic analysis reveals that the monaural HRTF matching did not perform well in providing listeners with spatial perception because hearing is greatly impacted by the placement of the microphone. The above analysis should not discount that the monaural HRTF matching is an adequate choice for the listener who is internal to the acoustic environment. These systems have the listener immersed within the environment and rely on the acoustic information being relayed to them. However, the intended audience of this thesis is external to the acoustic environment and where information is obtained by arbitrary array geometry. Advancing this HRTF match method would yield even greater disparity in listener’s spatial perception.

4.2.2 Reformulation of HRTF Matching

Removing the data-independencies from the beamformer leaves a few options. The most straightforward approach is the addition of a preprocessing control algorithm. Where the preprocessing control algorithm finds and separates the sound sources. The output of this process is the sound source arrival angle and the extracted sound source. The mapping of the extracted sound sources to match the characteristics of a human subject is done with a HRTF matching.

The HRTF matching modifies the spatial characteristics of the microphone array to match those of the human’s subject, HRTFs. The preprocessor control algorithm is required to modify the spatial attributes in a known way that will not modify the spatial characteristics of the array or do so in a known quantity. This means the HRTF matching is dependent on the modifications done by the preprocessor control algorithm. The HRTF matching of (4.7) then will be defined by the preprocessor’s beamformer output (the extracted sound). For the simple case of a single sound source, the HRTF matching is found from:
\[ Y(\omega, \theta) = w_M^H(\omega, \theta)w^H(\omega, \theta)X(\omega, \theta) = w_M^H(\omega, \theta)w^H(\omega, \theta)d(\omega, \theta)s(\omega, \theta) \] (4.14)

with \( Y(\omega, \theta) = \text{HRTF}(\omega, \theta)s(\omega) \). Then the HRTF matching is defined as

\[ w_M(\omega, \theta) = \left( \frac{\text{HRTF}(\omega, \theta)}{w^H(\omega, \theta)d(\omega, \theta)} \right)^H \] (4.15)

This process is illustrated in Figure 4.6, where the acoustic information is captured by the individual microphones and the preprocessing control algorithm finds and separates the sound source before the HRTF matching of (4.15) gives the sound sources the HRTF characteristics for that particular source direction \( \theta_s \). The system shown in Figure 4.1 is for a binaural setup, a HRTF matching that outputs both the left and right ear responses. The HRTF matching of (4.15) is referred to as the ‘binaural HRTF matching’ method.

The binaural HRTF matching process is depicted in Figure 4.6 for the head-mounted, Headphone Array but can be used for any microphone array geometry. Figure 4.6 again shows that the audience is external to the acoustic environment such that the mapped sounds are stored before being played to the listener.
4.2.3 HRTF Match with Head Movement

For head motion to be incorporated, the previously assumed same look direction of the desired and actual responses has to be changed. This is due to the fact that the head is moving but the array is stationary. A set of filters that correspond to different look directions of the human head needs to be implemented. The HRTF matching then would filter the response of the array corresponding to the fixed look direction of the array, $\theta_S$, to the non-fixed look direction of the listener’s head direction, $\theta_{\text{Head}}$.
\[ w_{MF}(\omega, \theta_s, \theta_{Head}) = \left( \frac{\text{HRTF}(\omega, \theta_{Head})}{w^H(\omega, \theta_s) d(\omega, \theta_s)} \right)^H \] (4.16)

The set of HRTF filters required for different look directions implementation challenges are discussed in Chapter 3.2.3
Chapter 5. Listener Test of Binaural HRTF Matching

The mapping of the microphone array’s response to match the characteristics of the human subject’s HRTF is done with a HRTF matching. The limitations of the monaural HRTF matching used on head-mounted systems were explored in Chapter 4, whereby being independent of the arrival angle the mapped response did not have enough spatial information for a listener to precisely localize a sound source. With the addition of a preprocessor control algorithm before the HRTF matching process, the overall process became dependent on the arrival angle. This process, referred to as a binaural HRTF matching, will be shown through a series of listener tests in this Chapter to illustrate that:

- The listener’s ability to locate a sound source with the binaural HRTF matching method is comparable to the creating a virtual 3D sound representation by convolving the HRTFs with a known sound.
- The psychoacoustic error of front to back confusion is caused by the use of foreign HRTFs.
- The addition of head movement increases a listener’s ability to locate a sound source.

The mapping of the microphone array’s response to match the characteristics of the human subject is the focus of this work and not the preprocessor control algorithm. The measurements taken by the array are done in a free-field (anechoic) environment with the location of the sound source known. Under these idealized conditions, the preprocessor control algorithm has been replaced with a simple DSB. The DSB will extract the desired source for the known direction of the sound source. The binaural HRTF matching will either use (4.15) or (4.16) to map the microphone array’s response to match the characteristics of the human subject. This complete process is shown in Figure 5.1.
The acoustic information in Figure 5.1 is captured by the Headphone Array before the preprocessor that has been replaced with a DSB extracts the sound source. The binaural HRTF matching of (4.15) or (4.16) then maps the microphone array characteristics to the human subject’s HRTF. Again, the audience of the listener tests is external to the acoustic environment such that the mapped sounds are stored before they are played to the listener. This investigation is accomplished by mapping the measurements taken from two arrays, the head-mounted, Headphone Array and the Headrest Array (see Appendix A).

![Schematic of the binaural HRTF matching](image)

Figure 5.1) Schematic of the binaural HRTF matching that maps the microphone array’s response to match the characteristics contained in the HRTFs.
5.1 First Listener Test of the Binaural HRTF Matching with Fixed Head Position

The first binaural HRTF matching listener test is a broad investigation to gauge the performance of the mapping process. The performance of the binaural HRTF matching is gauged by the listener’s ability to accurately determine the location of the azimuth of a sound source. The first listener test investigates the effect of fixed head positions, the type of sound used, and the number of microphones in the array has on the listener’s ability to locate the azimuth of a sound source. To accomplish this, the binaural HRTF matching maps the measurements taken from the microphone arrays for three different sound types (music, noise, and voice) from arrival angles that range from $0^\circ$ to $355^\circ$ in $5^\circ$ increments with varying numbers of microphones. The binaural HRTF matching mapped the microphone array’s response to match the characteristics of the human subject for the HRTFs of the KEMAR Manikin with a normal size pinna [19] and the HRTFs of Subject 1025 of the ‘Listen Group’ [20]. Two different measured sets of HRTF were used to see if bias toward one particular HRTF was noted in the listeners’ responses. The HRTFs of subject 1025 pertains to an actual person. The KEMAR manikin HRTFs do not pertain to an actual person. The KEMAR manikin, though, represents a statically average adult human. If you measure all the adult men and women heads, torsos, and pinnae, you will obtain the KEMAR manikin.

The first listener test tested the listener’s ability to accurately determine the location of the azimuth of a sound source. In a quiet study lounge at Virginia Tech, each participating test subject was told that they were taking a hearing test to determine how well they could determine the direction of a particular sound. The test sounds or mapped sounds were played to the test subject with the sound type, HRTF, and arriving angles randomly chosen. Each test subject was
asked to find the azimuth of the sound source by circling where in Figure 5.2 they believed the sound was coming from. The first listener test was broken down into two parts:

Part A) Binaural HRTF matching mapped the data from the head-mounted Headphone Array only, where 75 mapped sounds were played to 13 test subjects.

Part B) Binaural HRTF matching mapped the data from the Headrest Array only, where 43 mapped sounds were played to 10 test subjects.

Four subjects participated in both studies. The results of the first listener test are explored using Part A. Since the results of Part B mirror those of Part A, Part B is not explored. The results of Part B mirroring those of Part A indicate that the binaural HRTF matching is not as susceptible to the array’s geometry as the monaural HRTF matching is.

The performance of the binaural HRTF matching is gauged by the listener’s ability to precisely locate the sound source’s azimuth. The performance of the mapping process is shown in Figure 5.3 as ‘degrees off’ the actual location of the sound source. The smaller the degrees off,
the better the listener and the mapping process performed. The sound type that was used in generating the mapped response is shown in Figure 5.3 corresponding to different colors. The first listener test revealed that the type of sound did not have an impact on the test subjects’ ability to locate the azimuth of a sound source. The subjects’ responses follow a path that did not change when different sounds were introduced. This path is shown in Figure 5.3 as a black dashed line.

The path of this line with an analysis of the raw data reveals that the test subjects suffered from the psychoacoustic error of front to back confusion. Front/back confusion is caused by the listener believing that the sound is behind the listener but actually is coming from in front of the listener or that the sound is in front of the listener but actually is coming from behind the listener. Figure 5.3 does not take front/back confusion into consideration and is the reason for the changes in the mean response of the listener. By taking front/back confusion into consideration and correcting the listener’s confusion by placing the listener’s response in the correct front or back half plane, the listener’s mean response becomes constant at 20° off the actual arrival angle.

![Listeners’ responses of the first listener test for the Headphone Array. The colors represent a particular sound; red is music, blue is voice, and green is noise. Each dot corresponds to a different test subject’s response.](image.png)

Figure 5.3)
The issue of front/back confusion is an expected problem and is associated with creating a virtual 3D sound representation with HRTFs. In a normal listening scenario, front/back confusion is naturally overcome by head movement and the spatial modification of the pinna. The first listener test did not map the responses according to the actual position of the listener’s head (artificially constraining the head) and used foreign HRTFs, which introduces spatial characteristics that are dissimilar to the listener’s.

The percentage of front/back confusion for the first listener test is shown in Figure 5.4. The percentage of front/back confusion shown in Figure 5.4 reveals that for the sounds that are mapped to the front half plane (285°-75°), listeners are confused 50% of the time, except for a dip directly in front of the listener, 0°. For sounds that are mapped to the back half plane, listeners are only confused 25% of the time, except for a peak directly behind the listener, 180°. The respective dip and peek around 0° and 180° in Figure 5.4 are assumed to be a bias for sound coming from the front half plane where the main cues for determining azimuth (ILD and ITD) are not present. This assumption was validated by test subjects stating that when they were unaware of the sound’s origin, they assumed 0°.
The rationale for why there is more confusion of the sound mapped to the front half plane than in the back half plane lies with the nature of the human body’s HRTFs. The HRTFs generated by our body’s head, torso, and pinna are very person specific such that if you take what someone else is hearing and transplant that signal to another person, there will be a level of inaccuracy. The spatial modifications of the pinna are also greater for sounds originating in the front half plane than in the back half plane. Hence in the first listener test, the foreign HRTFs did not provide the listener’s auditory system with the information to conclude that the sounds originated in the front half plane.

The modifications of the human body HRTFs for different arrival angles are shown in Figure 5.5. In Figure 5.5, the azimuth angle of 0° corresponds to a sound source directly in front of the listener that is swept counterclockwise around the listener. The KEMAR left ear in Figure
5.5 corresponds to the normal size pinna and the KEMAR’s right ear corresponds to the large size pinna. One should note the difference caused by the two different pinna sizes at higher frequencies.

Figure 5.5) Spatial response of the KEMAR with a normal size pinna (left) and the KEMAR with a large size pinna (right) [19]

The final question investigated in the first listener test is the effect varying the number of microphones has on the binaural HRTF matching. The test subjects’ responses indicate that varying the number of microphones has no impact. Only one microphone is needed to map the microphone array’s response to match the characteristics of the human subject. However, the listener test tested the binaural HRTF matching and not the preprocessor. The preprocessor was replaced with known quantities; the location of the sound source was known prior to the binaural HRTF matching. The preprocessor that is required to find and locate the sound sources will have a direct impact on the performance of the mapping process, which is required to have more than one microphone. The effects of the number of microphones on the preprocessor control algorithm are discussed in Chapter 6.
5.2 Second Listener Test of the Binaural HRTF Matching with Head Movement

Although front/back confusion is an expected problem associated with a virtual 3D sound representation, the binaural HRTF matching could increase the occurrence of this psychoacoustic error. To pinpoint the factors that contribute to the front/back confusion, a second listener test was performed. The second listener test changed the factors that contribute to front/back confusion one at a time. These factors are the fixed head position, foreign HRTFs, and possibly the binaural HRTF matching.

The binaural HRTF matching of (4.16) that returns the head movement of the listener is a complex, real-time process. That process is not just computationally complex; it also requires additional hardware considerations. By using a simulated head movement, the fixed head movement constraint is eliminated, giving the listener more spatial knowledge while keeping the computational complexity low.

For a simulated head movement, the DSB output of the microphone array was mapped with the binaural HRTF matching that rotated around the listener’s head at a constant rate. The binaural HRTF matched the DSB output according to a look direction of $0^\circ$ with a clockwise or counterclockwise sweep to a final resting azimuth $5^\circ$, $10^\circ$, $15^\circ$, $20^\circ$, $30^\circ$, $45^\circ$, and $60^\circ$. Only the front half plane was investigated because the first listener test showed subjects had more difficulty locating sounds in the front half plane versus the back half plane. To accommodate a smooth transition between the HRTFs, a time-aligned cubic interpolation was used. The HRTFs are either a minimum phase generalization or time-aligned version. The delay represented by the ITD was extracted from the original HRTF and applied as an all-pass filter.

The sounds for the second listener test are generated with a virtual 3D sound representation with simulated head movement with either a binaural HRTF matching method or
a convolution method that convolves a known sound with the HRTFs. The known sounds are male and female voices that were captured by the microphone array for the binaural HRTF matching. The convolution method was added to the second listener test to determine whether the binaural HRTF matching process increases the occurrence of front/back confusion. This is because the sounds generated by the convolution method are absent from the spatial modifications of the array and mapping process.

The second listener test was broken down into two separate parts. The first part of the second listener test only used HRTFs from the KEMAR Manikin with a normal size pinna [19]. The second part used a subject-specified HRTF from the Listen Group’s subject data [20]. The subject-specified HRTF was chosen by the listener as the best matched HRTF available.

The second listener test was conducted in the same manner as the first listener test. In a quiet study lounge at Virginia Tech, each participating test subject was told that they were taking a hearing test to determine how well they could determine the direction of a particular sound. Each test subject was asked to find the final resting azimuth of the sound source by circling where in Figure 5.2 they believed the sound was coming from.

The goal of the second listener test was to pinpoint the factors that contribute to front/back confusion and how they pertain to the binaural HRTF matching. Each of the sounds in the second listener test simulated head movement by rotating a sound source around the listener’s head. The performance of the binaural HRTF matching is gauged by the listener’s ability to precisely locate the sound source’s azimuth. The performance of the mapping process is shown in Figure 5.6 as degrees off, with the mean response shown as a black dotted line. The simulated head movement improved the listener’s ability to locate the azimuth of a sound source by 20%, compared to the first listener test with a mean response that was 20° off the actual angle.
verse 16° off with simulated head-movement. Although the performance of the listener to locate a sound source improved, it did not resolve the front/back confusion.

![Figure 5.6](image)

Figure 5.6) Listeners’ responses of the second listener test shown as degrees off from the actual response. Although each dot corresponds to a different test subject’s response, there are many points that overlay each other.

The second listener test uses a virtual 3D sound representation generated with a binaural HRTF matching and a convolution of the HRTFs with a known sound using (3.1). By providing two different generated sounds, the binaural HRTF matching is verified to not introduce any unwanted front/back confusion. The different sounds provided comparable results. Test subjects answered the same over 70% of the time, with an overall median response difference of 7.5°. Some of the different answers are associated with front/back confusion. This occurred less than 10% of the time, with front/back confusion occurring more often in responses absent the binaural HRTF matching process.

The only remaining cause of front/back confusion is the use of foreign HRTF. Subject-specified HRTFs should lower the occurrence of front/back confusion, although this was not the
case. Using subject-specified HRTF showed no improvement or slightly more front/back confusion.

After compiling all of the test subjects’ responses, the results of the second listener test indicate that the cause of front/back confusion is the use of foreign HRTFs. The Listen Group database did not provide enough sampling for listeners to find a set of HRTFs that matched their own. This conclusion is based on discussions with the test subjects. Subjects that had issues with front/back confusion reported that their subject-specified HRTFs were accurate only for a small set of angles, i.e., they could not find a set of HRTFs that matched their own. Additionally, there were two subjects with less than 10% front/back confusion and one subject with 100% front/back confusion. A more detailed investigation into the study of the effects of generic versus individualized HRTFs [39] showed a significant increase in the performance of using individual HRTFs over generic HRTFs. It is the conclusion of the second listener test that the cause of front/back confusion is the use of foreign HRTFs.

5.3 Third Listener Test: Validation of Front/Back Confusion

The conclusion of the second listener test was that the cause of front/back confusion was the use of foreign HRTFs, i.e., HRTFs other than the listener’s. This conclusion can be validated by the testing of Subject 3; the subject in the second listener test with 100% front/back confusion. Although Subject 3 could be considered an outlier, the argument will be made for the importance of using a subject’s own HRTF. A third listener test was then designed for Subject 3 using the individual’s own HRTFs to verify the conclusion on the cause of front/back confusion.

The measurements of Subject 3’s HRTF were coarsely measured in comparison to the HRTF measurements of the KEMAR Manikin [19] and the subjects of the Listen Group [20]. This process is outlined in Appendix B. The extracted HRTFs of Subject 3 and those of the
KEMAR Manikin with a normal size pinna [19] were used to create a set of virtual 3D sound representation. These test sounds were generated by convoluting the HRTFs with a known sound using (3.1), with simulated head movement of 10°.

The third listener test focused on Subject 3’s ability to locate the azimuth of a sound source with a virtual 3D sound representation generated with Subject 3’s own HRTFs and a foreign HRTF. When Subject 3 was presented with sounds that were filtered with his own HRTF, the subject answered correctly as to what quadrant the sounds were mapped. When Subject 3 was presented with sounds that were filtered with a foreign HRTF, the subject’s answers only registered in the back half plane. These results verified the conclusion of the second listener test that the cause of front/back confusion is the use of foreign HRTFs.

5.4 Additional comment on the Second Listener Test

In the second listener test, the question of whether or not the listener perceived the motion of the sound was asked. For large head movements, sweeps of 45°, 60°, and 75°, test subjects always stated that they perceived the motion of the sound. For small head movements, the number of test subjects that perceived the motion of the sound decreased along with the sweep angle. An unexpected result was how the test subjects’ responded to how the sounds moved around their head. The common responses as to how they perceived the sounds traveling around their heads were:

- Moved in a straight line
- Moved in a diagonal line
- Moved in a rainbow arc (elevation component)
- The sound did not move at a constant angular rate
• The sound jumped from one angle to another

The problems of test subjects not perceiving the correct motion could be another issue caused by foreign HRTFs. However, the study of a virtual 3D sound representation exploiting more than one sound source theorized that single sound source environments provide unstable reference frames, which are unsuitable for perceiving the motion of a person or an audio cue [40]. This is mentioned because in the second listener test, the listeners were told to find the sound that started in the front and moved around their head in a clockwise manner to find the subject-specified HRTFs. Not a single test subject stated that the sound moved in a perfect clockwise manner. This could point to why the test subjects could not find a set of subject-specified HRTFs.
Chapter 6. Discussion of Process Limitations and Future Work

The aim of this thesis is to map the microphone array’s response to match the characteristics of a human subject. To accomplish this, a binaural HRTF matching was developed. The evaluation of the binaural HRTF matching in Chapter 5 mapped the microphone array data that was collected under ideal conditions of a free-field (anechoic) environment with the location of the sound source ‘known.’ In this analysis, the preprocessor control algorithm was not investigated and was replaced with known quantities. While this does not adversely affect our analysis, it does add an additional level of complexity and calls into question the statement that changing the number of microphones had no effect on the listener’s ability to spatially perceive the azimuth of a sound source.

In a realistic scenario, multiple sound sources with ‘unknown’ locations are possible, where the performance of the preprocessing control algorithm to locate and separate the sources has a direct correlation to the number of microphones in the array. These processes are impossible to do with only a single microphone. The added complexity and limitation to the binaural HRTF matching process are a limitation of the preprocessing control algorithm. The limitation of the preprocessing control algorithm though has a direct impact on the binaural HRTF matching. This Chapter presents a brief introduction into a realistic scenario to highlight the limitations of the binaural HRTF matching in this investigation and where future investigation is required.

6.1 Limitation of preprocessing control algorithm

The limitations of the preprocessing control algorithm and its effect on the binaural HRTF matching will be explored in the following scenario: A microphone array receives the
incoming signal of an unknown number of speakers. The preprocessing control algorithm is required to find the number of speakers and their location and to separate each of the individual speakers from the microphone array data.

6.1.1 Source Localization (Find the Number and Location of Sources)

The first step in the preprocessing control algorithm is to find the number and location of the speakers, a reciprocal process, i.e., one cannot be done without the other. Finding the location of the speaker’s Direction of Arrival (DOA) relies on the physical property in which signals coming from different locations in space will present different time delays from one microphone to the next. By steering the microphone array’s responses to different locations, peaks and valleys will occur in the Power Spectrum Density (PSD). The peaks in the PSD would indicate the presence of one or more speakers and a low value of PSD would mean the absence of a speaker [2]. This is the basic premise for virtually all DOA methods [41].

The PSD depends on the type of approach or beamformer used to steer the response. Though the Bartlett or DSB approach does not yield the most precise results for DOA estimation, it is a well understood and has already been derived in Chapter 4. For the Bartlett approach, the PSD is found by combining the Bartlett beamformer, (4.5), into the PSD given by (4.3):

\[ p(\omega_j) = a(\omega_j, \theta)^H Ra(\omega_j, \theta) \]  

(6.1)

where \( \omega_j \) represents the narrowband frequency.

The PSD of (6.1) is for narrowband signals, whereas the speaker voice is a broadband signal. The estimate of the DOA for broadband signals takes into account all frequencies (or a subset thereof) [42]. This is accomplished by taking the received broadband signals and
decomposing into narrowband components. By placing the previous \( L - 1 \) samples of each microphone into a spatiotemporal aperture matrix \( \overline{\mathbf{x}}(n) : ML \)

\[
\overline{\mathbf{x}}(n) = [\mathbf{x}(n), \mathbf{x}^T(n - 1), ..., \mathbf{x}^T(n - L + 1)]^T
\]  

(6.2)

where

\[
\mathbf{x}(n) = [x_1(n), x_2(n), ..., x_M(n)]^T
\]  

(6.3)

The decomposition is complete by transforming (6.2) into the frequency domain, \( \overline{\mathbf{X}}(\omega) \). The covariance matrix, \( \mathbf{R} \), is often required to be nonsingular, requiring the covariance matrix in (6.1) to be replaced by the sample covariance matrix:

\[
\overline{\mathbf{R}}(\omega) = \frac{1}{K} \sum_{k=1}^{K} \overline{\mathbf{X}}_k(\omega) \overline{\mathbf{X}}_k(\omega)^H
\]  

(6.4)

calculated over \( K \) frames.

The DOA estimate across the desired frequency band is performed with either a coherent or incoherent weighting of the PSD. Incoherently weighting the PSD is performed by averaging methods, i.e., arithmetic, harmonic, or geometric \([42]\). The DOA estimate of the speaker, \( \hat{\theta}_s \), is found by the peaks of steered PSD, incoherently weighted:

\[
\hat{\theta}_s = \max_\theta \left( \prod_{j=1}^{J} \mathbf{a}(\omega_j, \theta)^H \overline{\mathbf{R}}(\omega_j) \mathbf{a}(\omega_j, \theta) \right), 0 < \theta_s \leq 360
\]  

(6.5)

with geometric averaging.

The estimate of the DOA is limited by the performance and resolution of the beamformer. Beamforming suffers from the Rayleigh resolution limit \([43]\), which is mitigated only by increasing the width and number of microphones in the array \([44]\). Finding the precise location of the speaker is then dependent on the microphone array, as illustrated in Figure 6.1 and Figure 6.2. Figure 6.1 and Figure 6.2 Figure 6.1 shows the PSD for the head-mounted
Headphone Array with a single speaker coming from 50° and 150°. The plot shows the microphone array has a greater amount of difficulty locating the speaker coming from an arrival angle 50° compared to the arrival angle of 150°. In a natural weak-reverberated environment of a car, the Headrest Array could not locate or identify the correct number or location of two speakers.

Figure 6.1) PSD output of (6.5) for the head-mounted Headphone Array of a single speaker at an arrival angle of 50°
6.1.2 Source Separating

The map the microphone array’s response to match the characteristics of a human subject is done with HRTF matching. The HRTF matching modifies the spatial characteristics of the microphone array, requiring these characteristics to be intact in the extracted sound source. This is the reason why a beamformer, which maintains the spatial characteristics of the incoming signals, was used to extract the sources for a given arrival angle in Chapter 4.

However, beamforming suffers from the Rayleigh resolution limit [43], which is mitigated only by increasing the width and number of microphones in the array [44]. Subsequently, the ability of the beamformer to separate the given speaker is dependent on the microphone array.
Assuming that the preprocessing control algorithm has located the correct number and location of the desired sources, each speaker can be separated by a beamformer. The issue arises when the beamformer does not completely attenuate any or all the interfering sources from the desired speaker. The ramifications of interfering sources being present are that the binaural HRTF matching of the microphone array’s response to match the characteristics of a human subject will force the extracted sound source and the interfering source to have those characteristics.

This can be illustrated with the Bartlett beamformer extracting the source signals, \( S(\omega) \), according to the bearing \( \theta_S \), but fails to attenuate all of the interfering source, \( V(\omega) \), i.e.,:

\[
\mathbf{w}^H(\omega, \theta_S) \mathbf{X}(\omega, \theta_S) = S(\omega) + \alpha V(\omega)
\]

(6.6)

where, \( \alpha \) is some attenuation factor. For the left channel response the HRTF matching yields:

\[
\text{HRTF}_L(\omega, \theta_{Head})S(\omega) + \alpha \text{HRTF}_{MF}(\omega, \theta_{Head})V(\omega)
\]

(6.7)

where the interfering sources have an unknown directional component. For each of the separate speakers, a residual factor of the interfering sources could remain. The sequential mapping and overlay of the residual factors could provide an undesirable result that could impact a listeners’ ability to spatially locate the desired sound.

The performance of the array/beamformer array discussed in Appendix A.2.1 shows that the Bartlett beamformer and the microphone arrays have poor performance in focusing onto a look direction and rejecting all other directions.

**6.2 Conclusion of Limitation**

Although the map the microphone array’s response to match the characteristics of a human subject may only require a single microphone, it occurs only under ideal conditions of a single sound source with a known location. For non-ideal conditions, the preprocessing control
algorithm used to locate and separate a given sound source is dependent on the beamforming process, which is dependent on the width and number of microphones in the array.

### 6.3 Future Work

The investigation, processing and quantifying of the microphone array’s response in creating a virtual 3D sound representation has dealt mainly with an ideal acoustic environment, where only one sound source is present. Future study is needed for a non-ideal acoustic environment where multiple sound sources are present. Future studies should focus on the effects residual interfering sound sources have on listeners’ ability to localize sound sources in the binaural HRTF matching. These effects should also be weighed against the preprocessors performance with respect to the number of microphones.
Chapter 7. Conclusion

The aim of this thesis was to map the microphone array’s response to match the characteristics of a human subject. This mapping is necessary because as humans, we have a remarkable ability to locate sound sources with only two sensors, our ears. However, when listening to a response of a microphone array, this ability is lost. This inability to localize sound sources is caused by the array not containing the spatial characteristics of the human body, the HRTFs.

Generating the set of filters needed to map the microphone array’s response to match the characteristics of a human subject was first explored with the limitation of a head-mounted system with a monaural HRTF matching. These systems have the listener immersed within the environment and rely on the acoustic information being relayed to them to hear. The filtering of the microphone array’s response was setup to be independent of the arrival angles of the sound source.

The audience of the listener tests for this thesis was external to the acoustic environment, where the information was stored before it was played to the listener. Also, the aim of this thesis was on microphone arrays, not just a head-mounted system. The tests revealed that monaural HRTF matching did not provide listeners with enough spatial information to precisely localize a sound source. To provide the listener with the desired perceptual awareness, the data-independencies were removed for the HRTF matching. This binaural HRTF matching required the addition of a preprocessor control algorithm.

Though the binaural HRTF matching increased the listener’s ability to locate a sound source azimuth, it suffered from the common issues seen in creating a virtual 3D sound representation in the form of front/back confusion. The analysis revealed listeners were
extremely sensitive to foreign HRTFs, which created the front/back confusion. The addition of simulated head movement reduced source localization error by 20%. It did not eliminate the issues with front/back confusion.

The investigation of the binaural HRTF matching dealt with ideal conditions of having a single sound source with a known location, where the mapping of a microphone array’s response only required a single microphone. For non-ideal conditions, the preprocessing control algorithm used to locate and separate a given sound source is dependent on the beamforming process, which is dependent on the width and number of microphones in the array. Future study is needed for a non-ideal acoustic environment where multiple sound sources are present. Future studies should focus on the effects residual interfering sound sources have on listeners’ ability to localize sound sources for the binaural HRTF matching. These effects should also be weighed against the preprocessors performance with respect to the number of microphones.
Appendix A. Characteristics of Microphone Array

In this Appendix, the microphone arrays that have been used in the testing of this thesis, the head-mounted Headphone Array and the Headrest Array, are explored. The microphone array geometries were chosen to be used on both human subjects and non-human subjects. This was done to explore how the previously discussed mapping techniques of the head-mounted system respond to different array configurations.

Appendix A.1 Construction of the Headphone Array and Headrest Array

The Headphone Array, shown in Figure A.1, was built from a set of Sony MDR V500 Headphones containing 20 WD-61a microphones. Ten microphones were placed in a circular pattern on each headphone cup. The circular pattern and microphone spacing were designed to maximize the number of elements on each headphone cup. The array geometry of the Headphone Array was chosen to be similar in nature to a head-mounted array. All the test data that was measured by the Headphone Array was done with the array mounted on the KEMAR manikin as shown in Figure A.1

Figure A.1) Left to Right, Headphone Array and Sensor Construction, where the microphones receive their power from the DAQ.
The array geometry of the Headrest Array was chosen to not be similar to a head-mounted system. The headrest was also chosen for the possibility of testing in an automobile. The Headrest Array, shown in Figure A.2, was then built from a 2004 Audi A4 Headrest and is a ten element array with three WD-61a microphones placed on the front and back of the Headrest and two WD-61a microphones placed on each of the two sides of the Headrest Array. The individual elements of the Headrest Array are placed on the headrest in a random manner.

The difference in the construction between the sensors in the Headphone Array shown in Figure A.1 compared to the Headrest Array shown in Figure A.2 was done in an attempt to eliminate some or all of the sensor noise found in the WD-61a microphones.

Appendix A.2 Performance of the Headphone and Headrest Array

This Section provides a brief introduction into the most common measurements of Array/Beamformer performance. These performance metrics will evaluate the Bartlett beamformer ability to evaluate the performance characteristics of the Headphone Array and the Headrest Array for a broadband signal.
Appendix A.2.1 Beampattern

A way to illustrate the performance of the array beamformer is through examining the corresponding beampattern. The beampattern is a way to visualize how the beamformer, \( \mathbf{w} \), will pass the gain coming from the steered direction, \( \theta_s \), while suppressing signals coming from all other directions \( \theta \), such that the beampattern describes the sensitivity of the beamformer with respect to signals arriving from different directions [37]. The beampattern is defined as the squared magnitude of the spatial frequency response.

\[
BP(\omega, \theta) = |w^H(\omega, \theta_s)d(\omega, \theta)|^2
\] (A.1)

The beampattern is a function of the beamformer \( w^H(\omega, \theta_s) \) and array steering vector \( d(\omega, \theta) \), which are function of the array geometry and the source position. The beampattern is also a function of the number of sensors and the signal frequencies [7]. The beampattern of the Headphone Array with a DSB steered to \( \theta_s = 90^\circ \), is shown in Figure A.3 for the frequency range of human hearing.

The beampattern in Figure A.3 shows multiple lobes, the lobe corresponding to the highest amplitude is the main-lobe and all the other lobes are called side-lobes. The height of the side-lobes represents the gain patterns for noise and competing sources present along the directions other than the steered direction. In array and beamforming design, the hope is to lower the side-lobes as low as possible so that signals coming from directions other than the steered direction would be attenuated as much as possible [7]. In Figure A.3, the lack of attenuation across the azimuth range for low frequency indicates that the Headphone Array and DSB have poor performance.
Appendix A.3 Gathering Array Data

The basic experimental setup that was used to retrieve the responses of the array is shown in Figure A.4 for the Headphone Array. The experimental data gathered from each of the arrays was conducted in an equivalent manner and is outlined here:

- The Array is placed upon a rotatable platform, where the azimuth angle can be changed from 0 to 360 degrees. The elevation angle was held constant for all tests.
- Seven feet away from the Array, the source speaker is located at the same elevation as the Headphone Array on a manikin. The source speaker is a Mackie MR8 reference speaker.
- Two feet away from the reference speaker, between the source speaker and the Array, is a WD-61a reference microphone.
• The Array data were taken in an anechoic chamber and collected with a NI 9233 at a sampling rate of 50 kHz.
• The sampling rate of 50 kHz was selected instead of the standard 44.1 kHz sampling rate because the rate of 44.1 kHz is not available with the NI 9233.

![Array test setup for Headphone Array.](image)

**Appendix A.3.1 Finding the Acoustic IR**

The array steering vector and the array manifold matrix used throughout this thesis depends on the accuracy of the measurement of the Impulse Response (IR) or transfer function contained within. The IR is the time domain representation of the transfer function, a frequency dependent filter that shows how a sound source is filtered by the diffractive nature of the enclosure and the physical nature of the microphone array’s geometry.

The two standard approaches of finding the IR is by exciting the acoustic source (speaker) with an impulse or with a broadband noise of length longer than the IR. The disadvantage of using an impulse method is that it provides inaccurate reproduction of the phase and has difficulty reproducing nonlinearity aspects of the IR. The IRs used in this thesis were
measured by the broadband noise method. This method requires the addition of a reference microphone to act as the source.

The broadband random noise method of calculating the IRs can be done either in the time domain or the frequency domain. For the frequency domain, the IRs are found by the inverse Fourier Transformation of the transfer function, which is found by taking the measured microphone signals of the array and dividing them by the measured reference signal. The IRs in the time domain are found by cross correlating the measured microphone signal of the array and the measured reference signal. The IR is obtained mainly in the time domain, because the uncorrelated sensor noise will be captured in the frequency domain.

The IRs measurements are susceptible to noise. The most common means of reducing the effect caused by the additional noise is to average the IRs of multiple tests. An equalization process is often used to remove any of the non-ideal responses of the loudspeaker/microphone combination of the enclosure.
Appendix B. Measuring the HRTF

The direct measurement of the subject’s HRTF is done by sitting an individual on a chair and placing measurement microphones in the entrances of the blocked ear canals. The measurements are obtained at set azimuth and elevation intervals to achieve the desired spherical spatial resolution of the HRTF dataset. The measurements are obtained by driving a loudspeaker with a known broadband noise.

The measurement obtained from the HRTFs contains multiple sub-systems that need to be compensated to achieve an unbiased estimate of the HRTFs. The sub-systems are the transfer functions of the driving loudspeaker, the microphone, and the measurement microphones. For the free-field HRTF approach, the input or reference measurements are obtained by placing the measurement microphones at the origin (i.e., center of the head) with the subject being absent. Thus, the influence of the distance between the measurement loudspeaker and the microphone is excluded as well as the influence of the loudspeaker and measurement microphones.

Only Subject 3’s HRTFs were measured. Since only one subject’s HRTFs were measured, an increment rig was not built. Subject 3’s HRTFs were measured by:

- Two microphones were placed in Subject 3’s ears.
- Subject 3 sat in an office chair that was placed on a rotatable platform. Subject 3 was told to face straight ahead and stay as still as possible.
- The azimuth was changed in 5° increments form 0° to 180° (symmetry of the human subject was assumed). The elevation angle was held constant.
- The Mackie MR8 reference speaker played a known broadband random noise signal.
• The reference measurement was obtained by placing the ear microphones at the origin (i.e., center of the head) with the subject being absent.

• Subject 3’s HRTFs were measured in an anechoic chamber and collected with a NI 9233 at a sampling rate of 50 kHz.

Appendix B.1 Rationale for Blocked Ear Canal

Moller’s research investigated whether the ear canal influenced the directional dependence of the HRTFs [45]. Moller concluded that sounds near the entrance of the ear canal captured all the directional dependence and it was not necessary to model the inner ear (canal). Moller’s research also found that by blocking the ear canal, the individual difference between subject’s HRTFs is far lower than when the measurements were obtained with an unblocked ear canal. Moller’s research thesis has been taken as proof in the acoustic world that the ear canal is irrelevant for HRTF modeling.

Appendix B.2 Equalization of HRTFs

Equalization of the HRTF with the HRTF diffuse-field transfer function is a popular approach. Diffuse-field transfer function of the reference HRTF describes the distortion caused by the distortions in the sound field to the eardrum. Thus, by dividing the original HRTF by the diffuse field reference HRTF, the factors that are not incident-angle dependent, such as the ear canal resonance, are removed [46]. The diffuse field reference HRTF is obtained by averaging the power spectrum of all HRTFs for that ear and then returning it back to the average spectrum by taking the square root. This equalization process was carried out on the KEMAR HRTFs [19] and the HRTFs of the ‘Listen Group’ subjects [20], but was not done for the measured HRTF of Subject 3.
References


