Improvements in the Measurement and Optimisation of Head Related Transfer Functions for Binaural Ambisonics

Cal Armstrong

Nov. 2019

PhD

Department of Electronic Engineering
University of York
YO10 5DD

This Page Has
Intentionally Been
Left Blank

Abstract

In recent years, the desire for spatial audio has surged with the inclusion of such technologies within popular streaming platforms and content creation workflows. Often presented over headphones as binaural audio, spatial audio allows a listener to experience a sense of externalisation and realism over and above traditional stereo playback. It is particularly suited to Virtual Reality; head mounted displays are fast becoming an affordable option to present 3 dimensional visual content and it is only logical that coherent accompanying audio should also exist. However, the challenge comes in achieving a life-like auditory image at minimal computational cost.

Two things are needed to deliver high quality binaural audio: accurate measurement of the way in which humans interoperate a soundfield and a rendering engine capable of applying such methods to pre-prepared spatial auditory data. Head Related Transfer Functions (HRTFs), are individual filters that describe the transfer function between a free-field source and the signals that arrive at a listener's ears. Ambisonics, a data storage and audio reproduction format based around the spherical harmonic functions, has become one of the leading approaches to such rendering engines.

This thesis considers the capture and optimization of HRTFs for binaural-based Ambisonics. Spatio-temporal manipulations, a technique referred to as BiRADIAL, are shown to objectively improve the accuracy of binaural output through a perception-based spectral comparison model. A novel approach to HRTF measurement is then presented, capable of synthesising infinite far-field filters from just 50 real-world near-field measurements taken in under 7 seconds. Perceptual listening test results show an equivalence to the more traditional measurement approach despite the savings in time, cost and complexity.

This Page Has
Intentionally Been
Left Blank

Contents

List of Tables

List of Figures

Acknowledgements

To my supervisors,
Gavin Kearney and *Damian Murphy*

My cheerleaders,
Looloo and *Moosey*

The AudioLab

Aaron, me too.

This research was supported by a Huawei Innovation Research Partnership.

This Page Has
Intentionally Been
Left Blank

Author's Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Parts of this thesis have previously been presented at conferences and/or in journals in which I was the primary contributor to the work:

- Simultaneous HRTF Measurement of Multiple Source Configurations Utilizing Semi-Permanent Structural Mounts. Armstrong C., Chadwick A., Thresh L., Murphy D. and Kearney G. in *AES 143rd Convention.* 2017.

- A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database. Armstrong C., Thresh L., Murphy D. and Kearney G. in *Applied Sciences*, vol. 8, no. 11. 2018. DOI: doi.org/10.3390/app8112029.

- A Bi-RADIAL Approach to Ambisonics. Armstrong C., Murphy D. and Kearney G. in *AES International Conference on Audio for Virtual and Augmented Reality.* 2018.

and in which I was an equal contributor (between the 1$^{\text{st}}$ and 2$^{\text{nd}}$ Authors):

- A Perceptual Spectral Difference Model for Binaural Signals. Armstrong C., McKenzie T., Murphy D. and Kearney G. in *AES 145th Convention.* 2018.

- A Numerical Study into Perceptually-Weighted Spectral Differences between Differently-Spaced HRTFs. Young K., Armstrong C., Tew, A., Murphy D. and Kearney G. in *AES Conference on Immersive and Interactive Audio.* 2019.

This Page Has
Intentionally Been
Left Blank

1 Outline

Chapter Overview

This chapter will introduce the thesis, contextualise the work and state the hypothesis and accompanying objectives. A summary of the thesis is given including a non-technical overview of each chapter and identification of novel contributions.

1.1 Introduction

Spatial audio technologies are at the heart of immersive content creation for a wide range of applications from traditional film and television production through to music production and soundscape design. Popular digital audio workstations are ever-increasing their multi-channel capabilities to support spatial audio formats. There is also a growing proliferation of affordable spatial microphone arrays on the market accommodating immersive content creation at a consumer level. Similarly, in game audio, tools such as Google Resonance[1] are facilitating the creation of immersive and interactive audio within game design engines.

At the reproduction phase, spatial audio is usually delivered via a multi-channel loudspeaker array or headphones, the latter of which typically utilises binaural audio rendering, the focus of this thesis. Binaural audio attempts to deliver the perceptual localisation cues inherent in normal real-world listening in an effort to render 3-dimensional (3D) soundfields at the ears of the listener. The localisation cues of a source at a particular angle and distance to the head may be described in their entirety by a stereo filter referred to as a Head Related Transfer Function (HRTF).

HRTFs may be used to virtualise a collection of real loudspeakers such that they may be simulated as if they would be heard in real life, over a pair of headphones. This is known as the reproduction of a *virtual loudspeaker array.* The results can be very convincing if the process is calibrated and performed correctly; however, individualisation of localisation cues and difficulty in rendering accurate spatial soundfields over real reproduction arrays at low computation cost can result in inaccuracies in the rendered signals.

There are several techniques available to derive a set of loudspeaker signals that will reproduce a spatial sound scene. Popular methods include Vector Based Amplitude Panning (VBAP) (Pulkki, 1997), Wave Field Synthesis (WFS) (Berkhout, 1988) and Ambisonics (Gerzon, 1973). VBAP creates phantom sources by outputting source signals from the closest set out loudspeakers to the intended source locations. Alternatively, WFS and Ambisonics in general will attempt to analyse and reproduce the physical properties of the soundfield. Despite the options, Ambisonics has been adopted by the industry due to its flexibility, expandability and independence of its

[1] developers.google.com/resonance-audio

encoding and decoding procedures. Rotation matrices are also easily applied to the data to account for head movement in Virtual Reality (VR) applications. However, accuracy of high frequency reproduction is exponentially linked to an increase in complexity. The cause of these inaccuracies are primarily a result of the spatial separation of a person's ears.

Within a binaural renderer, and unlike in real life, there exists a complete independence of the signals sent to the left and right ears. It is therefore possible to consider optimisations applied within the renderer that are specific and individual to either ear. However, there remains an issue regarding the accurate and feasible capture of an individual's HRTFs. To date, HRTF measurement has been a relatively long and uncomfortable process for a subject. Capturing a high quality and dense set of acoustic measurements can take upwards of an hour depending on the technique used. During this time the subject is generally required to remain stationary and not always within the most comfortable position (Armstrong et al., 2018a). Although it is possible to reduce the time required for a set of measurements, e.g. to a matter of minutes (Masiero, 2012), doing so comes at the cost of a high signal-to-noise ratio and in general the accuracy of the measurements, for example, the precise relative positions at which the measurements are taken. Requiring a subject to remain completely still for a matter of minutes is still a significant challenge. It is therefore appropriate to consider novel methods to substantially reduce the capture time of personal HRTFs even further, for example to a matter of seconds, whilst maintaining a satisfactory level of quality.

The work of this thesis aims to tackle the issue of the reproduction accuracy of binaural Ambisonics by the manipulation of HRTFs used to reproduce virtual loudspeaker arrays. These techniques are applied within a new HRTF capture system to facilitate fast and convenient measurement of a person's HRTFs. The importance of individual measurements are considered alongside ear-specific optimisations. Objective and subjective evaluations throughout the thesis confirm theoretical predictions of increased rendering performance. A final perceptual listening test justifies the new approaches to HRTF measurement and concludes that the optimization techniques discussed within the following chapters positively effect the reproduction of popular musical stimuli.

1.2 Composition of Thesis

1.2.1 Hypothesis

The following hypothesis is considered:

> *Improvements can be made to binaural Ambisonic rendering workflows through the spatio-temporal manipulation of HRTFs within a feasible measurement procedure.*

Improvements reflect a perceptual increase in either the spatial or timbral quality of a system's output.

Binaural Ambisonic rendering refers to the binaural reproduction of an Ambisonic based spatial audio reproduction system.

Spatial manipulation refers to adjusting the position at which the HRTFs are measured.

Temporal manipulation refers to adjusting the time delays of the HRTFs.

A feasible measurement procedure relates to the speed and ease with which HRTFs are captured for an individual.

1.2.2 Objectives

To answer the hypothesis, the following objectives are defined:

1. To compile a comprehensive review of the human auditory system and Ambisonic reproduction technologies.

2. To investigate the performance of spatially and temporally manipulated HRTFs in the context of binaural Ambisonic rendering following a traditional HRTF capture workflow.

3. To deliver a fast and convenient HRTF capture system able to exploit the findings of Objective 2 to deliver optimized individual HRTFs for the end user.

1.2.3 Summary of Work

Following this introductory chapter, Chapter 2 consolidates the concept of binaural audio, HRTFs and the key localisation features of the human auditory system. Differences between anechoic and room responses are discussed and a brief explanation of binaural rendering is given. This chapter lays the foundations for the understanding and analysis of such measurements in the following chapters.

Chapter 3 then presents the SADIIE database, a novel collection of measured binaural Impulse Responsess (IRs) taken from humans and dummy heads in an anechoic and slightly reverberant environment. These measurements were taken in order to evaluate the perceived timbral quality of binaural rendering using both ones own measurements, as well as the measurements of other individuals. To that end it is shown that a person's own measurements are not always optimal. This is shown via a novel listening test in which participants were asked to evaluate the tibrel differences of musical stimuli rendered through individual and non-individual HRTFs.

Following an evaluation of direct HRTF rendering, alternative (and arguably more convenient) rendering techniques are considered. Chapter 4 provides a detailed and comprehensive summary of Ambisonics and the methods with which it may be used within a binaural rendering scenario. State of the art optimisations to the rendering strategy are discussed, namely the time-alignment of loudspeaker feeds, and in a novel conceptualisation it is shown how these optimisations result in a shift to the 'sweet spot' typically located in the center of the reproduction array.

Despite the many advantages, Chapter 5 highlights an important limitation of the Time-Alignment strategy for Ambisonics in the case of rendering far-field planer wavefronts. As a solution, a new rendering approach, referred to as Binaural Rendering of Audio through Duplex Independant Auralised Listening (BiRADIAL), is presented. This novel method optimizes the binaural rendering of Ambisonics via the duplication and repositioning of virtual loudspeakers around either ear. The method is compared directly to the Time-Alignment method.

As a way to compare the spectral output of the two rendering techniques a novel spectral comparison model is developed and evaluated to compare the timbre of two sounds based on the sensitivity and resolution of the human ear. It is shown how this model is able to predict differences in spectra more in line with actual human perception compared to the average difference between the Fast Fourier Transform (FFT) points of two spectra.

It is suggested that the optimizations to binaural Ambisonic rendering may be utilized to improve the HRTF capture process. In particular, the length of measurement and size of measurement rig are considered. In preparation for this, spectral differences of HRTFs are considered at various distances from the head. It is found that perceptual differences within certain frequency bands can exist outwards of 10m and therefore near field compensation should be considered in the synthesis of HRTFs for measurements taken within this radius.

Chapter 6 then covers the development of a Miniaturised Acoustic Response Chamber (MARC), a device for the fast-capture of binaural IRs designed specifically to support the implementation of individualised BiRADIAL rendering. Objective analysis is used to show that the measurements taken in MARC are of similar composition to those from the SADIIE database despite reducing the measurement length from over 1 hour to 7 seconds and the radius of the rig from 1.2m to 0.5m. This is achieved through the individual near field compensation of each loudspeaker element within the measurement array to better approximate the wave-front curvature of far-field sources given the variable radii of loudspeakers.

Chapter 7 brings together the work of this thesis with a final perceptual listening test. Time-Alignment and BiRADIAL rendering methods are compared with individual and non-individual HRTFs measurements from MARC and the SADIIE database. This listening test presents the first timbral comparison of musical sources rendered via binaural Ambisonics with individual and non-individual HRTFs. It is shown that in general neither localisation nor timbral preference was significantly impacted by opting for measurements taken in MARC as opposed to using those from the SADIIE database.

Chapter 8 includes the re-statement of hypothesis and provides a review of chapters, novel contributions and research questions. It identifies possible directions of future work and finally summarises the thesis with final remarks.

1.3 Summary

Binaural audio and virtual loudspeaker rendering has been introduced and Ambisonics has been suggested as a suitable spatial audio playback format. Following this introduction, the purpose of this thesis is then to explore whether or not improvements can be made to this workflow by the spatial and temporal manipulation of HRTFs. Such exploration shall begin by first considering the human auditory system in isolation before going on to optimize the ways in which spatial audio content is delivered.

This Page Has
Intentionally Been
Left Blank

2 Binaural Audio, Spatial Listening and Methods of Analysis

Chapter Overview

This chapter covers the basic principles of binaural audio and the key localisation features of the human auditory system (ITDs, ILDs, Spectral cues). It discusses the differences between HRTFs and BRIRs and briefly explains the concept of a binaural renderer.

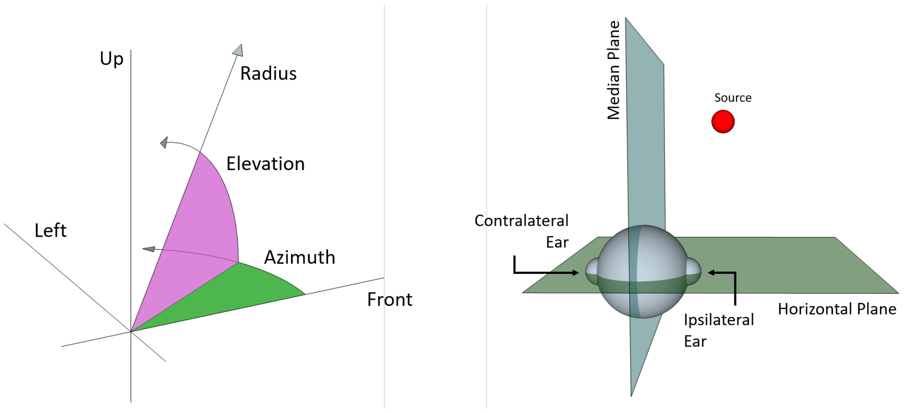
2.1 Introduction

The human auditory system is derived from a pair of spaced dynamic filters (the ears) whose responses are, in part, a function of the direction-of-arrival of a sound source (Blauert, 1997). These filters are known as HRTFs. HRTFs define the transfer function between a localised free-field anechoic source and the signals present at a listener's ear canals (Møller et al., 1995) and are the reason humans can locate a source within 3 dimensions.

Typical features include time of arrival and level differences between the ears as well as spectral colourations caused by the pinnae. Other features may include the likes of torso/body reflections. With these surfaces being further from the ear canal their resulting features are generally seen at lower frequencies than those from the pinnae. Despite being of less perceptual relevance in general, body reflections can still offer significant localisation cues for sources with reduced spectral content or from particular locations (Guldenschuh, Sontacchi and Zotter, 2008). Algazi et al., 2002a shows this to be true in particular for sources deviating from the median plane with no spectral energy above 3kHz. Further to HRTFs are Binaural Room Impulse Responses (BRIRs) which extend the anechoic transfer functions to include a room response (early reflections and reverberation). Collectively these measurements are referred to as binaural filters.

The key principle of binaural audio is the inclusion of these localisation features within a two-channel audio representation. Such material may then be presented to a listener over a pair of headphones in an attempt to recreate exactly (or very closely) the signals that would appear at a person's ears given a particular audio excitation in real life (Møller, 1992). It is a significant enhancement over the likes of stereo reproduction and is a natural accompaniment to virtual and augmented reality. Binaural audio spatialises its content in a realistic way to give the impression of externalised sources. However, in order to effectively work within this field it is important to understand the effects of such localisation features and develop methods with which to analyse and validate the resultant signals.

The reader should note that technically the abbreviation 'HRTF' refers to the frequency domain equivalent of an Head Related Impulse Response (HRIR) (time-domain), i.e. they may both be used to refer to the same measurement depending



(a) Planes

(b) Angles

Figure 2.1: Basic coordinate system where azimuth is measured anticlockwise about the horizontal plane and elevation is measured from the horizontal place (upward positive). Radius is measured outward from the origin. Ipsilateral and contralateral ears are defined by the placement of a source either side of the median plane.

on context. However, the term 'HRTF' has become more commonplace in industry when referring to these anechoic measurements in general. Whilst 'HRTF' will therefore be used in the general case, 'HRIR' will used when specifically referring to the time domain.

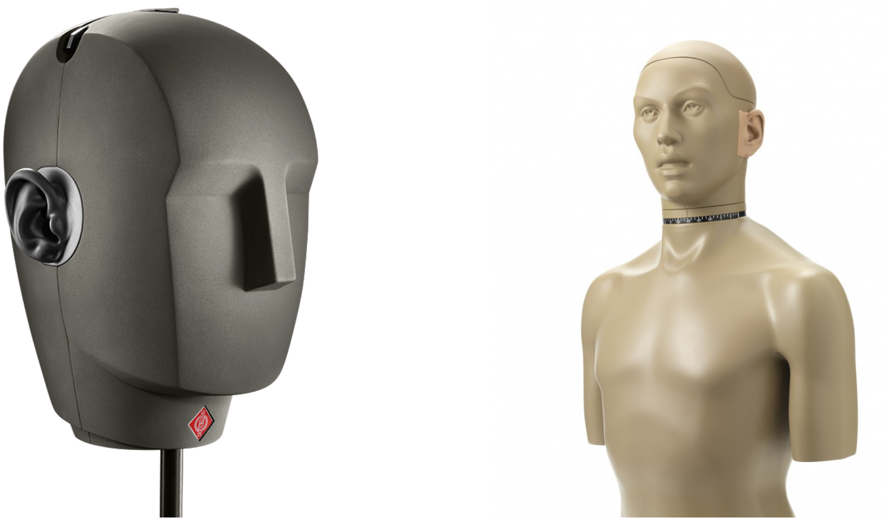
2.2 Coordinate System

The coordinate system used throughout this thesis is defined in Fig. 2.1a. It is based on a standard spherical coordinate system. Azimuth is measured anticlockwise about the horizontal plane. Elevation is measured from the horizontal plane such that the upward direction is positive. Radius is measured outward from the origin.

Fig. 2.1b describes 2 important planes within the specified geometry. The horizontal plane sits parallel to the floor and intersects the ears. It is used to describe 0° elevation. The median plane sits perpendicular to the horizontal plane and bisects the head front to back. It may be used to differentiate sides of the head. The terms ipsilateral and contralateral are used to mean same-side and opposite-side respectively. They are particularly useful to describe either ear independently based on the location of a source.

2.3 Binaural Dummy Heads/Mannequins

It is also important to mention here what is meant by a binaural dummy head/-mannequin (sometimes referred to as a Head And Torso Simulator (HATS)). These



(a) KU100 (Image from **neumann.com**) **(b)** KEMAR (Image from **gras.us**)

Figure 2.2: Examples of two commonly used binaural dummy heads/manakins

devices are commonplace within binaural and spatial audio research and are referred to often throughout this thesis. They are a type of anthropomorphic microphone, the development which over the years is presented by Paul, 2009. In simple terms, they consist of a model of a human head, and sometimes the shoulders/torso, with either inbuilt microphones or the space to place microphones within the ear canals.

Two very common production models are shown in Fig. 2.2. They are the KU100 dummy head[2] and KEMAR mannequin[3]. Each have built in microphones that sit within the head. Whilst the KU100 has a more abstract design, the KEMAR is designed with real median human measurements (Burkhard and Sachs, 1975).

The HATSs are capable of recording sound (in stereo, one channel for each ear) in a very similar way in which humans hear and therefore capture very similar auditory features, as discussed in Chapter 2.4. They are therefore useful for taking reliable and repeatable reference measurements for related research and, unlike fidgeting humans, the quality of their measurements do not deteriorate with lengthy measurement procedures. As such, the filter examples given throughout this chapter will be of measurements taken of a mannequin, not a human.

[2] en-de.neumann.com/ku-100

[3] gras.us/products/head-torso-simulators-kemar

2.4 Binaural Features

2.4.1 Overview

Human auditory localisation is based on three key features: Interaural Time Difference (ITD), Interaural Level Difference (ILD) and spectral peaks/notches. ITDs and ILDs are, in general, a function of the angle of a source from the median plane. They relate to the time and level difference of a source arriving at either ear respectively. Spectral peaks/notches on the other hand vary with both azimuthal and elevatory position and also depend on the shape of a person's pinnae. These features are individual to each person and each ear and characterise the way in which a person percieves spatial audio.

2.4.2 Duplex Theory and Frequency Biases

Particular localisation features have been shown to be dominant/exist only within certain frequency ranges. There is therefore a bias towards particular features depending on the frequency content of a signal. This was investigated by Lord Rayleigh as far back as 1907 and is known as Duplex theory (Strutt (Lord Rayleigh), 1907; Macpherson and Middlebrooks, 2002). It states that ITDs dominate at low frequencies ($<$ a few thousand Hz) whilst ILDs dominate above that. This is partially due to the diffraction of low frequency sound waves resulting in level differences between the ears hardly existing below 1000Hz and partially due to the issue of phase ambiguity.

Phase Ambiguity is the point at which, for periodic waveforms, the relationship between the phase difference between the ears and the time delay between the signals is no longer clear. This could be due to repeated wave cycles between the ears or indeed an inability to distinguish between the leading and lagging ear (Brainard, Knudsen and Esterly, 1992). The frequency at which this occurs is a function of ITD and therefore depends on angle. Ambiguities start from the frequency at which the period is twice that of the ITD or, alternatively, a path difference between the ears of a half wavelength. For example, a lateral source that presents an ITD of around 0.6ms will begin to encounter phase ambiguity at

$$f = \frac{1}{2T} = \frac{1}{2 \times 0.6\text{mS}} \tag{2.1}$$
$$= 833\text{Hz}$$

where f is the frequency, and T is the period. However, as the source approaches the median plane the ITD reduces and hence the maximum detectable frequency increases.

Further to ITDs and ILDs, pinna responses are generally only seen at frequencies $>$3-4kHz (Takemoto et al., 2012; Geronazzo, Spagnol and Avanzini, 2018; Algazi et al., 2002b) due to the physical size of the ears and proximity of the folds to the ear canal. On the other hand, shoulder and torso reflections have been identified as low as 700Hz and have been thought to indicate a sense of elevation (Algazi, Avendano and Duda, 2001).

2.4.3 Interaural Time Difference

ITD is defined as the time delay between a signal arriving at either ear as a result of varying path length. Fig. 2.3 shows a way to approximate these path lengths using a spherical head model and simple geometry. This model is developed by Aaronson and Hartmann, 2014 and is an expansion of the model by Woodworth, 1938.

α is defined as the 3D solid angle between the source and the median plane. Due to the normal placement of the ears with respect to and away from the median plane the model may be simplified to consider it as 2D. The path length may be found for the ipsilateral ear with advanced trigonometry

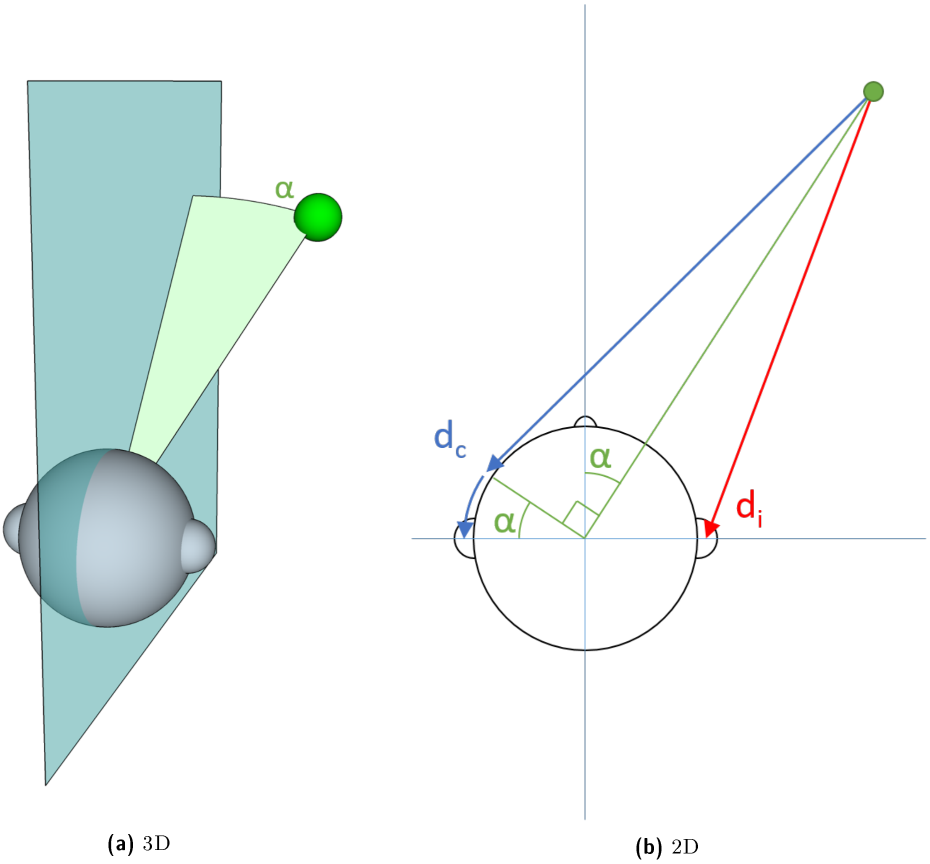
$$d_i = \sqrt{r_s^2 + r_h^2 + 2 \cdot r_s \cdot r_h \cdot \cos\left(\frac{\pi}{2} - \alpha\right)} \tag{2.2}$$

and for the contralateral ear with Pythagoras and equations for the arc length of a circle

$$d_c = \sqrt{r_s^2 + r_h^2} + r_h \cdot \alpha \tag{2.3}$$

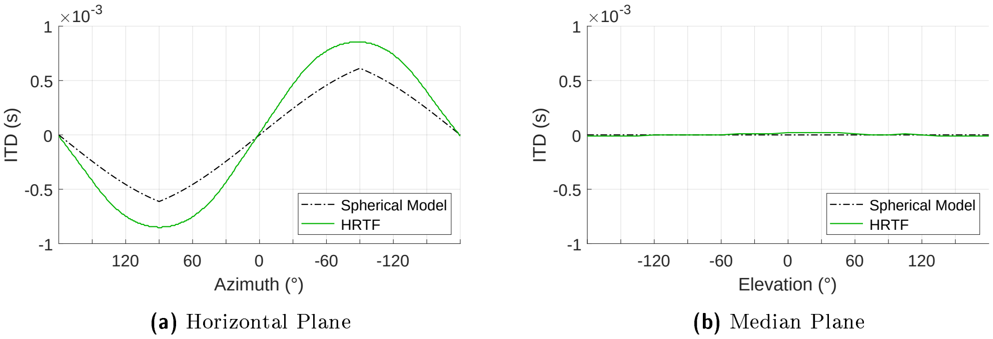
where:

- r_s is the radius of the source
- r_h is the radius of the head
- α is the solid angle between the source and the median plane in radians



(a) 3D **(b)** 2D

Figure 2.3: Considering a 3D source within a simplified 2-dimensional (2D) spherical head model for approximating ITD.



(a) Horizontal Plane **(b)** Median Plane

Figure 2.4: Example of the ITDs on the Horizontal (left) and Median (right) plane for the KU100 dummy head and a source at 1.2m radius (data from the SADIIE database, see Section 3.3). A positive value indicates that the sound has reached the right ear before the left.

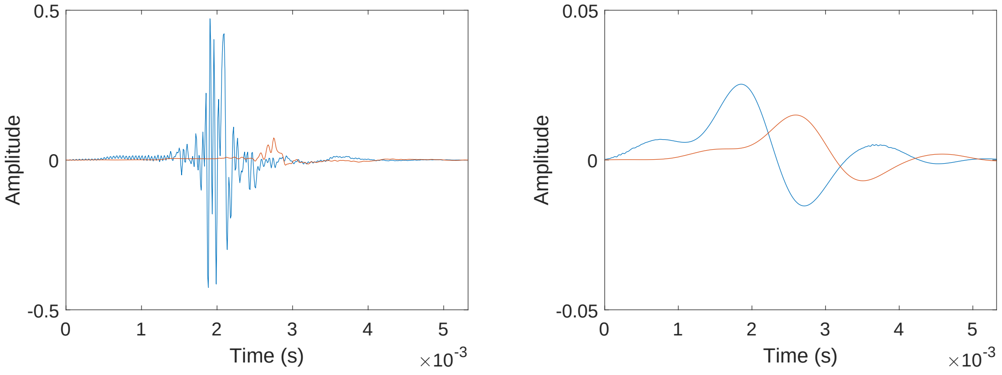
Fig. 2.4 plots the ITD for a real source moving around the horizontal axis of a KU100 dummy head at a radius of 1.2m. The output of the spherical head model is also shown for comparison. It can be seen that the model in general underestimates the ITD. This is due to nature of the Woodworth technique assuming the propagation of a high frequency source around a rigid sphere (Aaronson and Hartmann, 2014). Although alternative models incorporating diffraction characteristics are available, e.g. (Kuhn, 1977), they are inherently more complicated and often unnecessary for simple applications.

Whilst a relatively simple concept, ITD it is notoriously difficult metric to estimate from actual binaural IRs. This is partly because the theoretically clean time-domain waveform filters are distorted by the spectral cues of the pinnaes and so determining the exact time-of arrival of an impulsive signal within the left/right channels is almost impossible. Further, there may be frequency dependencies that vary depending on the linearity of the filters. Katz and Noisternig, 2014 undertook a comprehensive review of some of the most common methods of ITD calculation in 2014 but came up with no definitive findings as to the most accurate method. The methods tested included those from 3 families of measurements: comparison of time-of-arrival via onset threshold detection, calculation of the Maximum InterAural Cross-Correlation (MaxIACC) and group delay estimations of the entire filter. The work presented here implements a method based on the MaxIACC of the left and right channels as used in (Macpherson and Middlebrooks, 2002) and (Kistler and Wightman, 1992) and as documented in ISO standard 3382-1 (ISO, 2009).

It has been shown through Duplex theory that ITD cues are dominant only at low frequencies. It is therefore appropriate to optimise our calculation of ITD for this frequency band. This is done by imposing a minimum order (332 tap) FIR low pass filter with passband = 400Hz, passband ripple <0.1dB and -60dB stopband = 1250Hz. These values were empirically chosen to effectively remove unwanted and irrelevant high frequency noise from the signals. The effect of filtering a HRTF in this way is shown in Fig. 2.5.

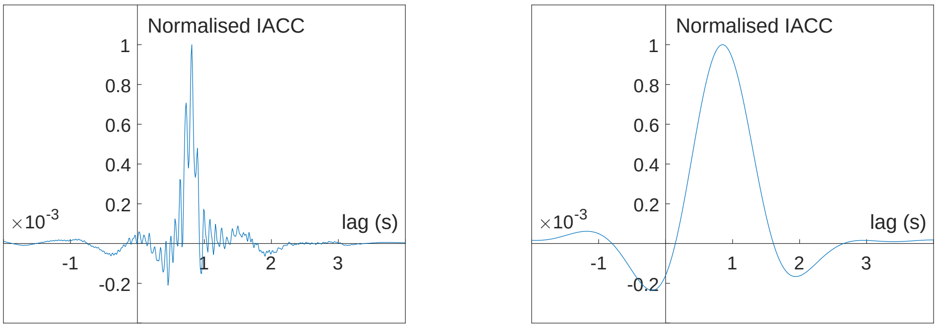
The cross-correlation, described as

$$R_{xy}[k] = \sum_{m=-\infty}^{m=\infty} x[m] \cdot y[m-k] \tag{2.4}$$



(a) Original **(b)** Low Pass Filtered

Figure 2.5: Example of an HRTF azimuth = 90°, elevation = 0° before and after low pass filtering. Note how the high frequency noise has been removed a more obvious time delay between the waveforms is visible. Blue: Left Signal. Orange: Right Signal.



(a) Original **(b)** Low Pass Filtered

Figure 2.6: Example of the InterAural Cross-Correlation (IACC) an HRTF azimuth = 90°, elevation = 0° with and without low pass filtering. Note the single clean and easily identifiable peak in the resulting waveform.

for discrete signals, is then calculated between the left and right channels of the binaural filter and the time delay (or lag) at which $R_{xy}[k]$ is maximised is defined as the ITD. As a precaution, a maximum delay value of 1.1ms is assured (it is expected that all ITDs should fall easily within the range ±1ms) to prevent unexpected correlations from occurring outside of the normal range and overestimating the ITD.

Examples of the output of the cross-correlation function for a raw and low-pass filtered HRTF are shown in Fig. 2.6. The low pass filtering has smoothed the InterAural Cross-Correlation (IACC) and hence improved the robustness of the calculation by eliminating false maxima in the result. Consequently, ITD curves plotted for angles about the horizontal plane are smoother and more reliable.

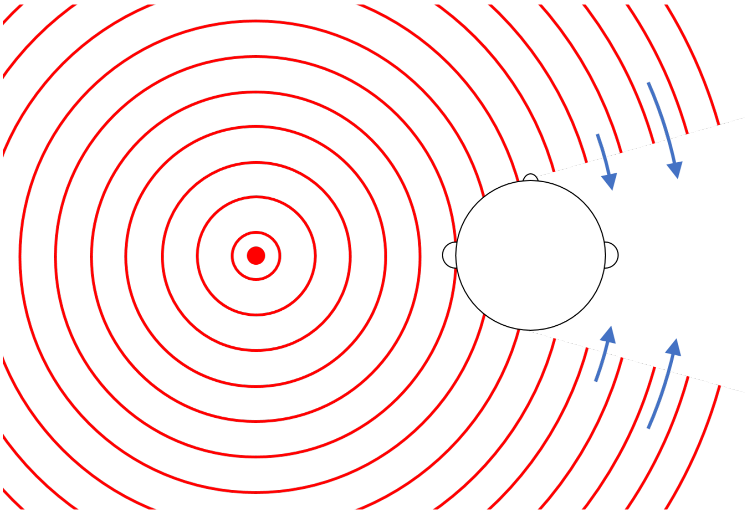
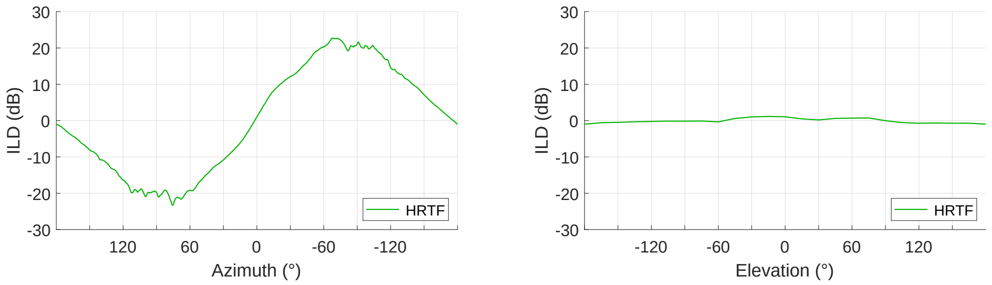


Figure 2.7: Shadowing of a source (with wavefronts shown in red) from the contralateral ear by the head. Blue lines represent the frequency dependent diffraction that will occur resulting in a frequency dependant ILD.



(a) Horizontal Plane

(b) Median Plane

Figure 2.8: Example of the ILDs on the Horizontal (left) and Median (right) plane for the KU100 dummy head and a source at 1.2m radius (data from the SADIIE database, see Section 3.3). A positive value indicates that the sound is louder in the right ear compared to the left.

2.4.4 Interaural Level Difference

ILD (typically quoted in dB) is defined as the level difference between either ear. In addition to the path difference, it is a result of the shadowing of the source by the head/torso as well as pinae, as demonstrated in Fig. 2.7. Due to frequency dependant diffraction around the head, the amount of shadowing will differ depending on the frequency content of the source. Further, individual anthropomorphic features all contribute to the overall perceived level difference at a particular frequency for an individual (Watanabe et al., 2016; Usher and Martens, 2007). Spherical models have been used to show that below approximately 400Hz the difference in level is negligible (<1dB) between the ears (Bernschütz, 2013).

Despite this, similar to ITDs, it is common and convenient to quote just a single numerical figure for ILD. To that end, Fig. 2.8 plots the ILD for a real source moving around the horizontal axis of a KU100 dummy head at a radius of 1.2m.

Analysis of ILDs can be challenging as the concept of 'level' is not well defined. It can be measured in a number of different ways from either the time domain waveform (e.g. RMS of the signal) or spectral response (e.g. average frequency bin weighting). Further, the result is highly frequency dependant. Amplitude differences in the low frequencies can be as low as 0dB (Bernschütz, 2013), whilst in the high frequencies can exceed 30dB, as shown in fig. 2.13c. That being said, there exist methods (detailed in the remainder of this section) for approximating and averaging the results over many frequency bands.

First it is necessary to consider what is implied by ILD. The metric is used to represent the general acoustic shadowing of a person's head/body from a contralateral source. It is therefore not fair to consider a single large spectral notch in the frequency response of one pinnae to constitute an exceptionally large ILD at that precise frequency. This would instead be a result of the pinnae folds, and not the acoustic shadowing of the body. Therefore it is necessary to consider averaging the ILD across the frequency spectrum. This is convenient as, again, it is helpful to consider a single representative value for the ILD across the entire spectrum.

Methods such as those proposed by Watanabe et al., 2016 suggest averaging the spectrum in 1/3 octave bands, or more generally based on a logarithmic scale. Mckenzie, Murphy and Kearney, 2019 adapt this solution to calculate a single value as the mean ILD calculated across 30 frequency bins of Equivalent Rectangular Bandwidth (ERB). The ERB is a measure of a person's ability to discriminate between nearby frequencies (Moore and Glasberg, 1995). It is the equivalent bandwidth of an auditory filter and is closely related to the critical bandwidth. It is dependant on center frequency and widens with an increase in frequency.

ERB may be considered as a measure of the perceptual masking of neighbouring FFT samples in human auditory perception. Consider 2 pairs of neighbouring frequency bins, the first pair centred either side of 100Hz and the second pair centred either side of 10kHz. It is quite possible that each sample in the first pair would sit independently within it's own critical band. Any discrepancies noted within either bin would therefore be of independent and equal relevance. However, due to the increasing bandwidth of auditory filters at high frequency it is likely that the frequency bins centred around 10kHz would fall within the same critical band. As the frequencies could not be perceptually discriminated it would be unfair to weight each one with equal relevance to the low frequency bins.

It is therefore proposed that ILD may be calculated by taking the mean value of a selection of frequency bins within an FFT representative of the frequencies between 1.5 and 20kHz but weighting the contribution to the average of each bin by its inverse ERB. In this sense, critical bands are being weighted with equal relevance, not FFT samples. The values 1.5 and 20kHz are chosen to correspond with Duplex theory (Strutt (Lord Rayleigh), 1907) and the limits of human hearing.

ILD is thus calculated

$$\mathrm{ILD} = \frac{\sum_{n_{\min}}^{n_{\max}} F_n^R \cdot \frac{1}{ERB_{fn}}}{\sum_{n_{\min}}^{n_{\max}} \frac{1}{ERB_{fn}}} - \frac{\sum_{n_{\min}}^{n_{\max}} F_n^L \cdot \frac{1}{ERB_{fn}}}{\sum_{n_{\min}}^{n_{\max}} \frac{1}{ERB_{fn}}} \tag{2.5}$$

$$n_{\min} = \mathrm{round}\left(\mathrm{nfft} * \frac{1500}{fs}\right)$$

$$n_{\max} = \mathrm{round}\left(\mathrm{nfft} * \frac{20,000}{fs}\right)$$

where:

- F^R and F^L are the left and right frequency spectra respectively in dB,
- nfft is the number of linear sample points in the FFT,
- f is the frequency vector that corresponds to the nfft sampling points,
- ERB is the ERB of each frequency sample defined as (Moore and Glasberg, 1995)

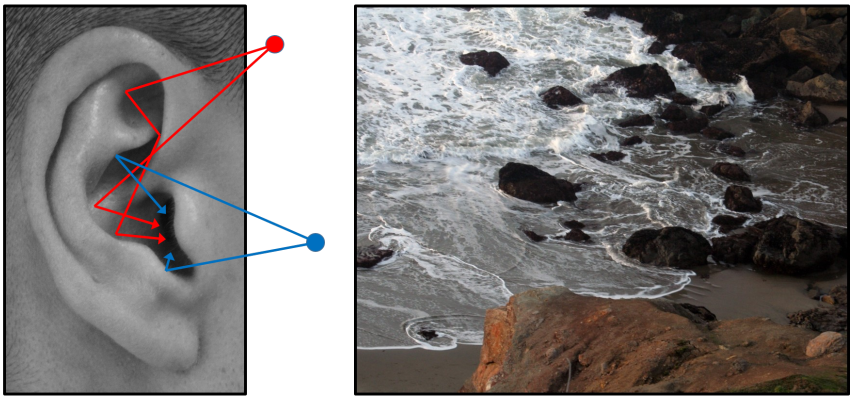
  $$\mathrm{ERB} = 0.108 \cdot f + 24.7 \tag{2.6}$$

  where:
  - f is the frequency.

2.4.5 Spectral Cues

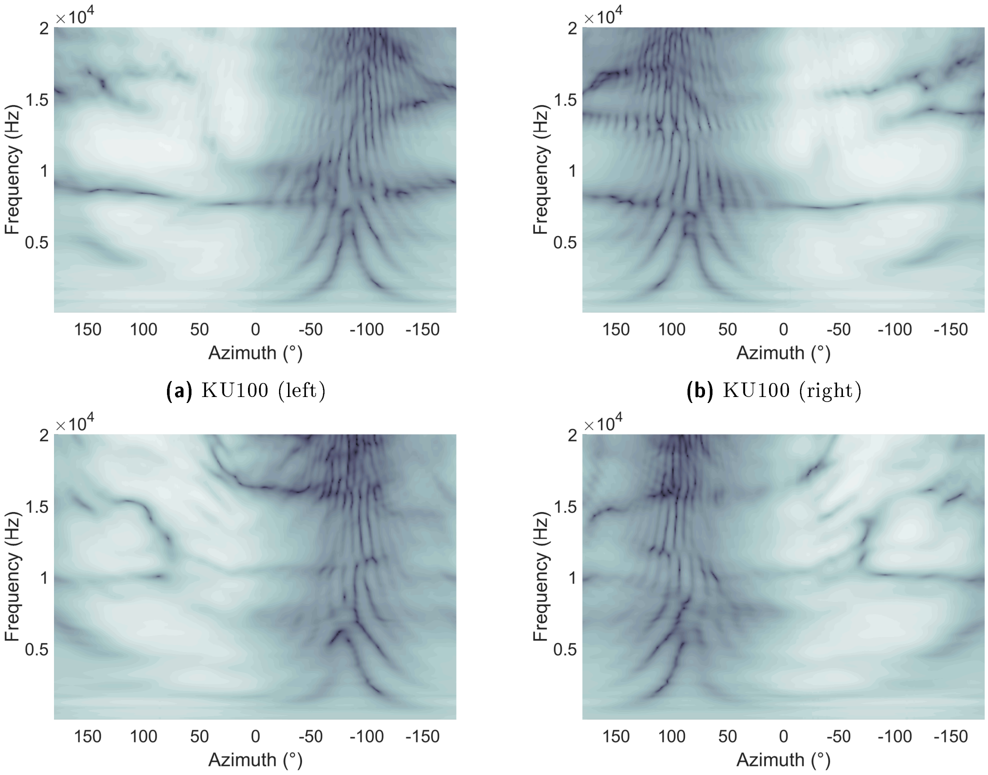
The spectral response of either ear is a direct result of the unique shapes of a person's pinnae folds. Variation in the location of a sound source results in unique patterns of reflections that are sampled at the ear canal. Consider the similarities of ocean waves refracting and superimposing around the shallow rocks in Fig. 2.9.

Specific examples of these frequency responses are provided and discussed in detail in Section 2.7. However, Fig. 2.10 demonstrates how the shape of these responses change for the left and right ears of two different subjects for a source moving about



(a) (b)

Figure 2.9: a) An example of the different reflection paths of multiple sources interacting with pinnae folds analogous to b) waves entering shallow rocks. Photo credit: Sharon Mollerus.



(a) KU100 (left) **(b)** KU100 (right)

(c) KEMAR (left) **(d)** KEMAR (right)

Figure 2.10: Example of the spectral responses of the left and right ears of the KU100 and KEMAR mannequins for a source panning around the horizontal axis (data from the SADIIE database, see Section 3.3). Amplitude is shown by colour. Black: -60dB, White: +20dB. Note the overall reduction in amplitude on the contralateral ears as a result of head shadowing.

the horizontal plane. Note the individual variation in the placement and shape of the spectral notches over and above the general amplitude differences between a source in the left/right hemisphere. Although is it possible for one person to learn the response of another person's pinnae (Hofman, Van Riswick and Van Opstal, 1998; Stitt, Picinali and Katz, 2019), it is these individualised features that are the biggest driving force behind individualised binauralisation.

Pinnae responses are complex and are not easily modelled without advanced image capture and simulation technologies such as (Genelec, 2017). However, significant work is ongoing in the field of Boundry Element Method (BEM) simulations for use with high resolution human/dummy scans for this specific purpose (Young, Kearney and Tew, 2018a; Jin et al., 2014). These techniques are not within the scope of this thesis, however, the reader is directed to the following for further reading in this area (Katz, 2001a; Katz, 2001b; Kreuzer, Majdak and Chen, 2009; Gumerov et al., 2010).

Another method of objective evaluation is to directly compare the spectra of two stimuli. Such analysis may identify differences in the spectral cues of the binuaral filters and indeed indicate the level of perceived timbral difference. It also inherently captures the frequency specific ILDs. As the spectral response of a HRTF does not change with time, calculating the differences with a single FFT is sufficient.

One method of comparison is to directly calculate the spectral difference (in dB across a number of frequency bands) between the spectral responses of two signals as in (Spagnol, 2015; Otani, Hirahara and Ise, 2009). This is referred to as the Absolute Spectral Difference (ASD). Some more advanced methods of ASD are given by Lee and Lee, 2011 who compares a selection of spectral distance algorithms in the context of perceptual differences in HRTF interpolation techniques. However, human auditory perception differs greatly in sensitivity depending on relative amplitude, frequency and temporal aspects (Yost, 2000). In trying to compare generic frequency response spectra destined for human listening therefore, the perceptual relevance of these differences must be considered. This is the motivation behind the Perceptual Spectral Difference Model (PSDM) presented later within the evaluation stages of this thesis in Section 5.7.

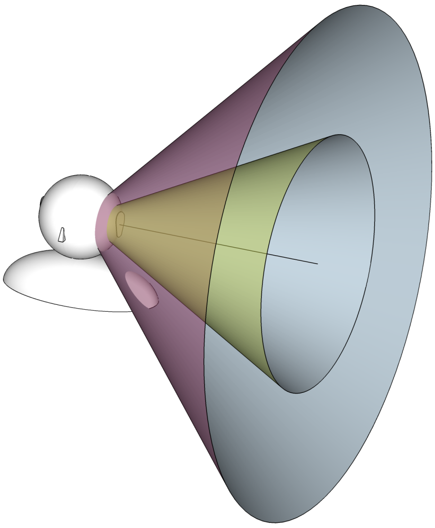


Figure 2.11: Example of 2 cones of confusion, one highlighted in purple, one highlighted in yellow. Along each outer ring ITDs and ILDs are approximately similar and localisation is determined by spectral changes only. There are an infinite number of Cones of Confusion at any angle from the interaural axis on either side of the head.

2.5 Localisation

2.5.1 Overview

Combining the features discussed in Section 2.4 allows a person to localize a sound to a point in 3D space. The ability with which this is managed is dependant on the actual location of the source, the frequency content of the source and a number of other points of confusion. These factors are discussed within this section.

2.5.2 Cones of Confusion

Cones of confusion are approximate rings of angles, as shown in Fig. 2.11, that result in similar ITDs and ILDs being perceived at the ears. They are generally defined as having equal radii and solid angle subtended between the source and median plane from the origin. It should be noted though that this is not an exact definition and differences in ITDs and ILDs do still exist in these paths due to asymmetries of the head (Searle et al., 1975; Middlebrooks, Makous and Green, 1989). That being said, the similarities in time and level differences can put a greater emphasis on spectral based localisation cues and as such perceptual placement of a source can be subject to confusion, most recently shown by Rudzki et al., 2019. Consider a source that is unfamiliar to a listener. It is impossible to tell whether the perceived frequency response is a result of the source or the pinnae folds. The most obvious example of this are sources that lie upon the median plane.

Further to the more general cones of confusion are the more specific cases of front-back confusion. This is where a subject is unable to determine in which hemisphere (front or rear) a source is located despite being able to determine both its angle from the median plane and elevation. The locations are mirror images of each other and belong to the same cone of confusion. Cases of front-back confusion are commonly discussed within the literature (Katz and Parseihian, 2012; Watanabe et al., 2016; Wightman and Kistler, 1989b; Wenzel et al., 1993).

2.5.3 Head movement

Considering the common causes of confusion, it is logical for head movement to aid in the accurate localisation of a source (Begault et al., 2000). As a listener moves their head they shift the relative location of sources around and between respective cones of confusion. By doing this, they give their brain the opportunity to observe dynamic localisation cues and resolve confusions. This has been confirmed by several studies, focused on resolving front/back confusions (Perrett and Noble, 1997; Iwaya, Suzuki and Kimura, 2003; Wightman and Kistler, 1999) as well as elevatory confusions (Kato et al., 2003).

The timbre of the source may be assumed to stay relatively stable over time and listeners can therefore attribute changes in the frequency response/ITDs/ILDs of the source to shifting localisation features. As such, it is very important to ensure that binaural rendering systems are capable of considering low latency head tracked applications.

2.5.4 Externalisation

Further to the angular localisation of a source, externalisation must also be considered. Externalisation is the impression that a source exists outside of the head and at a distance further than the headphone transducers. It has been shown in the literature to improve with the inclusion of early reflections and room reverberation (Begault, 1992; Durlach et al., 1992). This was studied extensively by Jot, 1999 in the late 90's who presented 'Spat', real-time spatial sound processing software, to provide better control over a sounds interaction with a virtual space. Further, Kearney et al., 2012 has shown that a sense of source distance can be achieved despite errors in the spectral reproduction providing there is an adequate direct to reverberant ratio. This is supported by studies from Møller and Sørensen, 1996 and Begault et al., 2000 that show no significant changes in perceived externalisation when sources are rendered with individual/non-individual HRTFs.

Typically, binauralisation of a source with only the direct HRTF (as discussed in Section 2.6) will therefore struggle to fully externalise that source. Durlach et al., 1992 summarizes a number of studies to confirm this and goes on to highlight the importance of correctly adjusting a source with respect to a person's head movement to improve externalisation. Pike, Melchior and Tew, 2016 went on to consider these effects within the context of different binaural renders and found significant differences in the experience of externalisation when using HRTFs but not in the case of BRIRs.

2.6 Binaural Rendering

2.6.1 Overview

Binaural rendering is the process of applying binaural filters to an audio source such that it may be perceived by a listener in a similar (and ideally the same) way as it would be in real life. It is a relatively simple process that may be achieved with rudimentary digital signal processing techniques. In its most basic form rendering of a source is generally restricted to the angles at which HRTF or BRIR measurements have been taken. Although interpolation methods may be used to approximate the transfer functions in between measurements these are prone to errors due to the complex and dense shifting of spectral features. Two rendering methods are therefore described, the first being a direct convolution technique suitable for static sources and the second being a virtual loudspeaker approach suitable for dynamic sources.

2.6.2 Direct Convolution

A source may be binaurally rendered using any binaural filter (Pike, Melchior and Tew, 2016; Treviño et al., 2011; Smyth and Smyth, 2016; Noisternig et al., 2003b) (e.g. a HRTF or BRIR). The transfer function (binaural filter), h, is applied to the signal, s, by means of digital convolution (Vorländer, 2008)

$$y = s * h \tag{2.7}$$

such that the output, $y(n)$, may be written

$$y(n) = \sum_{k=0}^{N-1} s(k) \cdot h(n-k) \tag{2.8}$$

By convolving a monophonic signal with an HRTF or BRIR and presenting the result directly to a listener's ears (usually via headphones) the source is simulated as if coming from the direction in which the IR was measured. When performed correctly using calibrated individualised measurements results can be indistinguishable from a real source (Langendijk and Bronkhorst, 2000).

If individual measurements are not available, approximations or general measurements may be made/taken. This may be appropriate if a single output signal should be rendered for a wide audience. However, such generic measurements can pose complex errors in the rendered signals, discussed further throughout Chapter 3 and in particular in Sections 3.2 and 3.4.2. That being said, it is a popular approach and an example of this implementation is in the rendering stages of Google Resonance[4] where measurements of a dummy head are used (Gorzel et al., 2019).

2.6.3 Virtual Loudspeakers

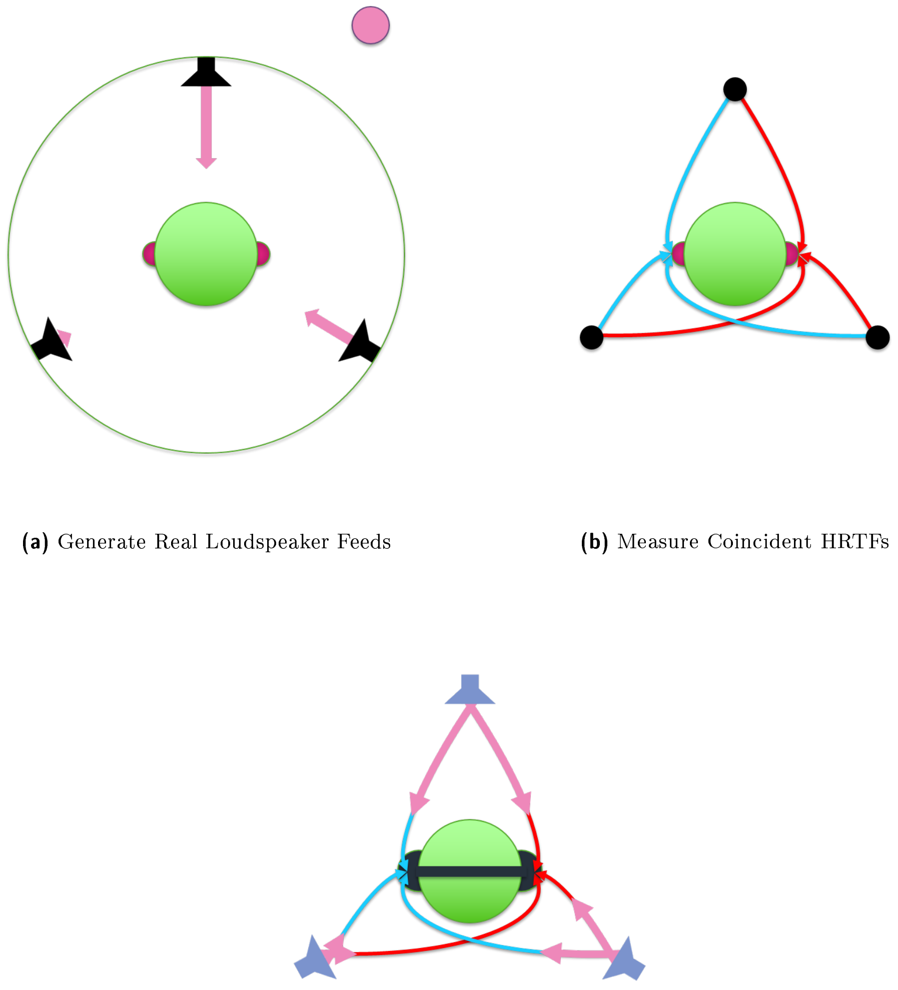
A virtual loudspeaker approach combines well defined methods for rendering a source at any location over a real loudspeaker array, with the ability to simply and easily binaurally render a static source with direct convolution. The workflow is described in Fig. 2.12. It may be expressed mathematically as

$$\sum_{l=1}^{L} \rho_l * h_l \tag{2.9}$$

for each stereo channel where:

- L is the number of loudspeakers,
- ρ are the loudspeaker signals,
- h_l is the binaural filter measured from the position of that loudspeaker.

[4] resonance-audio.github.io/resonance-audio/



(a) Generate Real Loudspeaker Feeds

(b) Measure Coincident HRTFs

(c) Convolve and Sum

Figure 2.12: Basic stages of a virtual loudspeaker binaural reproduction: (a) Loudspeaker signals are generated using standard real world decoding techniques; (b) binaural IRs are measured from the positions of the loudspeakers; (c) the loudspeaker signals are convolved with each binaural IRs and the resulting signals are summed for each ear.

A set of loudspeaker feeds are first generated using a standard loudspeaker based reproduction technique. The individual loudspeaker feeds are then each convolved with a binaural filter measured from the location of the loudspeaker and the results are summed together separately for each stereo channel. By taking this approach, the HRTFs (/BRIRs) required to render a source at any angle remain the same and the responsibility of spatially panning a source between multiple sample points is placed on the loudspeaker renderer, hence avoiding the need for any complex interpolation algorithm.

Early implementations of this technique include those by Mckeag and McGrath, 1996 and Noisternig et al., 2003a; Noisternig et al., 2003b which utilize the method to binaurally render Ambisonic signals (Gerzon, 1980), and are discussed further in Chapter 4. Whilst McKeag and McGrath consider optimizations to the method such as assuming symmetry of the head to reduce the number of convolutions required, Noisternig considers the more direct benefits over high numbers of computationally expensive time-varying HRTFs for multiple sources. Of course, alternative loudspeaker reproduction techniques such as VBAP or WFS could equally be used, although the benefits of Ambisonic reproduction are discussed further in Chapter 4.

2.7 Binaural Filter Examples

2.7.1 Overview

Having covered the basic features of binaural audio, real world examples of these measurements may now be explored. To that end the acquisition of such measurements is briefly detailed before going on to show examples of and the differences between HRTFs, BRIRs and rendered IRs.

2.7.2 Acquisition

Methods to measure binaural filters are well established in the literature (Stern, Brown and Wang, 2005; Møller, 1992; Rumsey, 2014). They are most commonly measured as one would measure a typical IR of a room. For example, using a swept sine technique from a set of static sound sources situated about a subject's head (Farina, 2000; Meng et al., 2008; Majdak, Balazs and Laback, 2007) and a stereo pair of microphones placed within a subject's ears. This would typically be followed by a significant amount of post-processing to trim and window the measurements and account for any frequency response of the measurement system. A detailed workflow is given in Chapter 3. Alternative methods have been explored via reciprocity (Zotkin et al., 2006) (loudspeakers within the ear canal and subject surrounded by microphones), however, these methods suffer significantly within respect to the low frequency response and Signal to Noise Ratio (SNR) due to restrictions in the size and sensitivity of the micro-loudspeakers. Alternative methods for the capture of the IR itself are also available, i.e. using pseudo-random white noise (Maximum Length Sequence (MLS) (Schroeder, 1979) and IRS (Dunn and Hawksford, 1993)) or a time-stretched pulse (Aoshima, 1981). However, the swept sine technique has proved popular due to its resistance to impulsive background noise and separation of harmonic distortions.

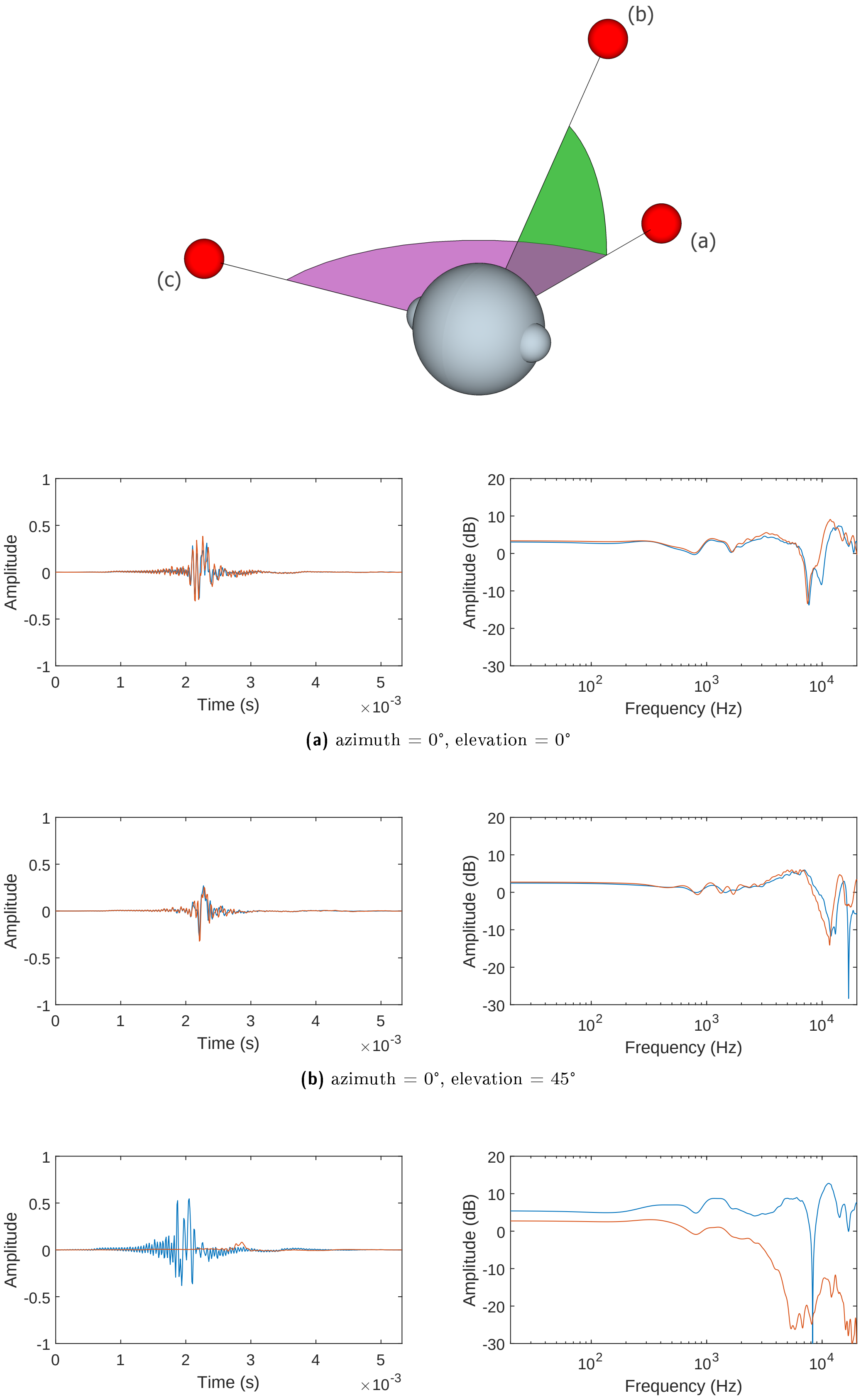
2.7.3 HRTFs

As an example of the type of responses seen from HRTF measurements, 3 source locations are plotted in the time and frequency domain in Fig. 2.13. From these plots varying interaural time and level differences may be identified as well as changes in the spectral responses.

The first response, Fig. 2.13a, is that of a source panned directly in front of the listener (azi = 0°, ele = 0°). Similar temporal positioning of the left/right peaks in the time-domain plot may be observed. The spectral responses are also quite similar, resulting from symmetrical source positioning respective of either ear. The results show a lack of any ITD or ILD. Similarities between the left and right channels are exaggerated in this example given the symmetrical nature of the KU100 dummy head. It is common to see some differences in human measurements due to the individual nature of peoples' ears.

The second response, Fig. 2.13b, is from a source panned in front of the listener at an elevation of 45° (azi = 0°, ele = 45°). Again, the responses of the left and right channels are almost identical due to the symmetrical positioning of the source and there is very little to no change in the ITD or ILD. However, when compared to Fig. 2.13a differences may be seen in the high frequency detail of the spectral responses.

The third response, Fig. 2.13c, is from a source panned directly to the left of the listener (azi = 90°, ele = 0°). As a result, significant differences are seen in the responses both with respect to the previous measurements and between the individual

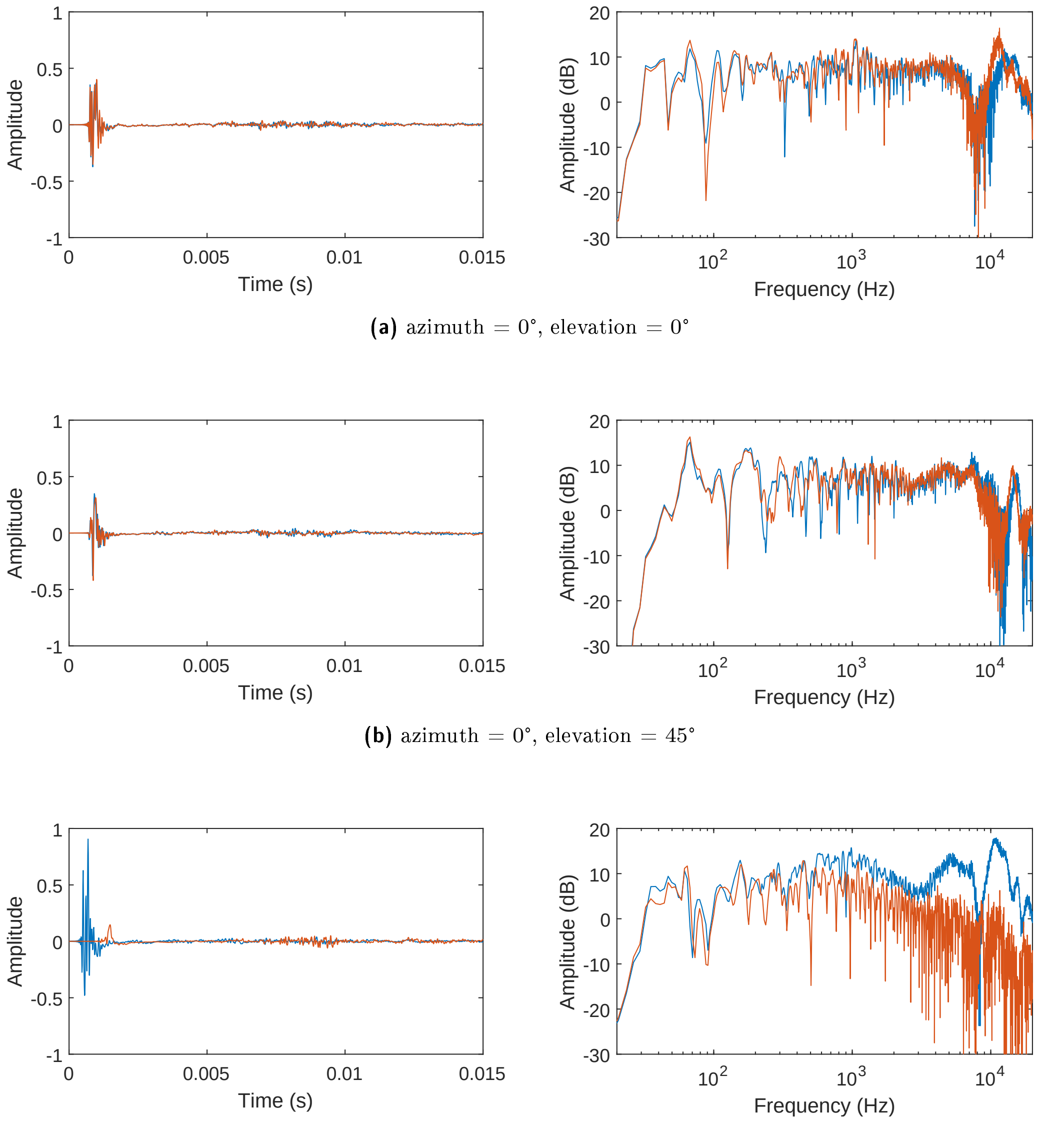


(a) azimuth = 0°, elevation = 0°

(b) azimuth = 0°, elevation = 45°

(c) azimuth = 90°, elevation = 0°

Figure 2.13: Time and Frequency domain plots of 3 stereo HRTFs measured at 1.2m radius as part of the SADIIE database of the KU100 dummy head, see Section 3.3. Blue: Left ear, Red: Right ear. Amplitude is shown on the y-axis (in dB for frequency domain plots).



(a) azimuth = 0°, elevation = 0°

(b) azimuth = 0°, elevation = 45°

(c) azimuth = 90°, elevation = 0°

Figure 2.14: Time and Frequency domain plots of 3 stereo BRIRs measured at 1.5m radius as part of the SADIIE database of the KU100 dummy head, see Section 3.3. Blue: Left ear, Red: Right ear. Amplitude is shown on the y-axis (in dB for frequency domain plots).

channels. The left channel (blue) has a far greater amplitude then the right channel (red) resulting in a large ILD of around 20dB in the high frequencies. Further, the temporal placement of the peaks has shifted. The left channel has moved forward in time ($\approx$ 2ms) and the right channel backward in time ($\approx$ 2.8ms) to give an ITD of around 0.8ms.

2.7.4 BRIRs

Similarly, 3 BRIR measurements are plotted from the same locations as before in the time and frequency domain in Fig. 2.14. In each case similarities may be drawn to the corresponding HRTF, however, the previously clean waveforms and spectra have been corrupted by the response of the room. This is most clearly shown in the extended time-domain responses of each measurement whilst bearing in mind that the room in question has received significant acoustic treatment and so the reflections in this case are still minimal.

In particular, similarities may be observed between the left and right channels of Figs. 2.14a and 2.14b. A separation of the channels in the frequency domain plot in Fig. 2.14c may also be observed, however, in this instance there is a reduction in ILD compared to Fig. 2.13c as a result of diffuse reverberation contributing to the total signal amplitude of both ears.

2.7.5 Binaurally Rendered Impulse Responses

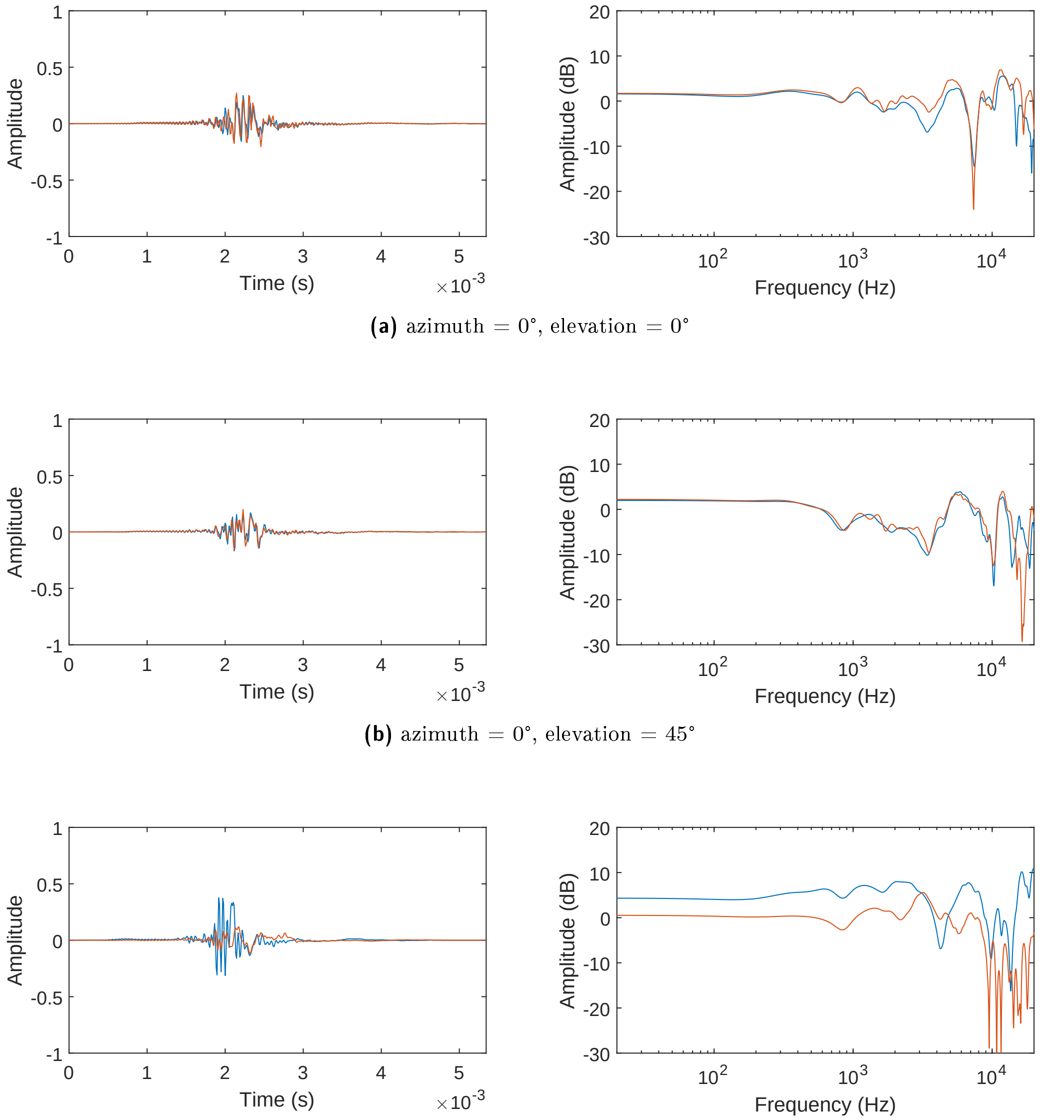
In the analysis of binaural renderers we must consider a third type of measurement, a rendered IR. Although binaural rendering usually refers to the reproduction of typical real world sources such as music or speech, there is nothing that prevents us from defining a more analytical source signal.

A source may be defined as a Dirac pulse, δ, such that

$$\delta = \begin{cases} 1 & t = 0 \\ 0 & t \neq 0 \end{cases} \tag{2.10}$$

The transfer function of the binaural renderer itself may then be isolated for the location at which the pulse was encoded. Depending on the accuracy of our binaural renderer, this transfer function will in general tend toward the raw HRTF or BRIR at that location.

The outputs of a binaural renderer are plotted in Fig. 2.15 for 3 Dirac pulses located in the same directions as the HRTFs and BRIRs presented in Sections 2.7.3 and 2.7.4 respectively. For completeness, the binaural rendering system in question is a single-band, basic, 3$^{\text{rd}}$ order, pseodo-inverse Ambisonic decoder reproducing the source over virtual loudspeakers (see Section 2.6.3) in a 26 point Lebedev grid configuration using KU100 HRTFs from the SADIIE database (see Section 3.3). The details of such a renderer are explained in Chapter 4 but are currently beyond the scope of this chapter. As the binaural renderer utilises the HRTFs of the KU100 dummy head, the output may be compared to the original HRTFs shown in Fig. 2.13.



(a) azimuth = 0°, elevation = 0°

(b) azimuth = 0°, elevation = 45°

(c) azimuth = 90°, elevation = 0°

Figure 2.15: Time and Frequency domain plots of 3 binaurally rendered stereo HRTFs reproduced with 3$^{\text{rd}}$ order Ambisonics over a 1.2m radius virtual array with KU100 HRTFs measured as part of the SADIIE database, see Section 3.3. Blue: Left ear, Red: Right ear. Amplitude is shown on the y-axis (in dB for frequency domain plots).

Interaural Time and Level differences are present in the time domain waveforms e.g. Fig. 2.15c, but are not as clear or as obvious as in Fig. 2.13c. Similarly, whilst the overall shape of the frequency spectra are similar they lack the clear and defined peaks and notches of the original HRTFs. Further, they contain entirely new spectral features that are a direct result of the colouration of the binaural renderer.

Note that these discrepancies between the binaurally rendered IRs and HRTFs vary from those that occur in the BRIRs. The time-domain waveforms remain short and the spectral responses remain smooth. Rather than introduce noise and reverberation the binaurally rendered IRs are more simply a distorted version of the original HRTFs. It is useful to be able to analyse and quantify these changes in order to assess the influence/transparency of the binaural renderer.

2.8 Summary

An overview of binaural features has been given including ITDs, ILDs and spectral cues. These features are encapsulated within binaural filters known as HRTFs or BRIRs. The perceptual weighting of these features has been presented in the context of duplex theory. The ways in which these features are used by the human auditory system to localise a spatial source is discussed in addition to the limitations of such discriminators (i.e. cones of confusion).

It is shown that by rendering a localised Dirac spike through a binaural rendering system, a third type of filter is synthesised, a binaurally rendered IR. These filters describe the transfer function of the binaural renderer for a source rendered at a particular location. Finally, real world examples of measured HRTFs and BRIRs are compared to rendered IRs and their differences are highlighted.

3 Measurement and Perceptual Analysis of Head Related Transfer Functions

Chapter Overview

This chapter presents the SADIIE database, a state of the art high-resolution multi-environment binaural IR database. Methods and workflows are given for the measurement and post-processing of the data (including HRTFs, BRIRs, headphone EQ filters and anthropomorphic data). A listening test is presented which compares the timbral performance of individual and non-individual HRTFs. Results find that the HRTFs of the KU100 are more generally preferred by subjects over their own individual measurements.

3.1 Introduction

Having discussed the principle features of a HRTF and considered methods of objective analysis, it is now appropriate to consider a subjective evaluation. As binaural audio continues to permeate immersive technologies it is vital to develop a detailed understanding of the perceptual relevance of HRTFs. HRTFs can vary significantly between individuals, measurement procedures, simulations and post-processing techniques. The quality with which a source is rendered depends on both the listener's experience and the exact measurements used. The use of non-individual measurements alters the way in which a person perceives a sound. However, it is unclear as to whether this could in fact benefit a listener in some respects (Nicol et al., 2014; Usher and Martens, 2007).

It is proposed that a listener's individual measurements may not be optimal in *every* case. In fact, there is no evidence yet to suggest that they are. Consider a hyper-real VR experience in which audio sources are accentuated beyond what a subject is used to in real life. To that end the measurement and post-processing of the SADIIE (SADIE II) database is presented. The SADIIE database is a state of the art collection of human and dummy head HRTF measurements and is a follow up to the original SADIE (Spatial Audio for Domestic Interactive Entertainment) database (Kearney and Doyle, 2015a). Following that, a listening test is conducted to evaluate the performance of both individual and non-individual HRTFs by rating a series of mono, stereo and binaural stimuli based on 4 pre-defined spatial audio attributes.

The work presented in this chapter has been published by Armstrong et al., 2017; Armstrong et al., 2018a.

3.2 Background

The response of a binaural filter is a result of physiological features and as such is unique to an individual. Although certain characteristics may be generalised, for example an increase in time delay as a source moves toward the contralateral hemisphere, other features such as the high frequency spectral notches caused by the pinnae are not so easily replicated.

It is necessary to explore further the dependency of individual HRTFs on spatial audio rendering quality before conclusions may be drawn regarding optimal rendering strategies. The phrase *individual HRTFs* is used here to refer to the unique HRTF measurements of a particular person. This phrase is used in place of other commonly used terms (e.g. personal, personalized, individualized) in an attempt to discriminate between real-world measurements and alternative techniques that simply aim to optimize a generic set of HRTFs based on a person's feedback, for example Katz and Parseihian, 2012. Techniques for doing this may involve manipulating the ITDs of the HRTFs measured about a dummy head to match that of a listener.

Previous studies have focused extensively on the impact of binaural rendering and individual measurements on source localisation (Wightman and Kistler, 1989a; Wightman and Kistler, 1989b; Hur et al., 2008; Seeber, Fastl and Others, 2003; Møller and Sørensen, 1996; Wenzel et al., 1993; Begault et al., 2000). In general, individual measurements are found to reduce front back confusions in comparison to generic measurements and can yield results similar to real world sources (Møller and Sørensen, 1996; Wenzel et al., 1993). However, results are not always consistent. Begault, when testing speech, attributed localisation accuracy and front back confusion far more to head-tracking then HRTF selection (Begault et al., 2000).

An underlying issue with each of these tests is that in every case the studies fail to fully consider alternative perceptual implications of the selected HRTFs (e.g. timbre). The measurement of a new database is therefore undertaken in order to gather HRTFs and evaluate the timbral performance of individual and generic measurements.

3.3 SADIIE Binaural Database

3.3.1 Overview

The SADIIE database is a state of the art binaural measurement database measured in 2017 and made available online: **york.ac.uk/sadie-project/database.html**. It includes HRTFs, BRIRs, headphone equalisation filters and associated anthropomorphic data. It collates over 60,000 binaural measurements of 20 subjects.

3.3.2 Data Summary

In total, measurements were initially taken from 31 subjects (22 male, 5 female, 2 non-binary, 2 dummy mannequins, ages: 20-63 [majority 20-30]). All subjects gave

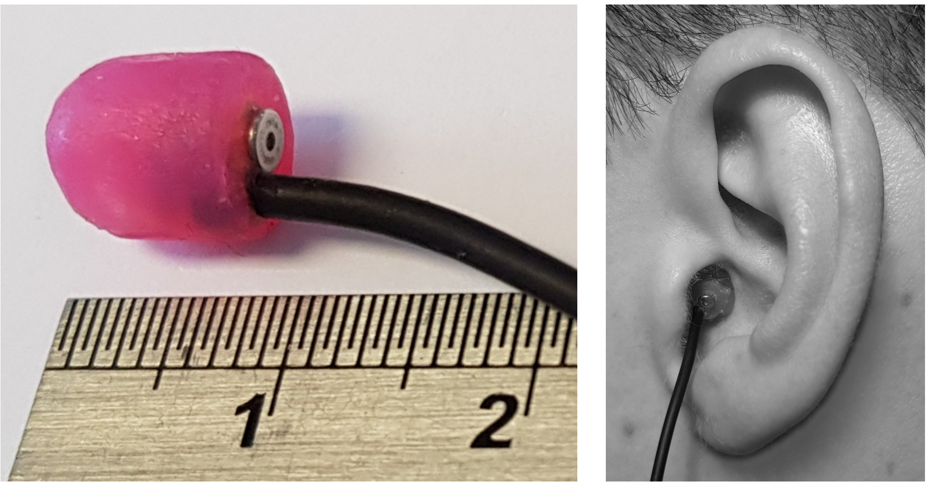
Table 3.1: A comparison of the number of points and minimum elevations measured by a number of popular human HRTF databases: ARI (Majdak, Goupell and Laback, 2010; Majdak, 2013), CIPIC (Algazi et al., 2001), ITA (Bomhardt, De La Fuente Klein and Fels, 2016), LISTEN (Carpentier et al., 2014), Orange Labs (Ospina, Emerit and Katz, 2015), RIEC (Watanabe et al., 2014), SADIE (Kearney and Doyle, 2015a), TU Berlin (Brinkmann et al., 2019), FIU (Gupta et al., 2010). (*) Note that RIEC actually took measurements down to -80°but observed errors due to the measurement set-up below -30°elevation.

	Number of Points	Minimum Elevation
SADIIE	2818/2114	-81°
ARI	1550	-30°
CIPIC	1250	-45°
ITA	2304	-66°
LISTEN	1680	-50.5°
Orange Labs	1560	-56°
RIEC*	865	-30°
SADIE	170	-75°
TU Berlin	440	-90°
FIU DSP Lab	72	-36°

their informed consent for inclusion before they participated in the study. The protocol was approved by the University of York Physical Sciences Ethics Committee. The measurements included

- (HRTFs) A fixed latitude-longitude distribution (equally spaced azimuth and elevation sampling) (NCAR, 2018)
- (HRTFs) 14 key Ambisonic loudspeaker configurations (listed in Section 3.3.3)
- (BRIRs) A 50 point Lebedev Grid (Lecomte et al., 2016)
- Headphone IR of Beyerdynamic DT990s (+ Headphone EQ filter)
- 3D anthropomorphic data (head scan + photos)

Alternative HRTF databases involving human subjects often suffer from a lack of measurements made at low elevations and a limited overall resolution, see Table 3.1. Measurements of dummy heads are more readily available (Gardner and Martin, 1995; Bernschütz, 2013), but are of course less applicable to individual HRTF experimentation. The SADIIE Database includes measurements down to an elevation of -81° and provides a minimum of 2114 measurements for each human subject. It also prioritises optimal spherical distributions for binaural based Ambisonics. In Ambisonics the virtual loudspeaker configuration utilised can have a significant effect on the timbre of the reproduced sound (Mckenzie, Murphy and Kearney, 2017).



(a) Binaural microphone **(b)** Placement within ear

Figure 3.1: (a) A Knowles FG-23329-C05 microphone housed inside a 3D printed capsule (scale in cm) and (b) the position of the capsule inside a participant's ear.

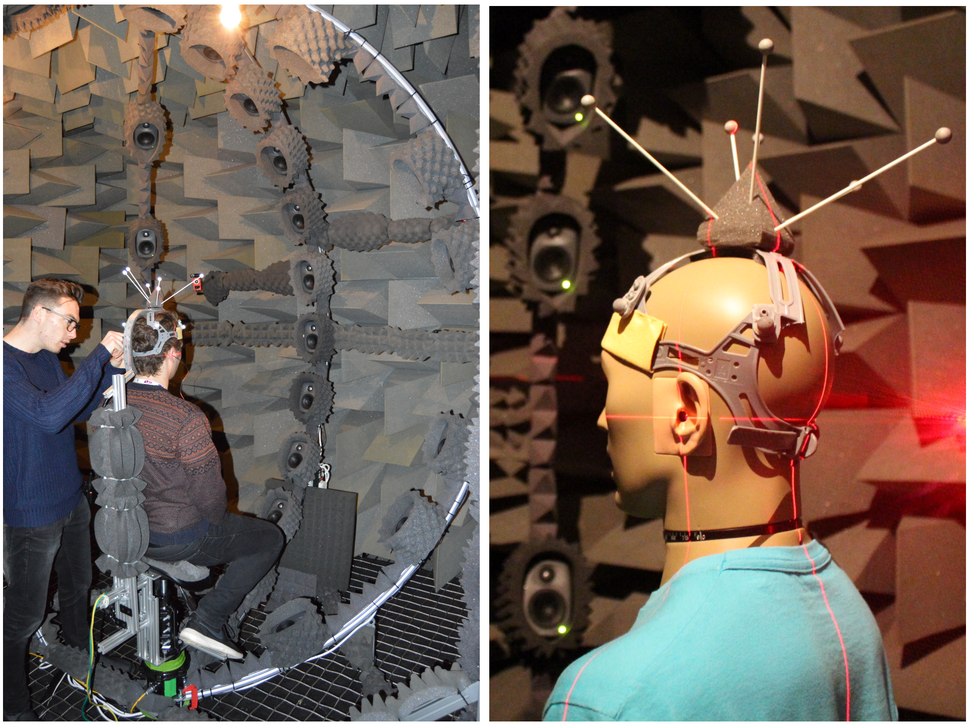
Interpolation of spatially sparse HRTF datasets to achieve a numerically optimal spatial sampling distribution can lead to colouration due to time and/or spectral based distortions in the HRTFs.

In the case of the KU100 and KEMAR mannequins recordings were made using their built in microphones. For human subjects, a pair of Knowles FG-23329-C05 microphones[5] were used and a blocked meatus approach was taken (Møller et al., 1995). The microphones were mounted inside 3D printed capsules and secured in the participants' ears with silicon putty, see Fig. 3.1. Once inserted, the microphones were not removed or re-positioned until all audiological measurements had been completed.

20 Subjects were admitted to the final database (15 male, 1 female, 2 non-binary, 2 dummy mannequins, ages: 20-63 [majority 20-30]). Inclusion was subject to the quality of their measurements determined by observational notes and analysis of spectral, ITD and ILD plots. The exclusions are detailed below. Qualifying datasets included those of the KU100 dummy head and KEMAR mannequin.

- 1 subject voluntarily stopped the measurement procedure part way though.
- 6 subjects were excluded due to excessive movement and shuffling in-between measurements.
- 2 subjects were excluded due to minor asymmetries in their ITD plots.
- 2 subjects were excluded due to unexplained discontinuities in their measurements, possibly a result of movement.

[5] knowles.com/series/dpt-microphones/subdpt-subminiature-microphones/series-fg-bfg



(a) Measurement rig **(b)** Laser alignment

Figure 3.2: (a) A subject being prepared for HRTF measurements. They are sat on a motor-controlled rotating 'saddle stool' and their head movement has been restricted by a motion tracked restraint. (b) An example of the cross-axis laser guides used to align a subject's interaural axis to the centre of the loudspeaker array.

3.3.3 Head Related Transfer Functions

Measurement

An HRTF measurement rig was designed and constructed in the fully anechoic chamber (measuring 2.5m x 1.5m x 3m) at the Audio Lab, University of York, U.K., see Fig. 3.2a (Armstrong et al., 2017). The set-up consisted of three static, vertical semi-circular arcs, each separated by 45° azimuth. 23 Genelec 8010 loudspeakers[6] were installed across the three arcs at 23 unique elevations and at a radius of 1.2m. In each case the loudspeaker was positioned with respect to its acoustic axis (Genelec, 2014). The 8010 was chosen for its small footprint and reliable frequency response (±2.5dB) from 74-20kHz. However, the non-coaxial nature of the loudspeaker still resulted in a frequency dependant lateral error of approximately ±35mm. This is equivalent to an angular displacement of 1.7°at a radius of 1.2m which is well within the understood range of human perception of elevation (Blauert, 1997).

Participants were sat on a height adjustable 'saddle stool', selected for its minimal acoustic occlusion. The stool sat on top of a rotating Yaesu G-2800DXC satellite

[6] genelec.com/8010

dish motor which was attached to the rig and could be controlled via a GS-232B serial interface[7]. The participant's feet were tucked underneath their body and were supported by a footrest. Their head was positioned using a laser alignment system such that in their initial orientation they were facing forward with their inter-aural axis aligned to the centre of the loudspeaker array.

Their head was restrained using the internal strapping of a commerciality available hard hat. The strapping was attached to a rigid back rest to help prevent unintentional head movement, as shown in Fig. 3.2a. Their head position was tracked in real time via a set of 10 reflective motion tracked markers (6 positioned asymmetricaly above the head and 4 positioned around the head) fixed to the top of the restraint, as shown in Fig. 3.2b. Four Optitrack Flex-3 Infra-Red motion capture cameras tracked the 10 point rigid-body to within $< 0.1^\circ$ via optical motion capture software, Motive[8] (Version 1.8.0 Final 64-bit).

Utilizing this head-tracking data, the participants were freely rotated about the horizontal plane to face a series of predetermined azimuthal positions. As the participant was rotated the relative azimuthal co-ordinate of each loudspeaker was thus redefined. A sequence of rotations was programmed such that the loudspeaker coordinates satisfied the intended measurement configurations. At each azimuthal position, the rotation was halted and a 2 second pause allowed any mechanical noise/oscillations to settle. An overlapped exponential swept sine wave technique (Majdak, Balazs and Laback, 2007) was used to quickly and efficiently measure the IRs from all 23 loudspeakers, regardless of their direct affiliation to a configuration.

The use of an exponential sinusoidal sweep is an effective technique to measure a source-receiver transfer function over a range of frequencies (Farina, 2000). The recorded signal is convolved with an inverse copy of the sweep to remove the time-smeared element of the input signal and re-align and normalize the various frequency components in the time domain. The inverse sweep must account for both the spread of frequencies in the time domain and the reduced input power of the higher frequencies given the exponential nature of the sweep. It may be calculated by first time-reversing the original sweep and then applying a frequency compensation of +6dB/Octave (High frequency boost). Note that this logarithmic frequency compensation may be performed as a linearly decreasing amplitude envelope applied directly to the time-reversed sweep. In Matlab this may be done:

[7] yaesu.com/

[8] optitrack.com/products/motive/tracker/

```
% Time reverse the sweep
sweep_rev = fliplr(sweep);

% Calculate the high freq boost as a low freq attenuation
% (number of ocataves = log2(max_freq/min_freq))
att = (-6*(log2(max_freq/min_freq)));

% Envelope in terms of dB
env_dB = 0:att/(length(sweep)):att-(att/(length(sweep)));

% Envelope in terms of amplitude
env_amp = 10.^(env_dB./20);

% Apply amplitude envelope
inverse_sweep = sweep_rev.*env_amp;
```

The process is known as de-convolution and results in the IR of the source being located at the moment the input sweep finishes. To save time, the sweeps output from the loudspeakers may be overlapped provided that there is no interference between the IRs once the signals are deconvolved (Majdak, Balazs and Laback, 2007).

24 second sweeps separated by 0.15s were performed with 0.1s fade in/out half-Hanning windows over the frequency range 200Hz-24kHz. The entire process was automated with control software written in Max MSP[9] and operated by technicians in an isolated control room via a dedicated Local Area Network (LAN).

Recordings were made via a RME Fireface 400 interface[10] at 96kHz sample rate and 24 bit resolution. Raw measurements were deconvolved using an unwindowed inverse sweep and the individual IRs were separated ensuring no overlap of the linear or harmonic distortion products of neighbouring sweeps in the deconvolution. IRs were trimmed to approximately 15ms before and 10ms after their peak amplitude to remove minor spurious reflections (assumed to come from the door frame of the anechoic chamber). An approximate SNR of 65dB was measured from the noise floor to the peak value of an IR measured from a frontal loudspeaker via a flat omnidirectional GRAS 46AE measurement microphone[11] positioned at the centre of the loudspeaker array.

[9] cycling74.com/products/max/

[10] archiv.rme-audio.de/en/products/fireface_400.php

[11] gras.dk/products/measurement-microphone-sets/product/140-46ae

Table 3.2: Distributions considered for HRTF measurement and their corresponding elevations. Approximations are indicated by the symbol ≈. Note that elevations correspond to the vertices (not faces) of each distribution.

	0°	±15°	±17.5°	±25°	±30°	±35.3°	±45°	±54°	±60°	±64.8°	±75°	±90°†
Lattitude-Longitude Distribution	●	●			●		●		●		●	●
Tetrahedron						●						
Octahedron (x4 orientations)	●											●
Cube						●						
Bi-Rectangle (x3 orientations)	●						●					
26pt Lebedev Grid	●					●	●					●
50pt Lebedev Grid	●		●			●	●			●		●
Icosehedron	≈	≈					≈			≈		
24-point Hardin and Sloane 7-Design		≈		●					●			
Pentakis Icosedodecahedron	●		≈		≈			●	≈			●

† -90° modelled by the interpolation of measurements made at -81°

Regarding the positions of the 23 loudspeakers, 12 elevations were measured at 15° intervals between -75° and 90°. These were necessary to measure the regular lattitude-longitude distribution (NCAR, 2018). Source localisation in the median plane is reported to be significantly worse than that in the horizontal plane (Blauert, 1997). A localisation blur of ±9° is reported for continuous familiar speech (Damaske and Wagener, 1969) whilst ±17° is reported for continuous unfamiliar speech (Blauert, 1970). Measurement intervals of 15° were considered to be of fine enough resolution to give an accurate representation of perceptual localisation cues without oversampling the subject unnecessarily.

A further 10 elevations were determined according to the common elevations coordinates of 11 typical Ambisonic loudspeaker layouts. An additional three layouts are composed of the same approximate elevations ±2°. A summary of the layouts and corresponding elevation coordinates is given in Table 3.2. At a distance of 1.2m an error of 2° translates to a speaker displacement of 4.2cm. This is small with respect to the size of the loudspeaker (18.1cm) and main driver (12cm). Considering such a minor displacement in elevation in relation to the resolution of the human ear, it is proposed that this error has little perceptual influence.

An elevation angle of $-90°$ could not be measured as the area was blocked by the installation of the chair and motor. Instead, a nearest alternative angle of $-81°$ was measured at each azimuth. In post-processing these measurements were interpolated to approximate a measurement at $-90°$ as follows: for each ear, measurements were time aligned by their peak amplitude to the average delay of the subset. A linear interpolation was then performed in the time domain by calculating the average mean amplitude of each sample. Due to the nature of HRTFs measurements made at such low elevations the majority of high frequency detail is occluded by the legs/torso/chair. It is therefore reasonable to use an interpolated measurement which will preserve mainly the low frequency cues.

64 azimuthal stoppages were required to measure the Ambisonic configurations. In addition, further regular stoppages were required to measure the fixed lattitude-longitude distribution. A 1° resolution was chosen for the dummy subjects. This required a total of 399 stoppages. It generated 8802 unique measurements and took over 3 hours to complete.

Unfortunately, this was too long to ask a human participant to sit still for, especially given the uncomfortable nature of the seat and head restraint. The horizontal resolution of the latitude-longitude distribution was therefore reduced on a subject by subject basis. 11 subjects (7 admitted to database) were measured with a 5° resolution. This required 127 stoppages, generated 2818 unique measurements and took approximately 1.25 hours. 18 subjects (11 admitted to database) were measured with a 10° resolution. This required just 95 stoppages, generated 2114 unique measurements and took approximately 1 hour.

Low Frequency Compensation

Due to the size of the loudspeakers' diaphragms and the low-frequency isolation limitations of the anechoic chamber, frequencies below 200Hz could not be reliably measured and were instead modelled. At low frequencies (<400Hz), analytical simulations such as those from Bernschütz, 2013 show that there is very little variation in the response of an HRTF. Even a listener's head barely effects (<1dB) the frequency spectra of the filters. It is therefore reasonable to adopt a low frequency model, similar to that presented by Xie, 2013, which extends a flat frequency response and linear phase response below approximately 400Hz.

Low frequency compensation was performed independently for each channel of each HRTF. A crossover frequency of 275Hz was chosen. This balanced the preservation of natural higher-frequency content with the need to accommodate a crossover filter's

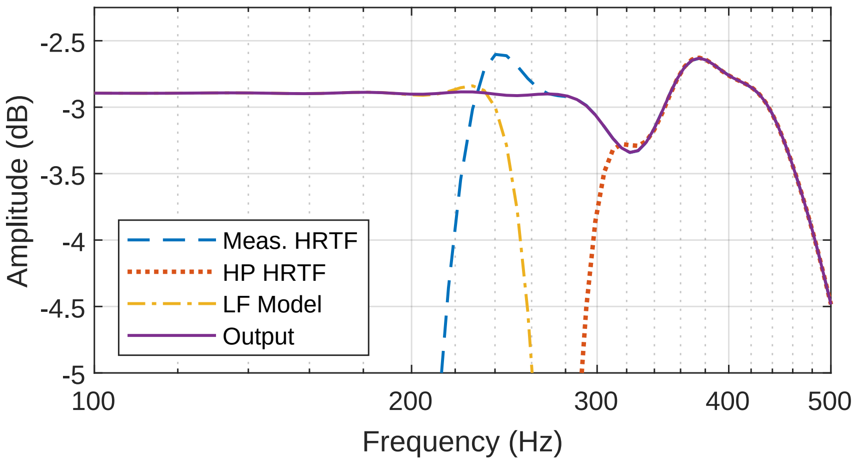


Figure 3.3: The 275Hz crossover of a measured HRTF and individually generated low frequency model.

low frequency roll off when applied to the HRTF signals which only included data down to 200Hz.

A Dirac pulse was generated with an amplitude and delay equal to the mean average amplitude and group delay of the signal to be extended between 250-300Hz. This data was calculated using a FFT and matlab's `grpdelay()` function respectively. The phase responses of both the original signal and Dirac model were then compared at the crossover frequency. Erroneous spikes in the group delay evaluation would slightly overestimate the true group delay of the signal at the crossover frequency. The delayed Dirac model would therefore consistently lag the signal. The Dirac model was therefore shifted *forward* in time to align the signals' phases. A forward shift ensured that the low frequency model remained well within any future amplitude windows applied to the HRTF around its peak.

A pair of 4096 tap FIR low/high pass crossover filters were used to overlay the low-frequency Dirac model with the valid high-frequency portion of the input signal. High order filters ensured that neighbouring frequencies were sufficiently attenuated to avoid de-constructive interference caused by slight phase misalignments in these regions. An example of the crossover between a measured HRTF and a corresponding low frequency model is shown in Fig. 3.3. A smooth transition between the High-Pass filtered HRTF and Low-Pass filtered low frequency model can be seen in the output signal. Note the small amplitude variation (<1dB) below 400Hz in the measured signal.

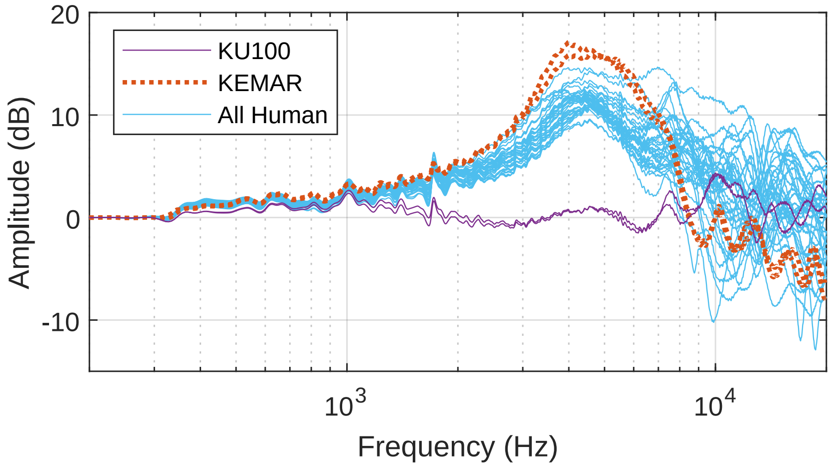


Figure 3.4: The diffuse-field response of each ear of each subject (total 40 responses) after free-field equalisation of both the average loudspeaker and respective binaural microphone responses. The responses of the KEMAR and KU100 dummy heads are distinguished from the other human responses. The plot shows a common broad peak at approximately 4KHz which may be explained by ear canal resonance.

Equalisation and Windowing

Correctly processed HRTFs require either diffuse-field or free-field equalisation to account for any frequency response variations in the measurement system. Diffuse-field equalisation attempts to remove all commonality from a set of measurements. It is typically performed by first calculating the average response of a set of measurements and then removing this response from each measurement individually. The result is that the average frequency response of the processed measurements is ideally flat. Free-field equalisation aims to remove only the direct response of the measurement system. It is typically performed by measuring the response of the complete measurement system (e.g. loudspeaker, microphone, interface) without any external influence (in the case of HRTFs, this would be the head/participant) and removing this response from each measurement.

The drawback of diffuse-field equalisation is that it would be possible to 'over-equalise' the measurement set. For example, if the shape of a participant's body/-head/ears meant that regardless of direction they heard with a slight high frequency boost, this boost in high frequencies would be removed by diffuse-field equalisation meaning that if the subject were to listen back to their own HRIRs they may perceive them as dull (or lacking in high frequencies) when compared to real world stimuli. The risk of over-equalisation is considered by analysing Fig. 3.4. The figure shows the diffuse-field response of each ear of each subject after free-field equalisation of both the average loudspeaker and respective binaural microphone responses.

The free-field responses of the loudspeakers were measured by placing a flat response GRAS 46AE measurement microphone at the centre of the loudspeaker array and measuring a set of sweeps from the loudspeakers as if measuring a set of HRTFs. These sweeps were then processed and their frequency responses evaluated. Binaural microphone responses were calculated as follows: a 20Hz-24kHz sine sweep was output from a Genelec 8040[12] loudspeaker in the anechoic chamber. The sweep was simultaneously recorded by a flat-response GRAS 46AE measurement microphone and each of the individual binaural microphones that were used as part of the database. Each binaural microphone was placed as close as possible to the measurement microphone. A FFT was taken of each recording and the spectral response of the measurement microphone was subtracted from that of each binaural microphone. This resulted in the spectral responses of each binaural microphone excluding the response of the loudspeaker. Inverse linear-phase FIR filters were computed using Kirkeby and Nelson regularization (Kirkeby and Nelson, 1999).

Referring back to Fig. 3.4, the free-field equalised diffuse-field responses of the KEMAR mannequin and human subjects all follow a similar trend, peaking at around 4kHz. The KEMAR response peaks highest at about 17dB. It is suspected that this peak is a result of early ear canal/cavum concha resonance. This would explain both the similarity and slight variation between subjects as the microphones could not always be placed at exactly the same depth within each participant's ears. It would also explain the amplitude of the KEMAR response whose microphones are housed deeper within the ear canal opening of the pinnae mould. (However, note that GRAS's 'Anthropometric Pinna' with improved ear canal modelling were *not* used during these measurements.)

In contrast, the response of the KU100 is relatively flat. This is to be expected as the dummy head is pre-calibrated with a diffuse-field equalisation filter in-factory (*KU100 Operating instructions*). The high frequency variation (>10KHz) may also be attributed to the microphone placement, for example, relating to the the angle of the microphone within ear or the placement of attached wire. The trouble is that at these frequencies the response if far more volatile and far less reliable.

The remaining diffuse-field response is therefore attributed to a feature of the measurement procedure. As such, it is fair to remove the response from the final HRTF

[12] genelec.com/support-technology/previous-models/8040a-studio-monitor

measurements. Consequently, within the SADIIE database all measurements are diffuse-field equalised. Whilst free-field equalisation remains arguably the more accurate method, there are several reasons as to why diffuse-field equalisation was chosen:

- a large enough set of data points was being captured to make diffuse-field equalisation viable;
- it would take into account the free-field response on the system in situ i.e. influences due to the placement of the microphone capsule within the ear canal;
- it would compensate for any generic response of the post-processing (e.g. windowing);
- it would equate the average frequency response of each dataset to provide a timbral consistency across the database (recall the KU100 is pre-calibrated with a diffuse-field equalisation filter (*KU100 Operating instructions*));
- it provides a compatible set of measurements for both loudspeaker and headphone reproduction by avoiding the over-reproduction of the transfer function of the external ear (Theile, 1984; Griesinger, 1989);
- it helps to ensure the reproduction of accurate tone colour considering the random directions from which many reverberant reflections could emanate from (Blauert, 1997).

The equalisation was performed in two stages: before and after a windowing operation imposed to reduce the tap length of the filters. For each stage the power average response of the dataset was calculated in the frequency domain for each ear of each subject. A Voronoi weighting was applied to the contribution of each measurement based on a solid angle calculation of neighbouring measurements. This ensured that clustered measurements did not over-represent a particular direction in the average. Inverse linear-phase FIR filters were calculated from the diffuse-field response using Kirkeby and Nelson regularization (Kirkeby and Nelson, 1999) to perform each equalisation.

Stage one was designed to compensate for the response of the measurement system. Input data was left unwindowed to preserve as much of the original signal content

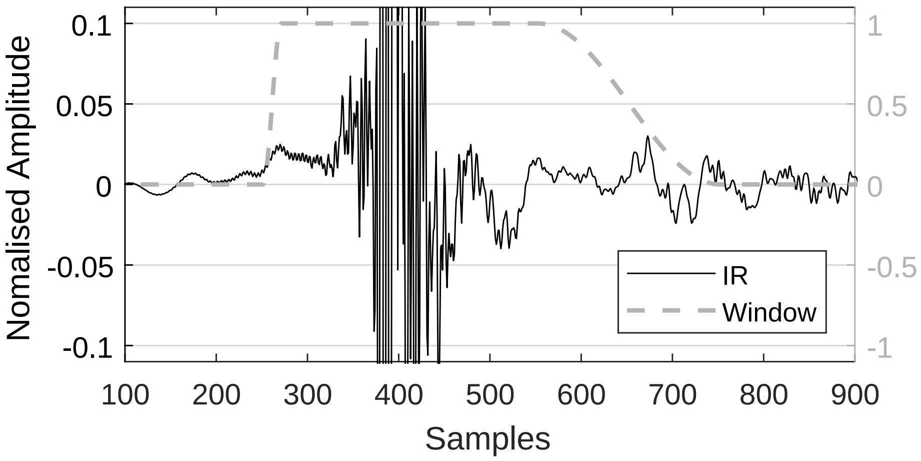


Figure 3.5: Windowing of HRTFs to 500 samples by means of a 20 sample half Hanning window and 130 sample pad before each peak and a 150 sample pad and 200 sample half Hanning window after. Note that the plot has zoomed in on the IR to illustrate the region of windowing.

as possible. 1/3$^{\mathrm{rd}}$ octave band smoothing was used to prevent overly-sharp peaks or notches appearing in the frequency response of the inverse filter and exacerbating the time-domain aliasing.

IRs (at fs = 96000Hz) were then windowed to 500 samples (approximately their final length) by means of a 20 sample half-Hanning window and 130 sample pad before each peak and a 150 sample pad and 200 sample half-Hanning window after, see Fig. 3.5. The proportions and relative position of this window affected both the preservation of the filter's frequency response and the final diffuse-field response of the dataset. By biasing the length of the fade out over the post-peak pad and shifting the window to preserve more of the pre-peak signal the diffuse-field variance can be reduced. However, accurate preservation of the frequency response generally required as much of the post-peak signal to remain as intact as possible. A sample window of length 500 was chosen to either match or exceed other preceding HRTF databases (Kearney and Doyle, 2015a; Bernschütz, 2013) to preserve key localisation features without allowing the length of the filter to extend beyond current real-time computational limitations.

Systematic frequency domain errors introduced by the windowing operation were compensated for by a second stage of diffuse-field equalisation. This equalisation was performed on the windowed IRs with 1/5$^{\mathrm{th}}$ octave band smoothing. As the first stage of equalisation had already considerably smoothed out the diffuse-field response a less smooth filter was required.

The IRs were grouped and trimmed as a set to 512 samples (96KHz) inclusive of 10 sample fade in/out half Hanning windows. It was ensured that at least 180

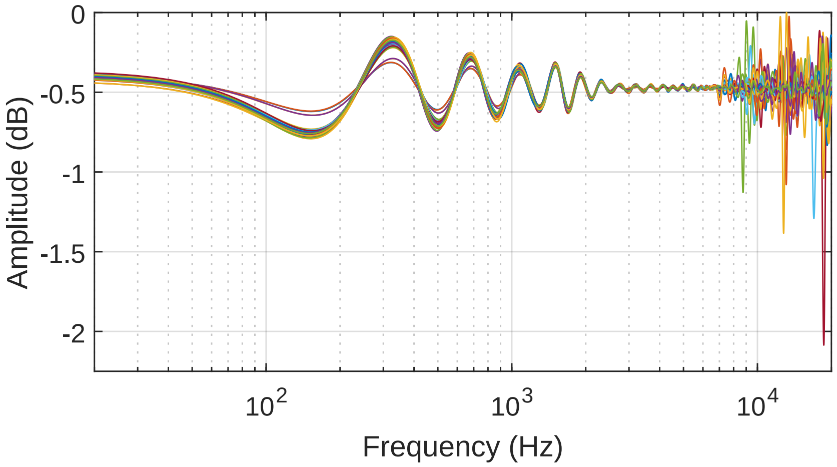


Figure 3.6: The final normalised diffuse-field response of each ear of each subject (total 40 responses) after all post-processing.

samples remained before each peak and at least 230 samples remained after. This left approximately 100 samples to account for the variances in peak onset time due to ITDs. The second windowing resulted in minimal disturbance to the time domain waveforms and therefore had virtually no subsequent effect on the frequency or diffuse-field responses.

Fig. 3.6 shows the final normalised diffuse-field response of each ear of each subject calculated after all post-processing: deconvolution and separation of original recordings, low frequency compensation, stage 1 diffuse-field equalisation, windowing, stage 2 diffuse-field equalisation, trim to 512 samples. Comparing the magnitudes of the frequency bins of each response, 95% fall within a 0.33dB range (approximately -0.35dB to -0.65dB). The pattern followed by each response below approximately 7KHz can be attributed to the windowing parameters.

3.3.4 Binaural Room Impulse Responses

Although not directly relevant to this work, it is rare to have to opportunity to measure comparative HRTF/BRIR binaural filters and hence there is a lack of this type of data within the industry. Therefore, a complimentary set of BRIRs were also measured for each subject directly after the HRTFs. Details of these measurements are included in the hope that they may be of use in future work.

Participants were led directly from the anechoic chamber to a treated listening room where BRIRs were measured from an acoustically calibrated 50 point Lebedev grid loudspeaker array, see Fig. 3.7. Measurements of this configuration are particularly useful as nested within it are the 6 and 26 point Lebedev grids (Lecomte et al.,

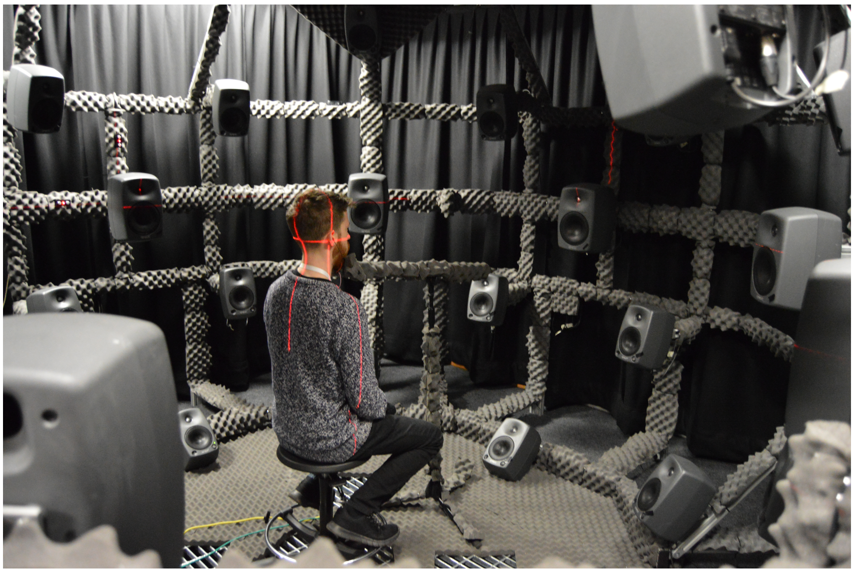


Figure 3.7: BRIRs of a 50 point Lebedev loudspeaker configuration being measured inside a treated listening environment.

2015). This particular array utilizes two types of loudspeaker. 40 Genelec 8030s[13] are supported by 10 Genelec 8040s, for low frequency reconstruction. The rig is enclosed by a thick curtain. Measurements of the KU100 and KEMAR mannequins produced reverberation times of around 50-65ms for a drop of 60dB, dependant on speaker location. This is a relatively short reverberation time and although it is not long enough to provide an abundance of late reflections, the measurements do still include the first few key early reflections (e.g. from the floor) that will aid with a sense of externalisation.

Participants were sat on a stool and their interaural axis was laser aligned to the centre of the array. A rigid, acoustically dampened chin rest was used to ensure the participant kept their head still throughout the measurement procedure. 3s exponential swept sine waves were played out of each loudspeaker one at a time over the range 20Hz-24kHz. Recordings were made via a MOTU UltraLite-Mk3 Hybrid audio interface[14] at 96kHz sample rate and 24 bit resolution.

After deconvolution, free-field equalisation of each of the microphone's frequency responses was performed using linear-phase FIR filters. Microphone responses were calculated as discussed in Section 3.3.3. By only equalising for the microphone responses (not loudspeakers) it ensured that the measurements most accurately represented the real-world listening conditions of the loudspeaker array. The BRIRs were trimmed to 0.3s inclusive of 10 sample fade in/out half-Hanning windows.

[13] genelec.com/support-technology/previous-models/8030a-studio-monitor

[14] motu.com/products/motuaudio/ultralite-mk3

3.3.5 Headphone Equalization Filters

Headphone Impulse Responses (HpIRs) are binaural measurements taken from the left and right transducers of a pair of headphones sat over a person's ears. Measurements of this type do not usually include any interaural crosstalk (i.e. the response of the left transducer measured from the right ear). They are therefore presented as a single stereo response where:

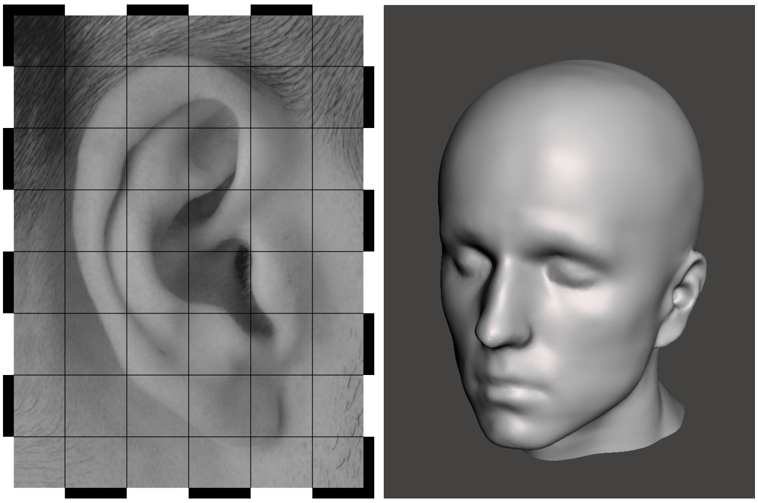
$$\text{Left Channel} = \text{Left transducer} \rightarrow \text{Left Ear}$$
$$\text{Right Channel} = \text{Right transducer} \rightarrow \text{Right Ear}$$

Headphone Equalisation (HpEQ) may be performed by implementing a filter with the inverse response of the HpIR. This compensates for the transfer function of the particular pair of headphones coupled to a person's outer ear and may help in ensuring an accurate reception of binaural signals (McAnally and Martin, 2002; Schärer and Lindau, 2009).

Without removing the binaural microphones participants were asked to put on a pair of open back Beyerdynamic DT990 headphones. 3s exponential swept sine waves were output from each transducer (one at a time) and were recorded through the MOTU interface. This was repeated 10 times. In between each pair of measurements the participant was asked to remove the headphones completely and place them back on their head. This helps to ensure a good average response as the headphones are seated onto the person's head slightly differently each time.

After deconvolution, the measured IRs for each ear were power averaged together in the frequency domain. An inverse FFT followed by a circular shift of half the FFT size brought the data back into a stable format in the time domain (i.e. a continuous waveform with a singular central peak). Again, a free-field equalisation of the binaural microphone frequency responses was performed and the HpIRs were trimmed to 2048 samples. Finally, the entire response was amplitude weighted by a full length Hanning window.

Linear Phase HpEQ filters were generated by inverting the frequency responses of the HpIR filters. This was done over the range 120Hz-24kHz with 1/5th octave band smoothing and Kirkeby and Nelson regularization (Kirkeby and Nelson, 1999). Responses were trimmed to 2048 samples and amplitude weighted by a full length Hanning window.



(a) Pinnae photo **(b)** 3D head scan

Figure 3.8: (a) A hi-resolution digital photograph of a participant's pinnae scaled to a 1 cm grid. (b) A 3D scan of a participant's head.

3.3.6 Anthropomorphic Data

3D head scans were captured using a Polhemus FastSCAN Laser scanner [15]. The data was captured as a point cloud and converted to a mesh using the products own software. Manual manipulation (performed in Autodesk Meshmixer[16]) formatted the meshes for publication. Head orientations were fixed such that the subject was always facing forward and their interaural axis sat on the same horizontal plane as the tip of their nose. These scans were not intended to be suitable for simulation purposes and as such their resolutions were not monitored. They were instead taken in order to approximate anthropomorphic measurements of the subjects at a later date if required.

High resolution digital photographs were also taken of each participant's pinnae. The photographs were scaled by placing a vertical and horizontal ruler in the same approximate plane as each pinnae. For each photograph the participant was asked to stand up straight and look directly forward. The rotational position of each ear is therefore representative of this head position. An example of each type of measurement is given in Fig. 3.8.

[15] polhemus.com

[16] meshmixer.com/

3.4 Listening Test: HRTF Preference

3.4.1 Overview

A listening test was conducted to investigate the existence of quantifiable timbral and/or spatial attributes within individual and non-individual HRTF measurements. 16 participants (13 male, 1 female, 2 non-binary, ages: 20-63 [majority 20-30]) all of whom were admitted to the SADIIE database were re-recruited for the test. All subjects gave their informed consent for inclusion before they participated in the study. The protocol was approved by the University of York Physical Sciences Ethics Committee. Participants were presented with a set of auditory stimuli over headphones and were asked to rate each one based on 4 attributes as defined in Section 3.4.2: brightness, richness, externalisation and preference.

3.4.2 Spatial Audio Rendering - Quality Assessment

The evaluation of spatial audio is a complex topic. One must be careful to define parameters that are descriptive enough to capture the essence of any given stimuli without overwhelming a listening test participant. It is necessary to consider many different aspects, for example: localisation, timbre, spatialization, naturalism and fidelity as well as the impact of listener preference. It is impossible to ensure that what may sound good to one person will sound as pleasing to the next.

Alternative work has explored HRTF preference through: methods of database optimization (Katz and Parseihian, 2012), the examination of perceptual repeatability and variability (Andreopoulou and Katz, 2016a; Schönstein and Katz, 2012) and the creation of global similarity metrics (Andreopoulou and Katz, 2015; Andreopoulou and Katz, 2016b). However, in each case perceived spatial localisation performance was used as an all encompassing metric for comparison. Impulsive or noisy stimuli was presented over a known trajectory and participants were asked to rate the spatial effectiveness of each sample. Considering a more general and real world scenario it is important to evaluate beyond just the spatial attributes of a rendered source (Huopaniemi, Zacharov and Karjalainen, 1999). Further to objective accuracy, it is beneficial to examine the impact of HRTF selection by the more general evaluation of spatial audio stimuli within a real-world context.

Previous work has standardised attributes for the subjective assessment of sound quality which are presented in the ITU Recommendation BS.1284-1 published in

2003 (ITU, 2003a; EBU, 1997). However, the recommendation lacks sufficient attributes for the assessment of spatial audio (Nicol et al., 2014). Early examples of subjective binaural evaluation (Huopaniemi, Zacharov and Karjalainen, 1999; Lorho et al., 2000) lack clarity and consider only general spatial or timbral colouration. Pulkki, Karjalainen and Huopaniemi, 1999 and Huopaniemi, Zacharov and Karjalainen, 1999 introduce perception based binaural models as a measure of binaural signal quality. Similar objective metrics have been published since (ITU, 2001; Thiede et al., 2000). Whilst such models are useful for monitoring authenticity, they operate by comparing a test signal to a given reference signal and as such do not directly assess a listener's Quality of Experience (QoE).

Other work has identified comprehensive lists of attributes tailored for the perceptual evaluation of spatial audio. Whilst the processes with which these lists were compiled vary by author, all result in a similar collection of holistic terms.

Berg and Rumsey, 1999 proposed a set of spatial attributes based on the Repertory Grid method in which subjects identify differences in triads of stimuli. Koivuniemi and Zacharov, 2001 presented a structured method for the development of any descriptive language. In an example, 12 expert listeners produce an exhaustive list of 8 spatial and 4 timbral attributes for evaluating different spatial sound reproduction systems. Lindau et al., 2014 developed the Spatial Audio Quality Inventory (SAQI) which presents 'a vocabulary containing all perceptual attributes'. It is derived from a focus group of 21 German speaking virtual acoustics experts. Lokki et al., 2012 focused on the acoustics of concert halls developing a broad list of attributes from the results of an individual vocabulary profiling experiment. Pearce, Brookes and Mason, 2017 examined the search terms used in online sound effect libraries and compiled a list of the most popular discriminators. Simon, Zacharov and Katz, 2016 were a little more specific and identified 8 qualities for describing the perceived differences between non-individual HRTF sets in binaural renderings. He first followed an individual vocabulary profiling procedure, similar to Lokki, before refining his terms through a series of focus groups.

The common elements of these lists were evaluated to identify 4 discriminatory attributes to be assessed within a listening test, see Table 3.3. The participants selected for this test were being recruited from a small pool of candidates (those measured for the SADIIE database). It was therefore important to gather as high a

Table 3.3: Spatial audio attribute scales and definitions as used in the perceptual listening test.

Attribute	Anchors	Definition
Brightness	Dark → Bright	The abundance of high (/low) frequencies.
Richness	Thin → Rich	A full and well balanced mix. Inclusive of all frequencies and with no obvious boosts or cuts.
Externalisation	In-Head → External	The locatedness of sources to distant points in space.
Preference	Unfavoured → Preferred	An overall plausibility of the soundfield.

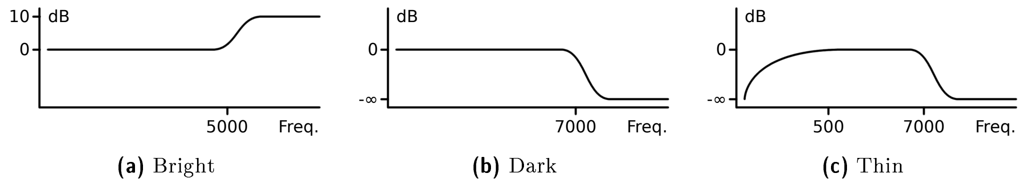
quality set of data from each participant as possible. An exhaustive list of attributes would have been an arduous task for any individual. To avoid a detrimental effect on the results due to listening fatigue, a smaller selection of attributes was chosen and participants were encouraged to think more carefully about the ratings given to each stimuli.

Koivuniemi and Zacharov, 2001, Lindau et al., 2014, Lokki et al., 2012, Pearce, Brookes and Mason, 2017 and Simon, Zacharov and Katz, 2016 all identify *brightness* (/darkness) as the abundance of high (/low) frequencies. The same definition is used here. A term to describe a sense of fullness is also included by each author. It is referred to here as *richness* (as in (Koivuniemi and Zacharov, 2001)). It is described as the sense of a full and well balanced mix inclusive of all frequencies and with no obvious boosts or cuts.

Regarding spatial attributes, a sense of *externalisation* is identified by Simon, Lindau, Koivuniemi and Berg and Rumsey, 1999. The same term is used here and is described specifically as the locatedness of sources to distant points in space. An overall feeling of realism or naturalness, is also identified by the same authors. These sensations are summarised here by the term *preference*, implying an overall plausibility of the soundfield.

3.4.3 Listening Test

The terms brightness and richness were described to each participant during a short training phase. Example audio files containing filtered excerpts of music were also played to illustrate the meanings of the terms. 3 excerpts were filtered according to the frequency response plots shown in Fig. 3.9. A high boost simulated a bright



(a) Bright (b) Dark (c) Thin

Figure 3.9: Frequency responses of the filters used to generate stereo anchor stimuli.

signal, a high cut simulated a dark signal and a low and high cut simulated a thin signal. 1 further example was left unfiltered to simulated a rich signal.

The term externalisation was not so easy to conceptualise. To verify the effectiveness of any example file would have required the verification of the spatial filters used to create such a file. This was in part the purpose of this study. Participants were instead provided with a graphical depiction of the soundfield and were advised that effective externalisation would be as if they were hearing the sources in real life. In considering preference, participants were asked to consider not only a sense of spatialness, but also a pleasant timbre and overall feeling of realism. In some ways it could be considered as an overall combination of the 3 other attributes but in a more personal sense rather than analytical. Participants were given the opportunity to ask any questions relating to the definitions of each term.

During the test, participants were able to freely switch between the set of stimuli over a continuous looped playback. Ratings were performed using the graphical interface shown in Fig. 3.10. Participants were required to drag a marker corresponding to a stimuli to a point on a graph. 2 graphs were used to represent the 4 attributes on continuous scales. Brightness and richness were represented by the x and y axes of one graph and preference and externalisation the axes of the other. The interface allowed participants to easily compare and adjust the ratings they were giving to each stimuli. They were instructed to make use of the entire range.

Noise bursts and other broadband signals are common stimuli used throughout listening tests, however, such unfamiliar and unnatural audio is inappropriate for this type of study. A common alternative is to use speech (Begault et al., 2000; Møller and Sørensen, 1996; Mattila, 2001). Whilst this is more ecological than noise, it is relatively band limited and lacks low frequencies especially. The stimuli used should represent examples of everyday binaural audio and as such should elicit the same or at least similar perceptual characteristics (Simon, Zacharov and Katz, 2016). For

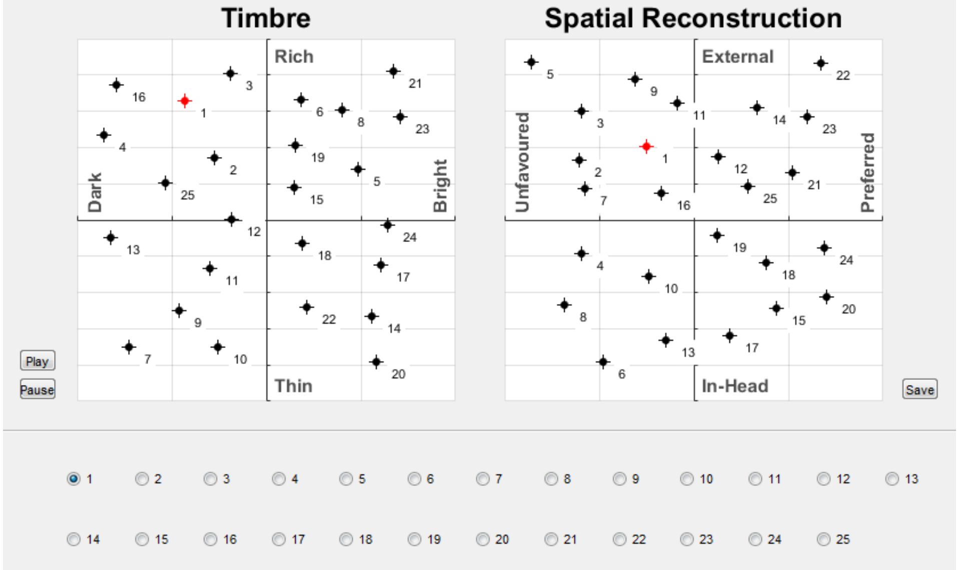


Figure 3.10: The graphical interface used by participants to rate audio stimuli. Selection of stimuli was made using the radio buttons at the bottom of the interface. A corresponding marker would be highlighted on each graph and participants were required to click and drag the markers using a computer mouse to where they felt was appropriate.

example, whilst it would be quite unusual to discuss the brightness of radio static, a similar discussion about the sound of a piano would be relatively common.

With this in mind, approximately a minute and a half of music was composed by the author in a jazz style using a range of non-reverberant VST MIDI samplers. These included a stereo drum set, stereo piano, flute, trumpet, trombone and double bass for a well balanced mix covering a large range of frequencies. The ensemble was binaurally spatialised by directly convolving individual audio sources with HRTFs spaced at 45° increments around the horizontal plane, starting at 0°, and summing the results for each ear individually. Stereo sources were convolved with adjacent HRTFs to mimic phantom source phenomena in real world listening (Pulkki, 1997). 20 binaural signals were produced using each of the 20 HRTF measurement sets admitted to the SADIIE binaural database, discussed in Section 3.3.2.

In addition, 5 anchor stimuli were presented: 4 stereo mixes and 1 mono mix. The stereo mixes were rendered by amplitude weighting the audio stems based on a constant power panning law. The mono mix was rendered by the equal summation of all sources. Of the 4 stereo mixes, 3 were degraded by the same filters as the example stimuli and as depicted in Fig. 3.9. 1 stereo mix was left unfiltered to simulate a rich signal. The mono mix simulated a non-spatial signal.

All 25 stimuli, normalised to RMS level, were presented to each participant in a random order over Beyerdynamic DT 990 PRO open-back headphones[17] via a Fireface UCX interface[18]. Participants were asked to adjust the volume of playback to a comfortable listening level i.e. a level at which they would normally listen. Personalised headphone equalisation was used in each case. Equalization filters previously measured as part of the SADIIE database, presented in Section 3.3.5, were used. The same pair of headphones were used for this test as were measured for the database.

3.4.4 Results

Each subject's ratings were normalized about 0 with respect to mean value and standard deviation as recommended by ITU-R BS.1284-1 (ITU, 2003a). The combined ratings for each attribute were then normalized to a maximum absolute value of ±1. The responses to each stimuli are presented as box plots in Fig. 3.11 in order of mean preference. Stimuli are identified by either the anchor they represent, or by the subject whose HRTF's were used to render the signals. To preserve anonymity, human subjects are referred to as H[3-20]. Included on the plots are the ratings given to each stimuli by the owner of the respective HRTFs. This is referred to as the *Personal Rating.*

The average values and narrow ranges of the *thin*, *dark* and *bright* anchors (stereo tracks) validate the participants understanding of the attributes. The results of the *mono* anchor are surprisingly optimistic. Despite averaging amongst the lowest scores in both externalisation and preference, the confidence intervals and error bars extend to well within the ranges of higher scoring HRTF sets. This indicates the significance of timbre in rendering systems.

The responses to each stimuli were tested for normality with a Lilliefors test which failed to reject the null-hypothesis of normality at the 5% significance level. The significance of the ratings given to each stimuli for each attribute were explored by one-way repeated measure ANOVA with post-hoc analysis. Violations of the assumption of sphericity were identified by Mauchly's tests and Greenhouse-Geiser corrections were applied in the calculations of p-values. Results are presented in Table 3.4. A p-value of below 5% indicates that it may be said with 95% confidence that the average results do truly vary. Greatest significance is seen with respect to preference, followed by timbral attributes: brightness and richness. A significant

[17] europe.beyerdynamic.com/dt-990-pro.html
[18] rme-audio.de/fireface-ucx.html

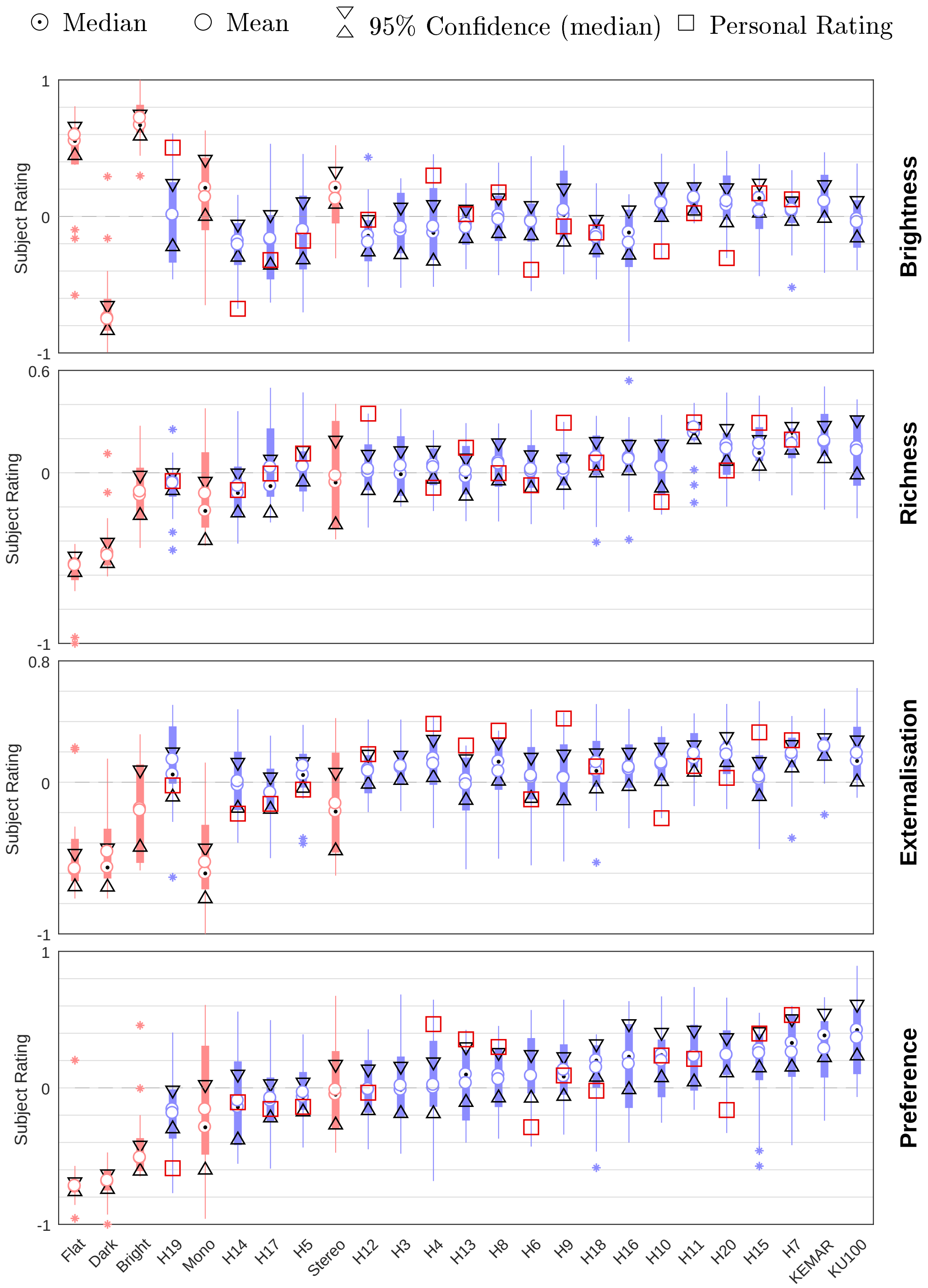


Figure 3.11: Subject ratings of stimuli in order of mean preference. A personal rating reflects a subject's rating of their own measurements.

Table 3.4: A Greenhouse-Geiser estimation of ϵ and the results of a corrected one-way repeated measure ANOVA applied to the ratings given to each stimuli with respect to attribute. A p-value below 5% was considered significant.

Attribute	Greenhouse-Geiser estimation of ϵ	p-value (with Greenhouse-Geiser correction) (%)
Brightness	0.418	1.1
Richness	0.435	0.78
Externalisation	0.448	9.1
Preference	0.461	0.065

Table 3.5: Significant differences found between individual stimuli with respect to preference, richness and brightness attributes. The values shown are the p-values (%) from a post-hoc pairwise mean comparison test using Tukey's Honestly Significant Difference procedure. A value below 5% was considered significant. The stimuli shown vertically received a higher rating in each case.

	Preference			Richness	Brightness		
	H19	H8	H17	H19	H18	H14	H16
KU100	0.20	4.2	4.7				
H20	1.0						
H7	1.4						
KEMAR	1.8				4.3		
H9	3.5						
H11				2.2	0.64	1.8	2.1
H15				4.4			
H10						0.95	

difference is not seen with respect to externalisation. Together with Fig. 3.11 these results reinforce that timbre must play a considerable role in HRTF selection.

A post-hoc pair-wise comparison of the mean ratings given to each stimuli was undertaken using Tukey's Honestly Significant Difference test procedure. This revealed 7 significant differences between stimuli with respect to preference, 2 with respect to richness and 5 with respect to brightness. 0 significant results were seen with respect to externalisation. A summary is given in Table 3.5. By virtue of the fact that diffuse-field equalised HRTFs were used throughout the test, these differences in timbral and spatial features must be attributed to the individual spectral notches of the HRTFs and not to any general frequency response of the individual.

The correlation of brightness, richness and externalisation with respect to preference is plotted in Fig. 3.12. The graph directly compares the attribute ratings given by each participant to each stimuli. Anchors and anomalies identified in Fig. 3.11 are excluded from this plot. Second order polynomial lines of best fit indicate a positive

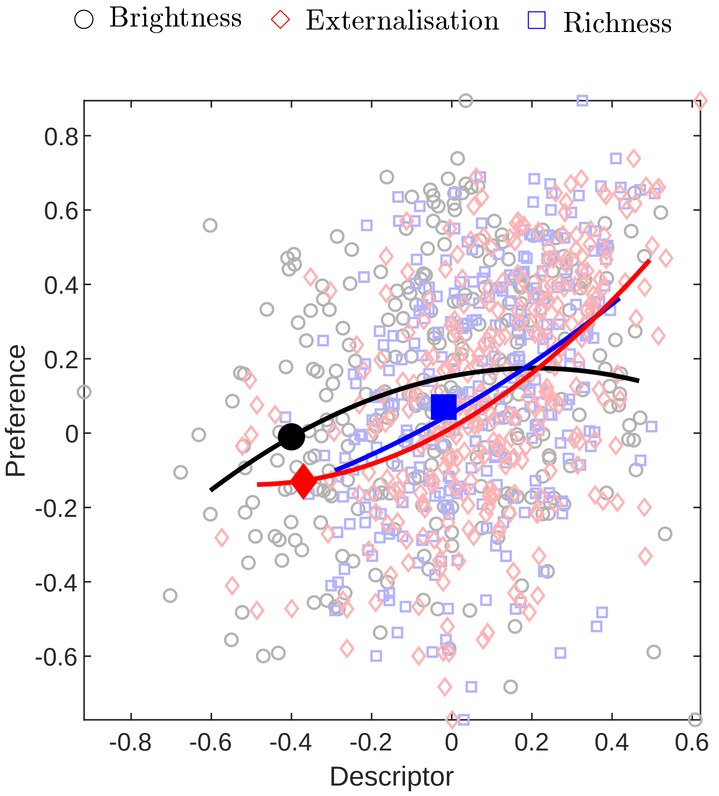


Figure 3.12: A comparison of the ratings given to each stimulus by each subject. Brightness, Richness and Externalisation ratings are plotted against Preference ratings to show correlation. Both Richness and Externalisation show a positive correlation whilst an overall preference for a more natural Brightness is indicated. Note that outliers have been excluded from this plot.

Table 3.6: Pearson's correlation coefficient values calculated between attributes (excluding anchors and anomalies).

	Brightness	Brightness*	Richness	Externalisation	Preference
Brightness	1	0.86	0.15	0.17	0.23
Brightness*	(0.86)	1	0.19	0.21	0.27
Richness	(0.15)	(0.19)	1	0.37	0.40
Externalisation	(0.17)	(0.21)	(0.37)	1	0.46
Preference	(0.23)	(0.27)	(0.40)	(0.46)	1

correlation between richness, externalisation and preference. A slight preference for neutral brightness can be seen.

Analyzing the correlation of such a mapping of brightness to preference is challenging due to its non-linearity. A new parameter is therefore defined: *brightness**, that is the deviation of the brightness rating from an optimal value of 0.2 (read from Fig. 3.12):

$$\text{brightness}^* \triangleq -|\text{brightness} - 0.2| \tag{3.1}$$

By doing this, correlations of brightness* may be considered such that a higher rating is indicative of preference. A summary of the Pearson's correlation coefficient values for each pair of attributes is given in Table 3.6.

3.4.5 Discussion

From these results three key findings are presented. The first is that there are significant differences in the attribute ratings given to particular HRTF sets by a general audience. The second is that there exists some correlation between these attributes, for example that of externalisation, richness and preference, but that this correlation is not high. The third is that individual measurements were not perceived by subjects to be of optimal performance.

Overall, it is the HRTFs of the dummy mannequins that are most preferred. Surprisingly, it is the HRTFs of the head without shoulders (the KU100) that received the highest rating for preference. This is despite averaging similar or less favourable ratings than the measurements of KEMAR, H11, H20 and H7 with respect to all other attributes, according to general correlations shown in Fig. 3.12. It it likely, therefore, that there exist other factors not identified in this study that have a stronger influence on overall preference.

There are two major differences between the measurements of the dummy heads and the human participants: movement and microphones. Despite best efforts, some movement is inevitable with human subjects, the binaural heads on the other hand remain perfectly still. Larger in-built microphones were also utilised for the dummy heads. Significant differences between the preference ratings of human measurements (H20, H7, H9) and (H19), however, indicate that microphone selection alone cannot be the sole cause of preference. It is therefore proposed that the stillness with which a participant sits could impact the quality of measurement and hence the performance of the HRTFs.

The strongest relevant correlation is seen between the attributes externalisation and preference with a Pearson's correlation coefficient value of 0.46. However, it is noted that this does not indicate a particularly strong correlation. A similar result is seen between richness and preference whilst brightness* appears to correlate relatively poorly with all other attributes. A slight preference for brighter timbres over darker timbres may be interpreted from Fig. 3.12. This is confirmed by the positive correlation coefficient ($\rho = 0.23$, see Table 3.6) calculated between brightness and preference.

These values indicate a slight correlation between the attributes tested in this study. However, it is in fact of more interest to note the lack of strong correlation. Such results show that HRTFs may be rated as highly preferable regardless of their timbral or spatial characteristics. Therefore, selection of an optimal HRTF remains a complex task and likely depends on the application.

A key result is the randomness with which a participant rated their own measurements. Andreopoulou and Katz, 2016a comments on the repeatability and hence reliability (or lack thereof) of HRTF ratings. They conclude that although in general HRTF rating is a difficult task, repeatability of results is significantly higher at the extreme ends of the response scales. This indicates that had individual measurements significantly out-performed non-individual measurements, as one might expect, this would have resulted in consistent ratings. However, this was not the case.

3.5 Summary

State of the art methods for HRTF measurement and post-processing of data have been discussed in context of the acquisition of the SADIIE database. The database represents one of the largest measured HRTF datasets for both human subjects and dummy mannequins currently available.

A listening test was conducted in which participants were asked to evaluate both individual and non-individual binaural renderings of a jazz ensemble on four scales: brightness, richness, externalisation and preference. Results show significant differences in the ratings given to particular HRTF sets at the 95% confidence interval. An overall preference for the measurement set of the KU100 dummy head is seen, followed closely by the the measurement set of the KEMAR mannequin. A slight preference is shown for rich and external stimuli of a neutral/slightly bright timbre.

Very little correlation is seen with respect to the responses given to stimuli generated with individual HRTFs and no obvious link is found between a person's individual measurements and their general preference towards them. This is despite almost certainly improving the localisational accuracy of the binaural reproduction (Seeber, Fastl and Others, 2003; Møller and Sørensen, 1996; Wenzel et al., 1993). It therefore seems reasonable that individual measurements are not always recommended throughout the literature (Nicol et al., 2014; Usher and Martens, 2007).

Within the test presented in Section 3.4 only 1 participant (H9) rated their individual measurements as sounding the most external. No participants rated their individual measurements as the most preferred and 81% of participants preferred the KU100 HRTFs to their own. That being said, it is not clear what it is about the KU100 HRTFs that are most appealing. Future work may consider refining the HRTF measurement technique in an attempt to rule out factors such as subject movement. This is somewhat addressed within the new HRTF measurement procedure introduced in Chapter 6.

The results of this test lead to questions regarding the future of HRTF measurement and binaural rendering. Source localisation, shown to improve with individual measurements, must be carefully balanced against timbral and spatial qualities of competing measurement sets. It is proposed that the exploration of faster HRTF measurement may lead to improved individual HRTF sets by ruling out errors in the measurements due to subject movement. Due to the improvements seen in localisation with individual measurement it is possible that these new, more accurate, individual measurements could eventually outperform the dummy mannequin measurements. One such technique is presented in Chapter 6. Alternatively, new approaches to binaural rendering may be considered within which HRTF performance should be re-evaluated. Binaural Ambisonic rendering is hence discussed in Chapter 4 with further listening tests evaluating the performance of HRTFs within this context presented in Chapter 7.

4

Ambisonic Reproduction and Binaural Rendering

Chapter Overview

This chapter summarises the entire process of binaural Ambisonic rendering. A detailed and comprehensive summary of Ambisonics and its integral encoding/decoding methods is given on both a mathematical and conceptual level. Competing industry standards are highlighted and recommended workflows are identified. The reason for the well known sweet spot is explained through spatial aliasing and state of the art methods of binaural rendering are presented.

4.1 Introduction

Having considered the impact of HRTFs on the perception of spatial audio through direct convolution, it is now appropriate to consider the more common ways in which HRTFs are used within state of the art binaural rendering technologies. Subjective performance can vary significantly between rendering schemes (Pike and Melchior, 2013; Pike, Melchior and Tew, 2016) and therefore the performance of HRTFs should best be considered within the correct context.

Direct convolution of source material with a HRTF is not necessarily the most convenient or efficient form of presenting spatial audio over headphones. The approach is computationally expensive and requires either dense sets of HRTFs to pan a source or complex interpolation algorithms to synthesise accurate filters from sparse data sets. Similarly, the technique does not lend itself well to head tracked applications as this simply increases the movement of sources relative to the listener. Further, the computational cost varies with the number of sources and as such rendering diffuse or wide spread sources becomes challenging.

An alternative approach using Ambisonics is therefore considered (Mckeag and McGrath, 1996; Noisternig et al., 2003b). Ambisonics is a method of spatial audio reproduction originally developed for loudspeakers that has since been reimagined for binaural reproduction. It utilizes a fixed set of HRTFs to render spatial audio and is capable of supporting head-tracking through inexpensive soundfield transformation matrices prior to decoding. As such, it is particularly applicable to VR applications and hence highly relevant to the current markets and industry. It is important then to understand how this process works and how HRTF filters fit into this workflow. Optimizations may then be realised based on this understanding and further objective and subjective evaluations.

4.2 Background

Despite being a frequent topic of spatial audio literature and a foreseeable future of immersive audio, Ambisonics is a subject that is still commonly misunderstood by many of those involved in the related fields. Complex origins buried deep within mathematical formulae may be partially to blame, but equally are the regular publications that continue to offer novel approaches to the soundfield reproduction technique. As a result, it has become difficult to define explicitly what *Ambisonics* is.

One must now consider a collection of techniques that each make use of a particular core set of functions known as spherical (3D) or cylindrical (2D) harmonics.

Ambisonics is an expandable, end-to-end mathematics-based approach to spatial audio reproduction over loudspeakers. It encompasses the encoding, storage and rendering of directional auditory data over multiple dimensions. It formulates the spatial sampling of an infinitesimal soundfield such that it may be resynthesised by a finite number of point sources. A collection of alternative reading may be found in the following (Daniel, 2001; Malham, 2003b; Hollerweger, 2006; Hollerweger, 2008; Kearney, 2010; Ortolani, 2015).

Ambisonics was developed throughout the 1970s. It was first conceptualised (Gerzon, 1973) and then formalised (Fellgett, 1975; Gerzon, 1975a) with further work discussing microphone capture techniques (Gerzon, 1975b) and practical reproduction methods (Gerzon, 1980). Ambisonics is a flexible system that, unlike traditional surround sound techniques (such as Dolby Digital 5.1 (Dolby, 2016)), does not require any knowledge of the playback system (e.g. loudspeaker layout) during the encoding process (Herre et al., 2014). Ambisonics provides a generic representation of a soundfield such that it may be reproduced, with limitations, over almost any loudspeaker configuration.

Ambisonic format data is made up of a series of time-domain mono audio streams known as channels. (Not be confused with traditional ‘channel-based’ audio signal representations such as stereo (Qualcomm, 2015).) Ambisonic Channels represent specific orthogonal scaled portions of a soundfield as defined by multi-dimensional trigonometric-based harmonic functions.

The number of Ambisonic Channels depends on the *order* of Ambisonics being used. Higher orders require a greater number of channels and generally provide an increased spatial resolution at the cost of data storage, computation, and increased complexity (Bertet et al., 2013). Whilst acceptable to ignore higher order components it is generally not possible to accurately ‘upscale’ lower order components, although research into approximating such data has been considered via plane-wave decomposition/expansion, e.g. Berge and Barrett, 2010; Wabnitz, Epain and McEwan, 2011.

Using almost identical principles to the mid-side microphone technique (Dooley and Streicher, 1982) Ambisonic Channels may be weighted and summed together to spatially sample the soundfield with a known directivity and directionality. The resulting polar pickup patterns are often referred to as virtual microphones. The algorithms used to calculate these channel weightings are known as Ambisonic decoders and are used to determine the loudspeaker signals of an arbitrary playback array.

In its primitive form (referred to as basic/mode-matching/holophonic reproduction) Ambisonics is a soundfield reconstruction technique, similar to Wave Field Synthesis (Daniel, Nicol and Moreau, 2003). Loudspeaker signals are theoretically generated such that the harmonic mode excitations within the centre of a playback array are best matched with those of the original soundfield. However, the method encounters significant limitations. Due to the finite sampling and limited angular resolution of the Ambisonic signal, spatial aliasing prevents accurate reconstruction outside of an area of space in the center of the array known as the sweet spot. Its size depends on the order of Ambisonics and position of the source and is found to decrease with frequency often to radii smaller than the human head (Zotter, Pomberger and Frank, 2009; Daniel, Rault and Polack, 1998).

To counter these issues, channel weightings may be introduced to alter the directivity of the virtual microphones. Careful derivation of these weights can, amongst other things, maximise the concentration of energy in the direction of the intended source position. Whilst these adjustments distort the mathematical restoration of the original soundfield, perceptual improvements in terms of localisation, especially for off-centre listening, have been shown (Kearney, 2010; Heller, Benjamin and Lee, 2012).

4.3 Notation

The following notation styles are introduced to improve clarity and coherence:

- Lower case variables will index corresponding upper case variables (e.g. $m = 0, 1, ..., M$).

- Bold lower case variables (e.g. $\boldsymbol{a}$) will represent single-dimensional matrices. Bold upper case variables (e.g. $\mathbf{C}$) will represent multi-dimensional matrices.

- Variables superscripted by a dash (e.g. P') will denote the un-normalised form of that variable. Explicit normalisation will be shown (e.g. P^{N3D}).

- Angles, $(\varphi,\ \vartheta)$, $(\vec{v})$, are defined by a modified spherical co-ordinate system. Azimuth, φ, is measured in an anticlockwise direction where 0° lies straight ahead. Elevation (latitude), ϑ, is measured such that 0° lies on the transverse (horizontal) plane and positive values indicate an upward direction. Note that in some literature elevation is presented as the colatitude (measured downward from the north pole). Often, this will transpire as swapping *sine/cosine* terms relating to the elevation.

4.4 Spherical Harmonic Functions

4.4.1 Overview

The spherical harmonic functions are 3D functions derived from the solution of Laplace's equation (Haber, 2012). They are used as the basis for sampling the soundfield in a particular direction. Their 2D alternatives are the cylindrical harmonic functions, however, as it is most relevant to modern Ambisonic applications, the 3D case will be prioritised.

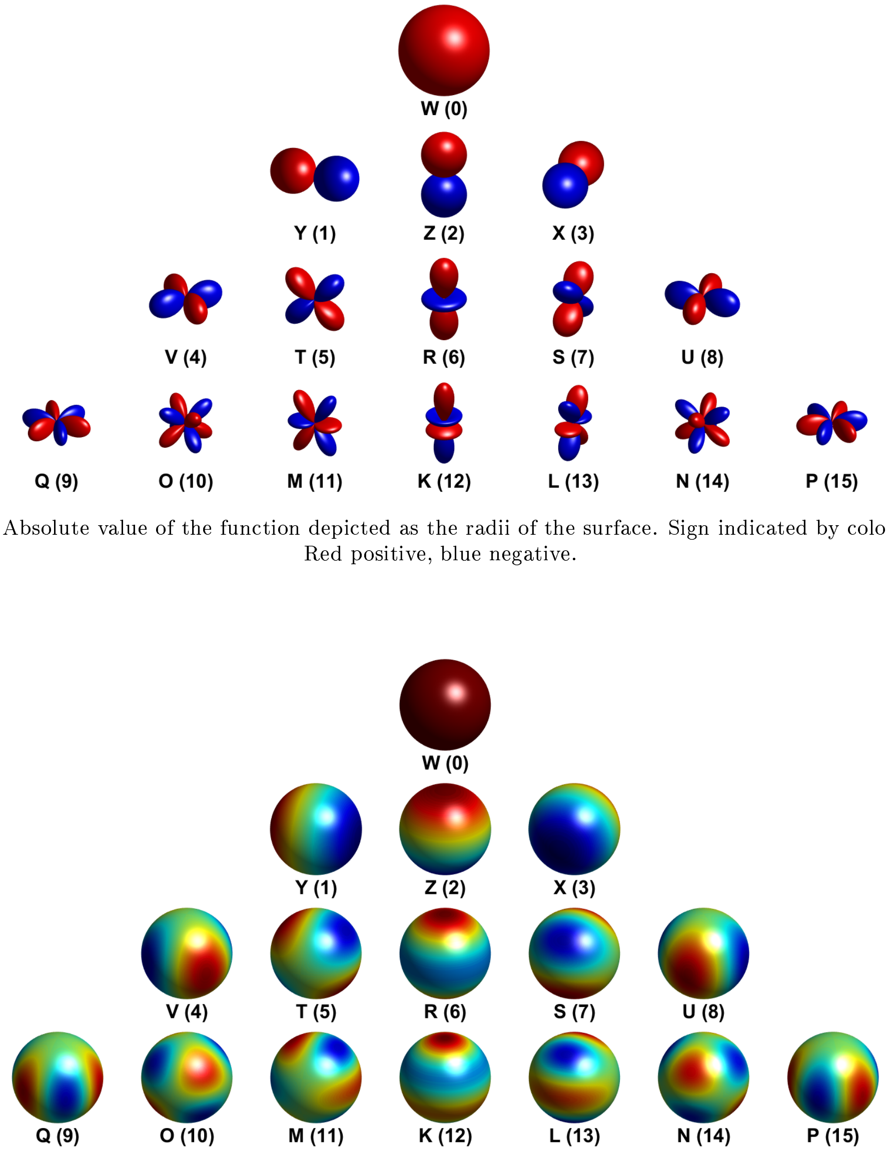
Spherical harmonics functions are defined by the multiplication of sinusoidal functions (defined by an azimuthal component) with associated Legendre polynomial functions (defined by an elevatory component) (Ceperley, 2016). They are commonly depicted in one of two ways, both shown in Fig. 4.1.

The harmonics are grouped by spatial sampling complexity and defined by parameters passed to the Associated Legendre Polynomials and sinusoidal functions during their derivation. The parameters are defined here as degree, $m \geq 0$, and index, $i = 0, 1, ...\ m$. A spin is also defined:

$$\sigma = \begin{cases} 1 & \text{if } i = 0 \\ \pm 1 & \text{if } i > 0 \end{cases} \tag{4.1}$$

related to the orientation of each function. Elsewhere in the literature it is common to omit σ and define $-m \leq i \leq m$. However, this notation is preferred as it highlights the symmetrical properties present in spherical harmonics.

In Fig. 4.1, an increasing degree is represented by row. Within each degree there are $2m + 1$ harmonic functions. It may be seen that the functions on the left observe



(a) Absolute value of the function depicted as the radii of the surface. Sign indicated by colour: Red positive, blue negative.

(b) Signed value of the function depicted by a graded colour map ranging from dark blue (most negative) to dark red (most positive).

Figure 4.1: The SN3D normalised spherical harmonic functions as are used in up to 3rd order Ambisonics labelled with alphabetical (FuMa) and Ambisonic Channel Number (ACN) notation as discussed in Section 4.5.4.

a 90° rotation about the z-axis compared to those on the right (e.g. ACN 1&3, 10&14). These harmonics are defined by the same index but opposite spin.

Ambisonic representations are defined by order, M, and include $(M+1)^2$ spherical harmonics components of degree $m \leq M$ i.e.

$$m = 0, 1, ... \ M, \quad 0 \leq i \leq m, \quad \sigma = \pm 1 \tag{4.2}$$

4.4.2 Notation

Unfortunately, notation within the literature surrounding these functions is far from standardised. Mathematical texts tend to favour the terms degree, l, then order, m, referring to the rows and columns of Fig. 4.1 respectively (Haber, 2012; Goldberg et al., 1967). However, this became confusing when Ambisonic literature began to use term 'order' to mean the order of Ambisonics (which in fact relates more directly to the mathematical 'degrees' or rows of spherical harmonics). Ambisonic literature then began to use the exact opposite terminology, order then degree, to refer to the same spherical harmonics. However, influential authors (e.g. Daniel, 2001; Zotter, 2009) were not able to agree on consistent alphabetical notation. This has resulted in the terms order and degree being termed m and n interchangeable throughout popular papers.

In an attempt to avoid confusion, the following standard is proposed and is used within this thesis:

- degree (m): referring to an individual row of Fig. 4.1
- order (M): referring to an order of Ambisonics that includes a collection of rows from Fig. 4.1
- index: (i): referring to the columns of Fig. 4.1

The term *index* has been chosen to replace the term *degree* in current Ambisonic nomenclature. The preferred notation leaves the term 'order' free to be used as the 'orders of Ambisonics' without getting mistakenly misinterpreted/confused with the indexing of the spherical harmonics. It also leaves the notation l free to be used for loudspeaker indexing. Further, matching the term 'degree' with the mathematical literature aids in the majority of technical derivations.

4.4.3 Derivations

The infinitesimal collection of spherical harmonics, ${Y^{\sigma}_{mi}}^{\mathrm{O3D}}$, of degree $m = 0, 1, ..., \infty$ form a complete orthogonal basis set such that

$$\int Y^{\sigma}_{mi}(\vartheta, \varphi) \ . \ Y^{\sigma'}_{m'i'}(\vartheta, \varphi) \ d\Omega = \delta_{mm'}\delta_{ii'}\delta_{\sigma\sigma'} \tag{4.3}$$

where $\delta_{a,b}$ denotes the Kronecker delta function which is 1 for $a = b$ and 0 otherwise and further there exists no other function, $f(\vartheta \ \varphi)$, that is orthogonal to all of $Y^{\sigma}_{mi}(\vartheta, \varphi)$ (Haber, 2012). Simply put, this means that the result of any function times another will integrate to 0 over the sphere. One could say that every function is equal and opposite to all others. Further, there exists no other function that is equal and opposite to those already included in the set.

Spherical harmonics are defined as follows:

$$Y^{\sigma}_{mi} = N^{\{\ldots 3\mathrm{D}\}}_{mi} P'_{mi}(\sin(\vartheta)) \times \begin{cases} \cos(i\varphi) & \text{if } \sigma = 1 \\ \sin(i\varphi) & \text{if } \sigma = -1 \end{cases} \tag{4.4}$$

$$= P^{\{\ldots 3\mathrm{D}\}}_{mi}(\sin(\vartheta)) \times \begin{cases} \ldots \end{cases} \tag{4.5}$$

Where a normalised Associated Legendre function is written

$$P^{\{\ldots 3\mathrm{D}\}}_{mi} = N^{\{\ldots 3\mathrm{D}\}}_{mi} P'_{mi} \tag{4.6}$$

and $N^{\{\ldots\}}_{mi}$ denotes the normalisation factor and P'_{mi} denotes the un-normalised Legendre Polynomial. Common normalisation factors include: SN3D (Schmidt Semi-Normalised), N3D (an orthogonal normalisation) and what will be referred to as O3D (an orthonormal normalisation)

$$N^{\mathrm{SN3D}}_{mi} = \sqrt{\epsilon_m \frac{(m-i)!}{(m+i)!}} \tag{4.7}$$

$$N^{\mathrm{N3D}}_{mi} = \sqrt{\epsilon_m \cdot (2m+1) \cdot \frac{(m-i)!}{(m+i)!}} \tag{4.8}$$

$$N^{\mathrm{O3D}}_{mi} = \sqrt{\epsilon_m \frac{(2m+1)}{4\pi} \frac{(m-i)!}{(m+i)!}} \tag{4.9}$$

where,

$$\epsilon_m = (2 - \delta_{i,0}) \tag{4.10}$$

and $\delta_{i,0}$ denotes the Kronecker delta function which is 1 for $i = 0$ and 0 otherwise. Note that the Schmidt Semi-Normalisation omits the Condon–Shortley phase factor of $(-1)^m$ commonly implemented in quantum mechanics.

Normalisation alters the relative scaling of each spherical harmonic function. The normalisations shown here weight the functions with respect to the surface integrals of each function squared over the unit sphere $r = 1$.

$$\int_{\boldsymbol{S}} (Y^{\sigma}_{mi})^2 dS = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_0^{2\pi} (Y^{\sigma}_{mi})^2 \cdot r^2 \cdot \cos(\vartheta) d\varphi d\vartheta \tag{4.11}$$

This is representative of the overall power of each function. Surface integration over the unit sphere sums the values of a function (in our case the (spherical harmonic functions)^2) at each point around the unit sphere weighted by the factor $\cos(\vartheta)$ to account for the areas of the infinitesimal surface elements. It is equivalent to calculating the volume between the surface $(Y^{\sigma}_{mi})^2 \cdot \cos(\vartheta)$ and the (φ, ϑ) plane.

N3D normalisation scales the spherical harmonic functions such that the surface integral of each function squared evaluates to 4π.

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_0^{2\pi} (Y^{\sigma}_{mi})^2 \cdot \cos(\vartheta) d\varphi d\vartheta = 4\pi \tag{4.12}$$

Orthonormal normalisation is very similar, but equates each integral to 1.

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_0^{2\pi} (Y^{\sigma}_{mi})^2 \cdot \cos(\vartheta) d\varphi d\vartheta = 1 \tag{4.13}$$

SN3D normalisation is slightly different in that it equates each integral to $\frac{4\pi}{2m+1}$ such that the summation of integrals within each degree, m, will equal 4π.

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \int_0^{2\pi} (Y^{\sigma}_{mi})^2 \cdot \cos(\vartheta) d\varphi d\vartheta = \frac{4\pi}{2m+1} \tag{4.14}$$

Generally speaking, SN3D normalisation reduces the weighting of higher order spherical harmonics whilst N3D and orthonormal normalisations maintain the overall power of each function. In practise, SN3D normalization may be preferred as it maintains a similar maximum value across all functions offering more consistent signal to noise ratios. Note that whilst reversible, depending on the decoding scheme implemented the normalisation of spherical harmonics may affect the reproduction stages of an Ambisonic renderer. More information is given in Section 4.6.10.

For completeness, the 2D cylindrical harmonics are defined for degree, $m \geq 0$, and spin, $\sigma = \pm 1$ (Daniel, 2001)

$$Y_m^\sigma = N_m^{\{...2\mathrm{D}\}} \begin{cases} \cos(i\varphi) & \text{if } \sigma = 1 \\ \sin(i\varphi) & \text{if } \sigma = -1 \end{cases} \tag{4.15}$$

and adapt the normalisation factors accordingly

$$N_m^{\mathrm{SN2D}} = 1 \tag{4.16}$$

$$N_m^{\mathrm{N2D}} = \sqrt{\epsilon_m} \tag{4.17}$$

$$N_{mi}^{\mathrm{O2D}} = \sqrt{\frac{\epsilon_m}{4\pi}} \tag{4.18}$$

4.5 Ambisonic Encoding

4.5.1 Overview

Ambisonic encoding covers the conversion of real world source and direction information to the intermediate storage format of Ambisonics (Ambisonic Format). In this form soundfields are decomposed into their spherical harmonic components and saved as Ambisonic Channels. This is achieved by one of two means:

- The manual weighting of individual audio sources by spherical harmonic coefficients

- The direct recording of real sound fields with directional microphones or beam forming technologies

4.5.2 Manual Weighting

To manually encode data into Ambisonic Format sources are weighted independently onto each Ambisonic Channel by a spherical harmonic coefficient. The value of each weight is determined by the value of the spherical harmonic function represented by that channel at the angle one wishes to encode the sound source at. This is done for all sources and all channels with multiple sources simply summed together on each channel.

$$\beta^{\sigma}_{mi} = \sum_{\text{for all } s} s \cdot Y^{\sigma}_{mi}(\vec{v}_s) \tag{4.19}$$

Where:

- β^{σ}_{mi} is the Ambisonic channel representing the spherical harmonic Y^{σ}_{mi},
- s is a source signal,
- $\vec{v}_s$ is the angle at which that source is being encoded.

4.5.3 Spatial Recording

To record directly into Ambisonic Format, one must aim to separately capture different regions of the soundfield with weightings that match each of the relevant spherical harmonic functions. This can either be performed directly with independent microphones, approximated using microphone arrays and summed responses (Gerzon, 1975b), or calculated with beam forming techniques.

For First Order Ambisonics (FOA), the relevant functions are omnidirectional and figure-of-eight shaped (Fig. 4.1) and can be captured directly using 1 omnidirectional and 3 figure-of-eight microphones respectively. Alternatively, and as is more commonly implemented, a tetrahedral array of microphones may be used to record the soundfield (Gerzon, 1975b). This is known as an A-Format Recording and is preferred as the configuration allows the microphones to be placed closer together and in a more co-incident fashion. An example layout is presented in Fig. 4.2. A simple set of equations are then used to convert the A-Format recordings into a more standard Ambisonic Format.

$$(W) \quad \beta^{1}_{0,0} = FLU + FRD + BLD + BRU \tag{4.20}$$

$$(X) \quad \beta^{1}_{1,1} = FLU + FRD - BLD - BRU \tag{4.21}$$

$$(Y) \quad \beta^{-}_{1,1}1 = FLU - FRD + BLD - BRU \tag{4.22}$$

$$(Z) \quad \beta^{1}_{1,0} = FLU - FRD - BLD + BRU \tag{4.23}$$

Where β^{σ}_{mi} is the Ambisonic channel representing the spherical harmonic Y^{σ}_{mi}. For

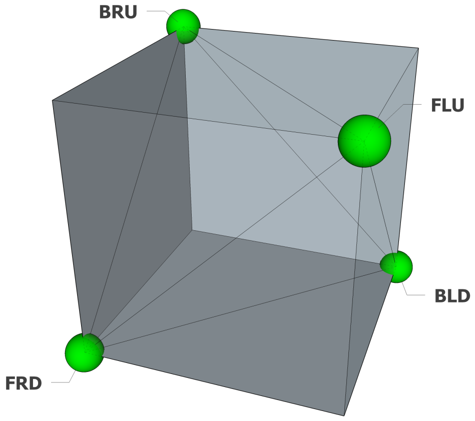


Figure 4.2: Example of a tetrahedral microphone configuration with microphone angles based on the vertices of a cube labelled as **F**ront, **B**ack, **L**eft, **R**ight, **U**p and **D**own.

completeness, FuMa notation (commonly used for FOA and discussed in section 4.5.4) is also given in brackets.

The Ambisonic signals should then be refined by means of frequency/phase correction filters to account for the slight spatial separation of the microphones. The responses of a set of exemplary filters for a microphone diaphragm spacing of 15mm are given in Fig. 4.3. Further details and the derivations of these filters are detailed by Farina, 2006 and in the patent for the original soundfield microphone (Graham and Gerzon, 1977). They may be calculated by

$$F_W = \frac{1 + \frac{1}{3}\frac{j\omega r}{c} - \frac{1}{3}\left(\frac{\omega r}{c}\right)^2}{1 + \frac{1}{3}\frac{j\omega r}{c}} \tag{4.24}$$

$$F_XYZ = \sqrt{6}\frac{1 + \frac{j\omega r}{c} - \frac{1}{3}\left(\frac{\omega r}{c}\right)^2}{1 + \frac{1}{3}\frac{j\omega r}{c}} \tag{4.25}$$

where:

- r is the radial distance of each capsule from the center of the array,
- $\omega = 2\pi f$ is the angular frequency,
- c is the speed of sound.

Unfortunately, for Higher Order Ambisonics (HOA) it is not possible to directly record the soundfield as microphones with complex enough pick up patterns do not exist. Instead, precise spherical arrays of microphones must be used in conjunction with beam forming techniques to approximate the correct pick up patterns. One such microphone is MH Acoustics' Eigenmike[19], see Fig. 4.4, which holds 32 microphone

[19] mhacoustics.com/home

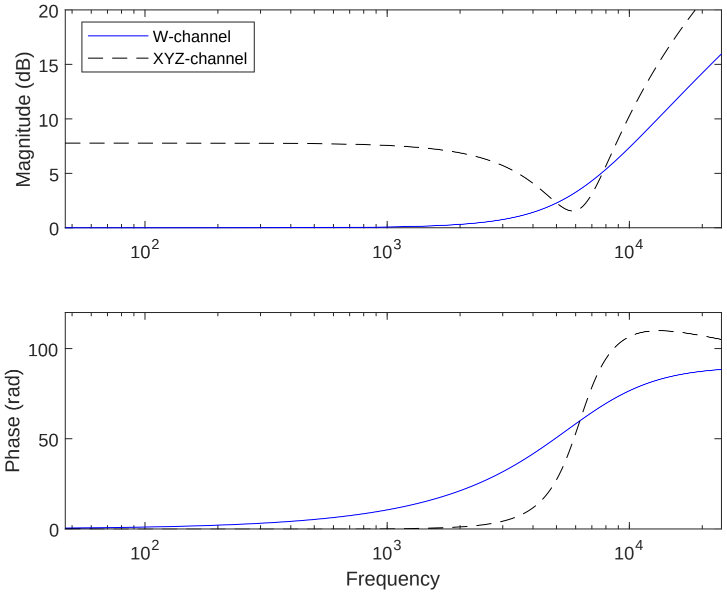


Figure 4.3: The theoretical responses of frequency and phase compensation filters for an A-Format tetrahedral microphone array of radius 15mm.

capsules and can approximate up to 4th order Ambisonic recordings. This topic have been studied extensively but is beyond the scope of this thesis. The reader is instead referred on towards relevant literature concerning the instabilities of such a spherical microphone array (Abhayapala and Ward, 2002; Gover, Ryan and Stinson, 2004) and a selection of solutions to the problem for: rigid spherical arrays (Meyer and Elko, 2002), directional microphone elements (Rahim and Davies, 1982; Meyer, 2001), double sphere arrays (Balmages and Rafaely, 2007; Jin, Epain and Parthy, 2014), open spherical shell arrays (Rafaely, 2008), double sided cone arrays (Gupta and Abhayapala, 2010) and by optimal sampling (Chardon, Kreuzer and Noisternig, 2014; Chardon, Kreuzer and Noisternig, 2015).

4.5.4 Competing Standards

There are a number of competing standards when it comes to Ambisonic Format data that relate to the sequential arrangement in which harmonic components are stored and the normalisation scheme used. Common examples include B-Format, ACN N3D and ACN SN3D where ACN (Ambisonic Channel Number) refers to the sequence of the Ambisonic Channels; N3D and SN3D refer to the normalisation; B-Format refers to a particular 1^{st} order format of specific sequence and normalisation. In general the standard used is entirely irrelevant, as long as the decoder is designed to accept the correct format.

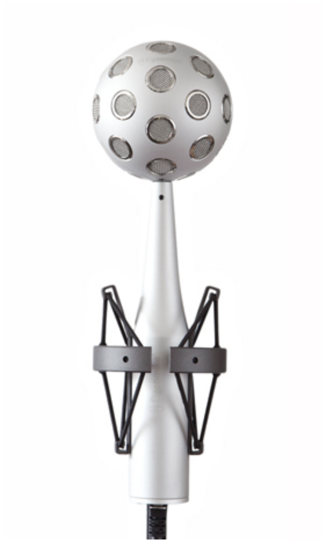


Figure 4.4: An Eigenmike: a 32 capsule microphone designed for HOA recording. (Image from **mhacoustics.com/home**)

In his original literature on the subject Gerzon proposed normalising the 1^{st} order harmonics such that X, Y and Z each had a gain of $\sqrt{2}$ in their directions of peak sensitivity. (W is omnidirectional and so has a gain of 1 in all directions.) This was in order that all four channels [W, X, Y, Z] each carried approximately equal average energy for any given soundfield recording (Gerzon, 1980) which in the days of magnetic tape recording was an important consideration to maximise the signal to noise ratios. This balance of amplitudes in the first 4 channels, alongside the ordering [W, X, Y, Z], is what is defined as B-Format.

A MaxN normalisation scheme is similarly but more generally defined such that each harmonic is scaled to have a maximum gain of ±1 (Daniel, 2001). Furse and Malham later proposed the generalised FuMa normalisation scheme (Malham, 2003a) which follows on from the MaxN normalisation scheme with an additional scaling of $\frac{1}{\sqrt{2}}$ on the W channel which provides backward compatibility with B-Format.

The B-Format alphabetic notation was also extended by Furse and Malham under the FuMa format (Malham, 2003a) (see Fig. 4.1) but the ordering or Ambisonic channels with respect to the normal order of the alphabet becomes quite illogical. It has now been almost entirely superseded by the ACN system, partially due to its expandability. A simple formula calculates the correct ACN of a harmonic based on its degree, m, and index, i

$$\text{ACN} = m \cdot (m+1) + \sigma i \tag{4.26}$$

4.6 Ambisonic Decoding

4.6.1 Overview

Ambisonic decoding covers the conversion of data from Ambisonic Format to a set of loudspeaker signals, the derivations of which may depend on the complete layout of each and every loudspeaker used for playback. There are a number of different ways in which this can be done and the following is by no means an exhaustive list. The key principle is in generating a decoding matrix, **D**, to weight and sum each ambisonic channel independently for each loudspeaker feed.

The 'standard' decoder may be assumed to be a Mode-Matching Pseudo-Inverse decoder as in (Poletti, 2000) which aims to restore exactly the original soundfield. Alternative decoders then each aim to preserve particular auditory qualities when presented with non-ideal irregular loudspeaker layouts. It follows therefore that by enforcing particular normalization schemes and regular layouts each method presented here will simplify to an identical decoding matrix. Fig. 4.5 is given at this time for reference and should become more clear as it is described throughout this section.

4.6.2 Loudspeaker regularity

The regularity of a loudspeaker array will in general effect a decoders' performance. A regular array will have loudspeakers that are evenly spaced. In 2D this is trivial: 3 loudspeakers that are spaced by 60°, 10 loudspeakers that are spaced by 36°and so on. However, in 3D finding appropriate configurations is a significant and ongoing challenge. Only 5 distributions are known to satisfy this criteria. These are known as the platonic solids.

Further, geometric regularity does not necessarily guarantee regularity in an Ambisonic sense. Daniel defines the requirement of regularity more precisely (Daniel, 2001, pp 175)

$$\frac{1}{L}\mathbf{C}^{\mathrm{N3D}} \cdot (\mathbf{C}^{\mathrm{N3D}})^T = I_K \tag{4.27}$$

Where I_K is the (K by K) identity matrix and the definitions of $\mathbf{C}$ and K are as given in Section 4.6.3. As a result, a dependency on the spherical harmonic coefficients and therefore Ambisonic order is seen (Moreau, Daniel and Bertet, 2006). That is to say, just because a distribution is regular for 1$^{\text{st}}$ order does not make it regular for 2$^{\text{nd}}$ or

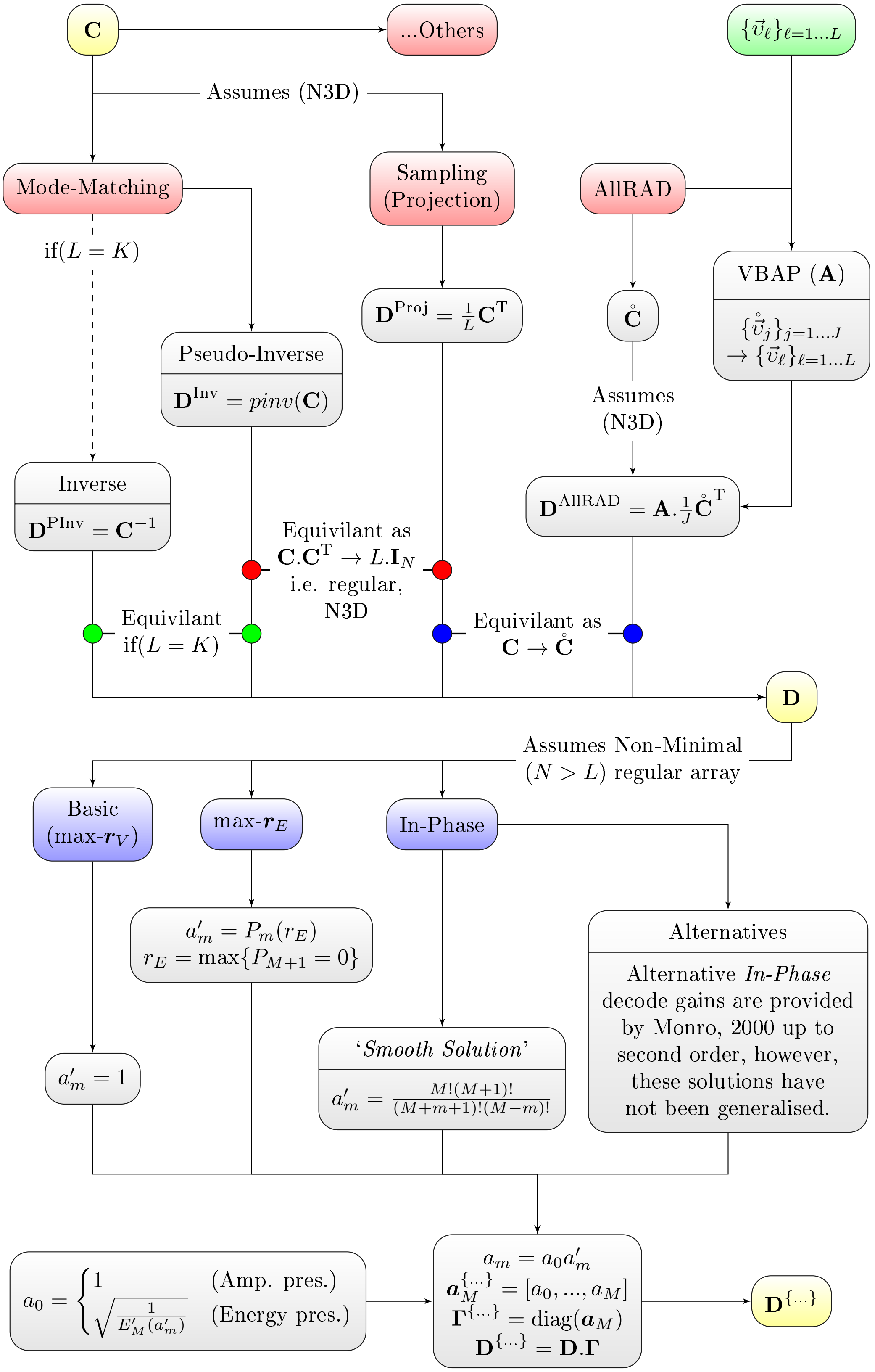

$\mathbf{C}$
...Others
$\{\vec{v}_\ell\}_{\ell=1...L}$
Assumes (N3D)
Mode-Matching
Sampling (Projection)
AllRAD
VBAP ($\mathbf{A}$)
$\{\mathring{\vec{v}}_j\}_{j=1...J} \rightarrow \{\vec{v}_\ell\}_{\ell=1...L}$
if($L = K$)
Pseudo-Inverse
$\mathbf{D}^{\text{Inv}} = pinv(\mathbf{C})$
$\mathbf{D}^{\text{Proj}} = \frac{1}{L}\mathbf{C}^{\text{T}}$
$\mathring{\mathbf{C}}$
Assumes (N3D)
Inverse
$\mathbf{D}^{\text{PInv}} = \mathbf{C}^{-1}$
$\mathbf{D}^{\text{AllRAD}} = \mathbf{A}.\frac{1}{J}\mathring{\mathbf{C}}^{\text{T}}$
Equivilant as $\mathbf{C}.\mathbf{C}^{\text{T}} \rightarrow L.\mathbf{I}_N$ i.e. regular, N3D
Equivilant if($L = K$)
Equivilant as $\mathbf{C} \rightarrow \mathring{\mathbf{C}}$
$\mathbf{D}$
Assumes Non-Minimal ($N > L$) regular array
Basic (max-$\boldsymbol{r}_V$)
max-$\boldsymbol{r}_E$
In-Phase
$a'_m = P_m(r_E)$
$r_E = \max\{P_{M+1} = 0\}$
Alternatives
Alternative *In-Phase* decode gains are provided by Monro, 2000 up to second order, however, these solutions have not been generalised.
‘*Smooth Solution*’
$a'_m = \frac{M!(M+1)!}{(M+m+1)!(M-m)!}$
$a'_m = 1$
$a_0 = \begin{cases} 1 & \text{(Amp. pres.)} \\ \sqrt{\frac{1}{E'_M(a'_m)}} & \text{(Energy pres.)} \end{cases}$
$a_m = a_0 a'_m$
$\boldsymbol{a}_M^{\{...\}} = [a_0, ..., a_M]$
$\boldsymbol{\Gamma}^{\{...\}} = \text{diag}(\boldsymbol{a}_M)$
$\mathbf{D}^{\{...\}} = \mathbf{D}.\boldsymbol{\Gamma}$
$\mathbf{D}^{\{...\}}$

Figure 4.5: Summary and workflow of common Ambisonic decoding techniques and matrix weighting solutions.

Name	Vertices	Regularity ($M \leq$ ...)
Tetrahedron	4	1
Octahedron	6	1
Cube	8	1
Icosahedron	12	2
Dodecahedron	20	2

Table 4.1: A summary of the Ambisonic orders to which the vertices of the platonic solids may be considered to be of regular distribution.

3$^{\text{rd}}$ order. This is summarised for the platonic solids in Table 4.1. Alternatively, t-designs have been shown to be regular up to Ambisonic order M given the condition $t \geq 2M + 1$ (Zotter, Frank and Sontacchi, 2010).

4.6.3 Definitions

It may be said that the ideal decoding of an Ambisonic Format signal is to derive the individualised input signals for an array of loudspeakers such that the encoded soundfield is exactly restored during playback (Berkhout, 1988). When discussing the decoding of Ambisonics it is therefore common to consider the following:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} \qquad \boldsymbol{\rho} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_L \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} Y_{00}^{1}(\varphi_1, \vartheta_1) & Y_{00}^{1}(\varphi_\ell, \vartheta_\ell) & \dots & Y_{00}^{1}(\varphi_L, \vartheta_L) \\ Y_{mi}^{\sigma}(\varphi_1, \vartheta_1) & Y_{mi}^{\sigma}(\varphi_\ell, \vartheta_\ell) & \dots & Y_{mi}^{\sigma}(\varphi_L, \vartheta_L) \\ \vdots & \vdots & \ddots & \vdots \\ Y_{Mi}^{\sigma}(\varphi_1, \vartheta_1) & Y_{Mi}^{\sigma}(\varphi_\ell, \vartheta_\ell) & \dots & Y_{Mi}^{\sigma}(\varphi_L, \vartheta_L) \end{bmatrix}$$

Where:

- $\boldsymbol{\beta}$ is a column vector of the individual Ambisonic Channels (W, Y, Z, X, etc.) of an Ambisonic Format file. It has length:

$$K = \begin{cases} 2M + 1 & (2D) \\ (M+1)^2 & (3D) \end{cases} \tag{4.28}$$

  where M is the order of Ambisonics being used and K is the number of Ambisonic Channels;

- $\boldsymbol{\rho}$ is a column vector, length L, of the decoded loudspeaker signals where L is the number of loudspeakers;

- $\mathbf{C}$ is a K by L matrix of the spherical harmonic value coefficients for each Ambisonic Channel in the direction of each loudspeaker. It is known as the *re-encoding* matrix.

The re-encoding Matrix is a particular example of a more general matrix of spherical harmonic coefficients

$$\mathbf{Y}_M^{\{\vec{v}...\}...} = [\boldsymbol{y}_M(\vec{v}_1), \boldsymbol{y}_M(\vec{v}_2), \boldsymbol{y}_M(\vec{v}_3), ...] \tag{4.29}$$

such that K spherical harmonic coefficients for each angle, $\vec{v}$, are referred to as the column vectors

$$\boldsymbol{y}_M(\vec{v}) = [Y_{0i}^{\sigma}(\vec{v}), ... \underbrace{Y_{mi}^{\sigma}(\vec{v}), ..., Y_{mi}^{\sigma}(\vec{v})}_{2m+1}, ... \underbrace{Y_{Mi}^{\sigma}(\vec{v}), ..., Y_{Mi}^{\sigma}(\vec{v})}_{2M+1}]^T \tag{4.30}$$

An L by K *decoding* matrix, $\mathbf{D}$, is further defined that comprises the Ambisonic Channel weightings for each loudspeaker signal for any given decoding strategy such that

$$\boldsymbol{\rho} = \mathbf{D} \cdot \boldsymbol{\beta} \tag{4.31}$$

Eqn. 4.31 is known as the *decoding equation*.

4.6.4 Velocity and Energy Vectors

Gerzon introduces two objective measures of soundfield restoration in his initial paper on periphony. These are known as the Makita (or velocity) vector, $\boldsymbol{r}_V$, and the Energy Vector, $\boldsymbol{r}_E$. The location of the Makita vector is 'the direction in which the head has to face in order that the interaural phase difference is zero'. Similarly, the direction of the Energy Vector is 'the direction the head has to face in order that there be no interaural amplitude difference at high frequencies' (Gerzon, 1980). These two measures relate directly to the human auditory system's localisation cues: ITD and ILD respectively.

The vectors may be calculated by summing together independent vectors pointing in the directions of each loudspeaker within the playback array. The length of each of these independent vectors should be proportional to the amplitude gain of that specific loudspeaker in the case of the Makita vector, or the energy gain in the case of the Energy Vector.

$$\boldsymbol{r}_v = \sum_{l=1}^{l=L} g_l^{\text{Amplitude}} \cdot \hat{\boldsymbol{v}}_l \tag{4.32}$$

$$\boldsymbol{r}_E = \sum_{l=1}^{l=L} g_l^{\text{Energy}} \cdot \hat{\boldsymbol{v}}_l \tag{4.33}$$

Where

- L is the number of loudspeakers;
- g is the loudspeaker gain;
- $\hat{\boldsymbol{v}}$ is the unity vector in the direction of a loudspeaker

Gerzon also defines a total playback amplitude/energy gain as the sum of the magnitude of each of the independent vectors. The length of the Makita and Energy vectors should ideally equal the total playback gains, as would be the case for a single loudspeaker reproduction. In the case that the length of the vectors do not equal the total playback gains this can lead to instability in the source imaging.

Gerzon proved that for FOA, assuming loudspeakers are placed in either diametrically opposite pairs (the *Diametric Decoder Theorem*) or spaced at equal angles (the *Regular Polygon Decoder Theorem*) then $\boldsymbol{r}_V$ and $\boldsymbol{r}_E$ will always remain coincident (Gerzon, 1977; Gerzon, 1992a). This is the case for regular point distributions other than a tetrahedral layout and can be extended to higher order Ambisonics providing there is always an adequate number of loudspeakers to support the respective decode (Heller, Benjamin and Lee, 2012).

By optimising for the velocity vector the restoration of low frequency ITD cues is optimised. Likewise, by optimising for the energy vector high frequency ILD cues are optimised.

4.6.5 Ambisonics as a Direct Panning Function

Generally, Ambisonics is considered as a discrete two-step process in which audio sources are first encoded into spherical harmonics and then decoded into loudspeaker signals. However, one may instead choose to consider only the total weighting between a source of given direction and its playback amplitude through each independent loudspeaker as a single step. As for most practical applications holophonic (i.e. the physically correct) sound field reproduction is not feasible over extended listening area anyway, Ambisonics is often seen as a somewhat complex 3D panner and is on occasion referred to as such throughout the literature (Poletti, 2000; Ward and Abhayapala, 2001; Zotter, Frank and Sontacchi, 2010; Zotter, Pomberger and Noisternig, 2012).

Zotter describes a continuous panning function, $g(\langle \vec{v}, \vec{v}_s \rangle)$, whose values describe the ideal weighting of a source on a continuous distribution of loudspeakers, $\{\vec{v}_\ell\}_{\ell=1\ldots\infty}$, given the source direction $\vec{v}_s$. Note that only the orientation of these virtual panning functions changes with panning direction $\vec{v}_s$; the shape does not.

As practical limitations demand, discretization of this panning function for a finite set of loudspeakers placed at $\{\vec{v}_\ell\}_{\ell=1\ldots L}$ results in a discrete set of gains $\hat{\boldsymbol{g}}$. Discrete gains, $\{\hat{g}_\ell\}_{\ell=1\ldots L}$, refer to the actual gains by which a source is weighted for each loudspeaker. These gains can be directly calculated from a normalised continuous panning function as described by Zotter and Frank, 2012, however, will only be mathematically justifiable for a regular layout. For irregular layouts alternative decoding approaches should be taken as are discussed later in this Chapter.

4.6.6 Mode Matching

The Mode-Matching (or holophonic) methodology is defined with the aim to restore the original soundfield within the reproduction array. The loudspeaker signals, $\boldsymbol{\rho}$, are calculated by equating the summed spherical harmonic mode excitations of the L loudspeakers to the spherical harmonic mode excitations of the original soundfield (Ambisonic Channels), $\boldsymbol{\beta}$, (Zotter, Pomberger and Noisternig, 2012; Poletti, 2000). The error is then minimised in a least squared sense by a matrix inversion/pseudo-inversion.

The ideal reconstruction (known as the *re-encoding equation*) may be written

$$\boldsymbol{\beta} = \mathbf{C} \cdot \boldsymbol{\rho} \tag{4.34}$$

Eqn. 4.34 represents the summed *re-encoding* of each loudspeaker signal as a source in the direction $\vec{v}_\ell$ into Ambisonic Format via the re-encoding matrix, $\mathbf{C}$, equated to the original Ambisonic Format signal, $\boldsymbol{\beta}$. By simple rearrangement, it follows that the loudspeaker signals may be derived

$$\boldsymbol{\rho} = \mathbf{C}^{-1} \cdot \boldsymbol{\beta} \tag{4.35}$$

such that

$$\mathbf{D}^{\mathrm{Inv}} = \mathbf{C}^{-1} \tag{4.36}$$

This is precisely referred to as Inverse Mode-Matching. Taking the inverse of a matrix is, however, only possible if the matrix is square i.e. in the case of $\mathbf{C}$ if $L = K$.

Alternatively, in the case $L \neq K$, one should consider the Moore-Penrose Pseudo-Inverse of $\mathbf{C}$ (Penrose and Todd, 1955). Taking the Pseudo-Inverse is an approach that extends the principles of inverting a matrix to non-square matrices (Courrieu, 2008; Golub and Kahan, 1965). This is precisely referred to as Pseudo-Inverse Mode-Matching.

$$\mathbf{D}^{\mathrm{PInv}} = \mathbf{C}^{\dagger} \tag{4.37}$$

where $\mathbf{C}^{\dagger}$, the Pseudo-Inverse of $\mathbf{C}$, is defined as

$$\mathbf{C}^{\dagger} = \begin{cases} \mathbf{C}^T \cdot (\mathbf{C} \cdot \mathbf{C}^T)^{-1} & L \geq K \\ (\mathbf{C}^T \cdot \mathbf{C})^{-1} \cdot \mathbf{C}^T & L < K \end{cases} \tag{4.38}$$

Conveniently, for square matrices, $\mathbf{D}^{\mathrm{PInv}} \equiv \mathbf{D}^{\mathrm{Inv}}$. A Pseudo-Inverse Mode-Matching decoder may therefore always be implemented irrespective of the matrix dimensions without any fear of a loss of accuracy.

The perceptual quality of a Mode-Matching decoder depends in part on the condition number of $\mathbf{C}$ which can be related to the regularity of the loudspeaker layout. Hollerweger, 2006 describes how the energy vector of an Ambisonic decode will only align with the sources encoded direction if $[L > K]$ and if the loudspeaker array is either regular, or semi-regular (i.e. the more general case that $(\mathbf{C} \cdot \mathbf{C}^T)$ is at least diagonal) (Daniel, 2001). For a short summary of solutions to the discrete spherical harmonic transform, its interpolation and approximation the reader is referred to Noisternig, Zotter and Katz, 2011.

4.6.7 Projection

The derivation of projection decoding originates from the simplification of Pseudo-Inverse Mode-Matching decoding for regular loudspeaker configurations and N3D (N2D (2D)) normalisation. It can be shown that where $L \geq K$ the pseudo-inverse of matrix $\mathbf{C}$ is trivialised as the term $(\mathbf{C} \cdot \mathbf{C}^T)$ becomes the diagonal matrix

$$(\mathbf{C}.\mathbf{C}^T) = \begin{cases} L \cdot I_{2m+1} & (2D) \\ L \cdot I_{(m+1)^2} & (3D) \end{cases} \tag{4.39}$$

where I_K is the (K by K) identity matrix. The result is the simplified decoding matrix

$$\mathbf{D}^{\text{Proj}} = \mathbf{C}^T \cdot (\mathbf{C} \cdot \mathbf{C}^T)^{-1} \tag{4.40}$$

$$= \mathbf{C}^T \cdot \begin{bmatrix} L_{[1,1]} & \dots & 0_{[1,K]} \\ \vdots & \ddots & \vdots \\ 0_{[K,1]} & \dots & L_{[K,K]} \end{bmatrix}^{-1} \tag{4.41}$$

$$= \frac{1}{L} \cdot \mathbf{C}^T \tag{4.42}$$

Projection decoding represents the spatial sampling of the Ambisonic Channels without any consideration of the placement of any other loudspeaker. This makes it a somewhat more robust and reliable technique in the face of irregular loudspeaker layouts. However, the advantages in matrix stability must be weighed up against the possible directional distortions and energy balance issues of a rendered source (Hollerweger, 2006; Zmölnig, 2002).

4.6.8 ALLRAD

All Round Ambisonic Decoding (AllRAD), presented by Zotter and Frank, 2012, applies principles of All Round Ambisonic Panning (AllRAP) to general Ambisonic decoding. The technique first derives a decoding matrix for a theoretical, perfectly regular loudspeaker layout suitable for the order of Ambisonics being implemented. The theoretical loudspeaker signals are then treated as virtual sources and are panned about any available playback rig using a VBAP approach (Pulkki, 1997). Note that as the real playback rig becomes more similar to the theoretical virtual rig this method also simplifies to that of a standard Ambisonic decoder and therefore ideally a projection decoder due to the regular layout.

Individually, neither Ambisonics nor VBAP provide optimally flexible solutions for loudspeaker playback. Ideal Ambisonic rendering requires very specific loudspeaker layouts whilst VBAP suffers from inconsistent source energy spread and arguably less flexible encoding/decoding procedures (Zotter, Frank and Sontacchi, 2010). By combining the techniques, however, both methods may be taken advantage of. The downside is that this technique somewhat ignores the holophonic rendering possibilities of a standard Ambisonic reproduction and instead treats the intermediate virtual loudspeaker signals as the result of a simple panning function.

Spherical designs or t-designs (discussed by Chen, 2009; Bajnok, 1991; Hardin and Sloane, 1996) are suggested as the most appropriate regular loudspeaker layouts. Decoding to such an array is trivial; referring to Eqn. 4.42 the decoding and re-encoding matrices for the virtual rig are defined as follows

$$\mathring{\mathbf{D}} = \frac{1}{J}\mathring{\mathbf{C}}^{\mathrm{T}} \tag{4.43}$$

$$\mathring{\mathbf{C}} = [\boldsymbol{y}_M(\mathring{\vec{v}}_1), \boldsymbol{y}_M(\mathring{\vec{v}}_j), \dots \boldsymbol{y}_M(\mathring{\vec{v}}_J)] \tag{4.44}$$

Where

- J denotes the number of virtual loudspeakers, indexed by j;
- Intermediate variables relating to the virtual loudspeakers are accented by a ring

Rendering the J virtual sources over a real playback rig of L loudspeakers requires an L by J VBAP gains matrix, $\mathbf{A}$, such that

$$\mathbf{D}^{\mathrm{AllRAD}} = \mathbf{A} \cdot \frac{1}{J}\mathring{\mathbf{C}}^{\mathrm{T}} \tag{4.45}$$

$$\mathbf{D}^{\mathrm{AllRAD}} = \mathbf{A} \cdot \mathring{\mathbf{D}}^{\mathrm{AllRAD}} \tag{4.46}$$

Details of the derivation of matrix $\mathbf{A}$ may be found from Zotter and Frank, 2012; Pulkki, 1997.

4.6.9 Alternative Techniques

Ambisonics is by no means limited to the basic decoding techniques presented here. Some alternatives specifically developed to directly compensate for the limitations of irregular arrays are briefly listed here for the reader's convenience.

Energy Preserving

Energy Preserving decoding was first introduced by Zotter, Pomberger and Noisternig, 2012. It is a technique that addresses the dependence of total playback energy on source panning direction for Pseudo-Inverse and Projection decodes and irregular loudspeaker layouts.

Eqn. 4.31 and Eqn. 4.34 stipulate that

$$\mathbf{C} \cdot \mathbf{D}^{\mathrm{PInv}} = \mathbf{I} \tag{4.47}$$

where $\mathbf{I}$ is an Identity matrix. The Singular Value Decomposition (SVD) of re-encoding matrix, $\mathbf{C}$, may be found

$$\mathbf{C} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}} \tag{4.48}$$

where:

- $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices,
- $\mathbf{V}^{\mathrm{T}}$ is the transpose of $\mathbf{V}$,
- $\mathbf{S}$ is a diagonal matrix of non-negative real numbers.

Hence, the decoding matrix, $\mathbf{D}$, may be calculated

$$\mathbf{D}^{\mathrm{PInv}} = \mathbf{V}\mathbf{S}^{-1}\mathbf{U}^{\mathrm{T}} \tag{4.49}$$

$$\mathbf{D}^{\mathrm{Proj}} = \left(\frac{1}{L}\right)\mathbf{V}\mathbf{S}\mathbf{U}^{\mathrm{T}} \tag{4.50}$$

It is shown by Zotter, Pomberger and Noisternig, 2012 how variations in playback energy are attributed exclusively to the range of singular values in $\mathbf{S}$. A decode matrix is therefore proposed (Zotter, Pomberger and Noisternig, 2012, Eqn. 30)

$$\mathbf{D}^{\mathrm{EP}} = \mathbf{V}\mathbf{U}^{\mathrm{T}} \tag{4.51}$$

which removes the singular values.

It is suggested here to include a linear scaling factor

$$\mathbf{D}^{\mathrm{EP}} = \sqrt{\frac{1}{L}}\mathbf{V}\mathbf{U}^{\mathrm{T}} \tag{4.52}$$

such that for regular arrays and N3D normalisation

$$\mathbf{D}^{\mathrm{EP}} \equiv \mathbf{D}^{\mathrm{Proj}} \equiv \mathbf{D}^{\mathrm{PInv}} \tag{4.53}$$

Constant Angular Spread

Constant Angular Spread decoding (Epain, Jin and Zotter, 2014) introduces a completely new analytical method for deriving decoding matrices. As a strategy, it aims to preserve: the total playback energy; the alignment of energy vector $\vec{r}_E$ with source direction; the angular spread of energy with respect to panning direction.

A set of optimal target loudspeaker gains

$$\hat{\mathbf{G}}^{\text{opt.}} = [\hat{\boldsymbol{g}}(\vec{v}_1), \hat{\boldsymbol{g}}(\vec{v}_q), \ldots \hat{\boldsymbol{g}}(\vec{v}_Q)] \tag{4.54}$$

are first derived for a large number (Q) of source directions by means of a *completely independent method* (Multiple-Direction Intensity Panning). A decoding matrix, $\mathbf{D}^{\text{CAS}}$, is then analytically computed such that the *mismatch* between $\hat{\boldsymbol{G}}^{\text{opt.}}$ and $\hat{\boldsymbol{G}}^{\text{CAS}}$ is minimised in the least-square sense where $\hat{\boldsymbol{G}}^{\text{CAS}}$ are the loudspeaker gains derived from the Ambisonic encoding of source directions, $\{\vec{v_q}\}_{q=1\ldots Q}$, and subsequent decoding of the Ambisonic channels through decoding matrix $\mathbf{D}^{\text{CAS}}$.

$$\hat{\boldsymbol{G}}^{\text{CAS}} = \mathbf{D}^{\text{CAS}} \cdot \mathbf{Y}_M^{\{\vec{v}_q\}_{q=1\ldots Q}} \tag{4.55}$$

Simply, the Ambisonic decoding matrix is computed to best match the results of the Multiple-Direction Intensity Panning method.

Further approaches

Further approaches include:

- Decoding for Irregular Loudspeaker Arrays Using Interaural Cues (Treviño et al., 2011)
- Decoding for Irregular Arrays of Loudspeakers by Non-Linear Optimization (Heller, Benjamin and Lee, 2010; Wiggins, 2004)
- Comparison with and without mode matching using the hemisphere (Zotter, Pomberger and Noisternig, 2010)

4.6.10 Effects of Normalisation

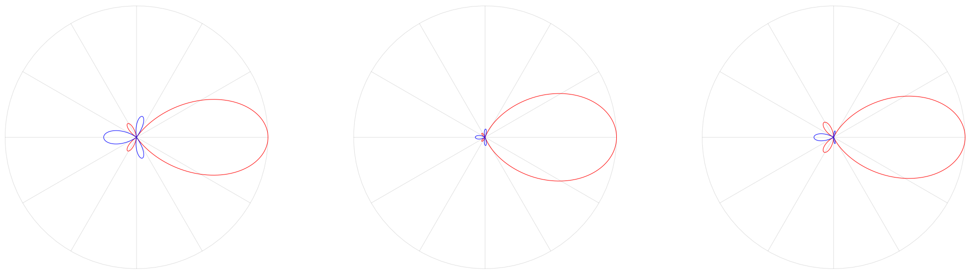
It is necessary to mention here how improper consideration of the input normalization scheme can significantly effect the output of the decoder. Normalization scales the spherical harmonic components in a particular fashion. This happens during the encoding stages, and then again in the decoding stages in the calculation of the re-encoding matrix, **C**. Crucially though, in the case of Inverse/Pseudo-Inverse decoders, as **C** is inverted to calculate the decoding matrix, **D**, so are the normalization factors. As the Ambisonic signal is multiplied by the decoding matrix the normalization and inverted normalization factors cancel and the effect of normalization is nullified.

If contradictory normalization schemes are used by the encoder/decoder then this cancellation effect is improperly calculated. Similarly, if a non-inversion based decoder is used, for example a projection decoder, then this calculation never takes place at all. By not cancelling the effects of the normalization scheme the ambisonic data itself is effectively weighted. Further, there is the potential for this weighting to be applied twice, once during the normalized encoding of the Ambisonic data and a second time during the normalized calculation of the re-encoding matrix.

The effect is similar to that discussed in Sec. 4.8. The degree dependant weighting of the Ambisonic channels results in change in shape of the virtual microphone pickup patterns. It is interesting to note then that this mistake is not necessarily immediately obvious. For example, encoding in SN3D format and decoding assuming an N3D format will result in the reduced weighting of higher order components. This is the basic concept of the Maxr_E weighting scheme, discussed in Section 4.8. Fig. 4.6 shows the coincidentally similar (although not identical) results. Both 4.6b and 4.6c show a reduction in rear lobing at the expense of slightly widening the frontal lobe. However, it is essential to be aware that one is a mistake and the other is a carefully calculated optimization.

4.6.11 Near-Field Compensation

Ambisonics is based on the assumption of plane wave theory. Mathematically encoding a source into spherical harmonics as described in Sec. 4.5.2 in fact assumes that the source must have a planar wavefront. Likewise, this means that in calculating the re-encoding matrix the loudspeaker sources are also be assumed to be planar.



(a) N3D to N3D (basic) **(b)** N3D to N3D (max-$\boldsymbol{r}_E$) **(c)** SN3D to N3D (basic)

Figure 4.6: Virtual microphone pickup patterns of 3rd order projection decoders. a) and b) show correctly normalised basic and max-$\boldsymbol{r}_E$ decoders respectively. c) shows the result if SN3D normalised data is input to an N3D normalised basic decoder. Note that in the case of both b) and c) the rear lobing is reduced at the expense of a slight widening of the frontal lobe.

For this to be true in the case of a point source it must be of infinite distance. This is of course impossible in either case.

It was shown in Eqn. 4.19 that the Ambisonic components, β^{σ}_{mi}, of a plane wave signal, s, of incidence (φ, ϑ) may be defined

$$\beta^{\sigma}_{mi} = s \cdot Y^{\sigma}_{mi}(\varphi, \vartheta) \tag{4.56}$$

Daniel goes on to describe that for a radial point source of position $(\varphi, \vartheta, r_s)$ it is necessary to consider the near-field effect filter which relates HOA to the Fourier-Bessel-Expansion of the sound field (Daniel, 2003), Γ_m, such that

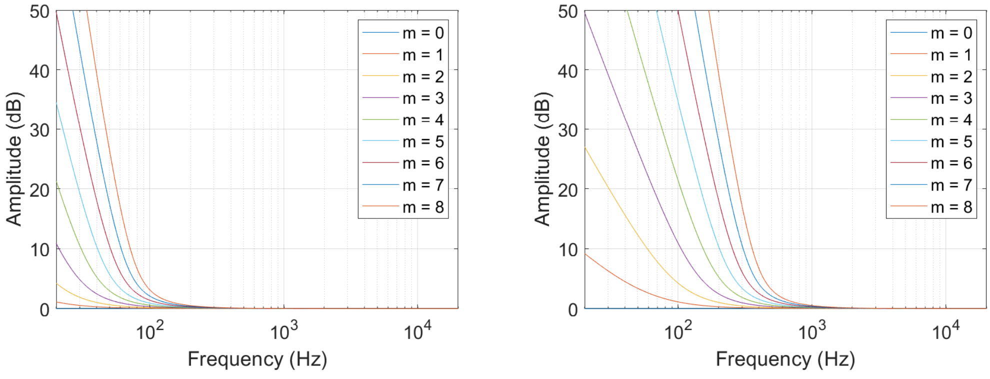
$$\beta^{\sigma}_{mi} = s \cdot \Gamma_m(r_s) \cdot Y^{\sigma}_{mi}(\varphi, \vartheta) \tag{4.57}$$

$$\Gamma_m(r_s) = k \cdot d_{\text{ref}} \cdot h^{-}_{m}(kr_s) \cdot j^{-(m+1)} \tag{4.58}$$

Where

- $k = 2\pi f / c = \omega / c$ is the wave number;
- d_{ref} is the distance at which the source, s, was measured - it is a compensation factor that derives from the equation (pressure $=$ 1/distance);
- $h^{-}_{m}(kr_s)$ are the spherical Hankel functions of the second kind (divergent);
- $j = \sqrt{-1}$.
- $\Gamma_m(r_s)$ is the degree dependant filter that simulates the effect of a non-planar source

Daniel goes on to simplify this filter by considering s to be the pressure field measured at the origin i.e. $d_{\text{ref}} = r_s$. By doing this, the need to compensate for the attenuation, $1/r_s$, and delay, $\tau = r_s/c$, of the signal may be ignored. These features may be



(a) 5m **(b)** 1m

Figure 4.7: Bass boosting amplitude response of the near-field effect filter for degree, $m = 0, 1, ... 8$, for a source simulated at a) 5m b) 1m

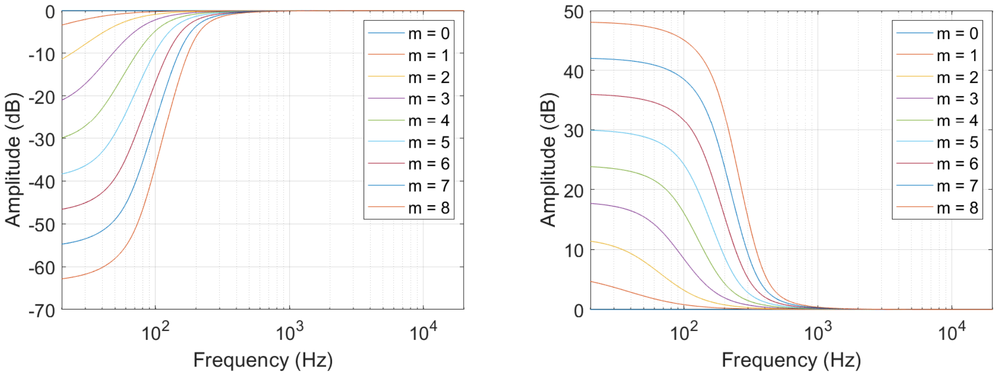
defined by the 0^{th} order (pressure only) near-field effect filter Γ_0. Hence, this factor may be removed from the general filter to show that

$$\beta^{\sigma}_{mi} = s \cdot F^{\tau}_{m} \cdot Y^{\sigma}_{mi}(\varphi, \vartheta) \tag{4.59}$$

$$F_m = \frac{\Gamma_m}{\Gamma_0} = \sum_{i=0}^{m} \frac{(m+i)!}{(m-i)!i!} \left(\frac{-jc}{\omega r_s}\right)^i \tag{4.60}$$

where F_m are the degree dependant transfer functions which model the near-field effect (wavefront curvature) of a signal originating from the point $(\varphi, \vartheta, r_s)$ having been measured from the origin. The filter should be applied to the source in the spherical harmonic domain. Generally speaking, the filters apply a phase shift and bass-boost to sources as they approach the origin and have a greater effect on higher order components.

Unfortunately, for high order representations and sources close to the origin this boost effect becomes unstable and tends towards infinity as shown in Fig. 4.7. Fortunately, there is a simple work around. In Ambisonics one must consider the near-field properties of both the original source and the reproduction loudspeakers. Whilst a near-field effect filter must be applied to the source (simulating a finite radii), a near-field compensation filter must also be applied to the loudspeakers (simulating an infinite radii). Consider the need to simulate the sources as if they were closer to the listener and the loudspeakers as if they were further from the listener. Near-field compensation filters may be introduced as simply the inverse of the near-field effect



(a) Source: 5m, Loudspeakers: 2m

(b) Source: 1m, Loudspeakers: 2m

Figure 4.8: Shelf filtering amplitude response of the combined near-field effect and compensation filters for degree, $m = 0, 1, ... 8$, for a loudspeaker radius of 2m and a source simulated at a) 5m b) 1m

filters. By combining the filters

$$H_m = \frac{F_m^{source}}{F_m^{loudspeakers}} \tag{4.61}$$

the unstable nature of the low frequency boost/cut cancel to produce a stable bass boost/cut shelf filter as shown in Fig. 4.8.

In the case that the source is being rendered at the same radius as the loudspeakers the two filters will cancel entirely. The near-field effect that must be applied to the source is the inverse of the near-field compensation that must be applied to the loudspeakers. This is why an Ambisonic system that does not consider near-field compensation at all will render a source 'on the loudspeaker array' i.e. at the same radius.

The combination of filters may be considered as defining the radius of the rendered source as an offset from the radius of the reproduction loudspeakers. A derivation is given by Daniel, 2003 that defines the filters as a series of second (and first) order sections which may be applied to any given signal in the general case. Updated regularizations are given by Favrot and Buchholz, 2012. Note that in the case where a soundfield has been recorded directly, near-field effect filters need not be applied. However, near-field compensation filters may still be required if the loudspeaker signals cannot be assumed to be planar. If this were the case, one could define a non-intrusive near-field effect filter (for example, defining a source radius of 20m) that would barely effect the signal but prevent any instability in the final filter. An example is given in Fig 4.9 which compares the ideal infinite cut filters to those

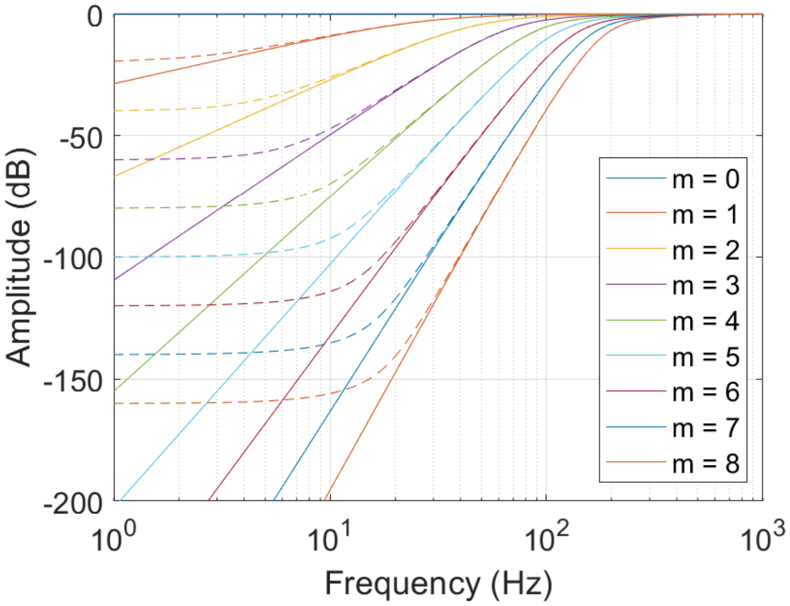


Figure 4.9: A comparison of ideal (solid line) infinite near-field control filters to those stabilised by non-intrusive near-field effect filters (dashed), defined by a source radius of 20m, for degree, $m = 0, 1, \ldots\ 8$, for a loudspeaker radius of 2m

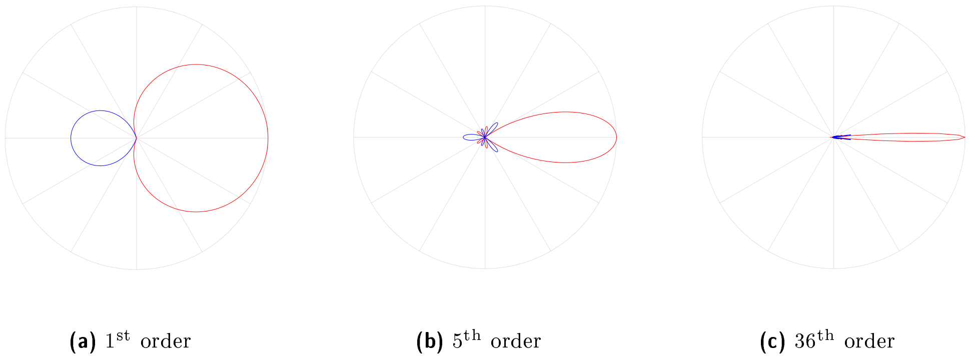
stabilised by non-intrusive near-field effect filters. Hardly any difference is seen above 20Hz and even then only for the higher order components.

4.7 Spatial Aliasing and the Sweet Spot

4.7.1 Overview

Ambisonic rendering in general defines a process of reproducing a soundfield from a finite number of fixed points about a sphere with a particular angular resolution. The angular resolution is dependent on the Ambisonic order and narrows with an increase in the number of spherical harmonics utilized. This is shown in Fig. 4.10. The graphs may be viewed in one of either two ways. Firstly, they may be considered as the gains with which a point source would be panned to a number of loudspeakers depending on their relative locations to the source. Alternatively, they may be considered as the gains with which each loudspeaker samples the soundfield about its location. The processes are one and the same.

Depending on the resolution therefore, it is common that a single point source may be sampled/output from multiple loudspeakers within a reproduction array. The result is a spatial blurring of the soundfield. A wide frontal lobe in particular means that a single source panned in any direction will in general be output from a collection



(a) 1^{st} order

(b) 5^{th} order

(c) 36^{th} order

Figure 4.10: 1^{st} (left), 5^{th} (middle) and 36^{th} (right) order Ambisonic representations of a point source. Positive values are shown in red. Negative (out of phase) values are shown in Blue.

of neighbouring loudspeakers. The wider the lobing, the greater the spread of the source.

Outputting a signal through multiple loudspeakers leads to signal repetitions that are vulnerable to comb filtering as soon at the path difference between the different speakers and the sampling point begins to vary. The outcome is that accurate reconstruction of a soundfield is limited to a small area of space in the center of the reproduction array known as the sweet spot. Inaccuracies outside of this region are a direct result of comb filtering, caused by sampling the soundfield with a finite angular resolution. This is known as spatial aliasing.

4.7.2 The Effects of Discrete Sampling

In the ideal case an infinite number of sample points (loudspeakers) is required at an infinitely fine angular resolution ($M = \infty$). Theoretically, a perfectly accurate reproduction of the original soundfield may then be achieved. Of course in reality neither condition is possible and it is necessary to consider the implications of finite alternatives.

As the number of loudspeakers is reduced spaces are created in between the sample points. Sparsely sampling a soundfield with a high angular resolution will lead to gaps where sources are panned in between loudspeakers. 3 cases are investigated:

- Under sampling considering the angular resolution
- Ideal sampling matched to the angular resolution
- Over sampling considering the angular resolution

The sampling is matched to the angular resolution by ensuring a regular distribution and setting the criteria

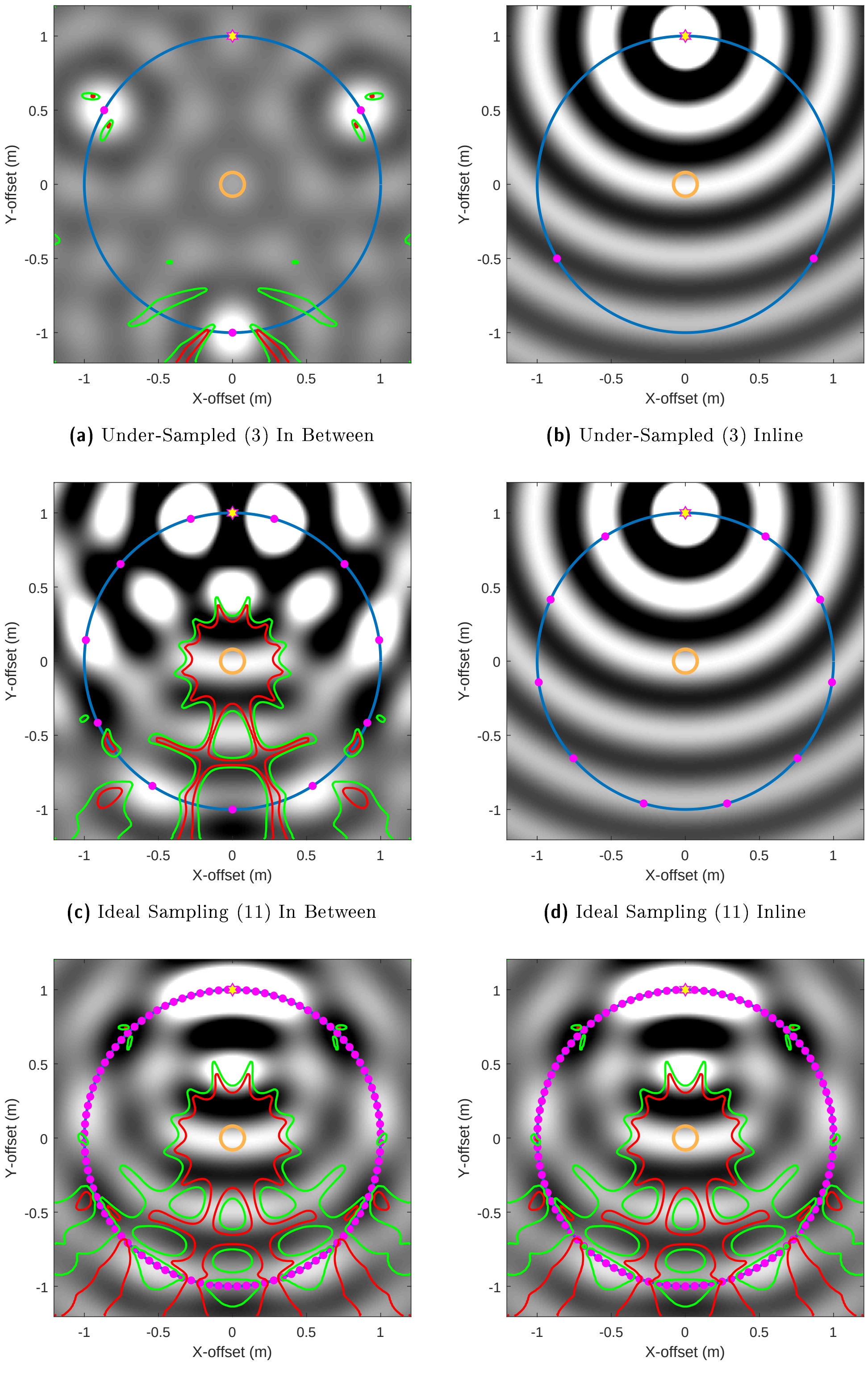
$$L = K = \begin{cases} (2M+1) & \text{(2D)} \\ (M+1)^2 & \text{(3D)} \end{cases} \tag{4.62}$$

In the following examples a 2D reproduction is considered as it is trivial to derive a regular distribution in 2 dimensions.

Fig. 4.11 shows a 5th order Ambisonic source rendered over 3 regularly distributed loudspeaker arrays (3, 11 and 100 speakers). 2 source positions are shown: exactly in between 2 loudspeakers and directly inline with a loudspeaker.

It is shown show that in the under-sampled case where a source is panned between the loudspeakers (Fig. 4.11a) an accurate source signal is not reproduced. The correct wave pattern is not generated nor is the amplitude of the reproduced signal comparable to the original. In this case a *gap* is seen in the reproduced soundfield. This is resolved in Fig. 4.11c by increasing the number of loudspeakers to close the gap. Here it is seen that the signal is primarily output from the two neighbouring loudspeakers (bright white spots) and an area of accurate reconstruction is shown in the center of the array. As the number of loudspeakers is increased (Fig. 4.11e) a significant overlapping of the angular spectrum is seen being output from a cluster of loudspeakers near the source (solid white curve). Whilst this does not significantly interfere with the area of accurate reconstruction significantly higher levels of destructive interference are seen outside of the central position.

Fig. 4.11b and 4.11d show the reproduction of an under and ideally sampled source panned directly inline with a loudspeaker. In both cases perfect reconstruction is evident. This is because only the frontal loudspeaker (the one aligned with the source) is outputting any signal. The others align with the nodes of the virtual microphone pickup pattern/angular resolution. This is due to the fact that the panning function actually takes on the form of an angular sinc function and the positions of the neighbouring loudspeakers correspond to the zero-crossings of this function when regularly distributed. In the case of the ideally sampled array this happens naturally as a result of defining the number of loudspeakers in accordance with Eqn. 4.62. In the case of the under-sampled array this is more precisely the result of the Pseudo-Inverse decoder manipulating/stretching the panning function.



(a) Under-Sampled (3) In Between

(b) Under-Sampled (3) Inline

(c) Ideal Sampling (11) In Between

(d) Ideal Sampling (11) Inline

(e) Over-Sampled (100) In Between

(f) Over-Sampled (100) Inline

Figure 4.11: 2D 5th order Pseudo-Inverse Ambisonic reproductions of a 750Hz sinusoidal point source ($\varphi = 0°, \vartheta = 0°, r = 1$) with a zero-to-peak amplitude of 1 at the center of the array. Loudspeakers (pink dots) have a radius of 1m. Amplitude is plotted on a capped colour scale: -1 to 1 black to white respectively. An orange ring of radius 8cm indicates the approximate size of a human head. Areas of accurate reproduction are highlighted by contours: $<$ 20% error (green); $<$ 10% error (red). Reproduction is performed on under, ideally, and over sampled arrays. 2 source positions are considered: directly in between and inline with the loudspeakers. Note that Figs. 4.11b and 4.11d both represent perfectly accurate restorations (no visible contours).

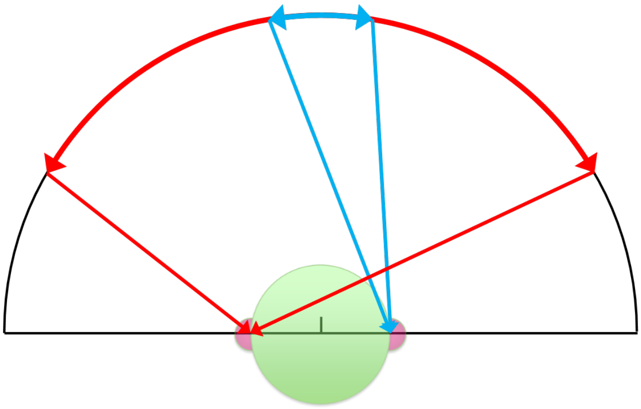


Figure 4.12: An example of an increasing difference in path length seen from an off-center listening position as a result of an increase in source spread. Note the blue paths of similar length and red paths of different length.

Very little difference is seen as the position of a source is altered in the over sampled case. Generally speaking this is because there are an almost identical number of loudspeakers outputting the same signal as before spread over the same area.

4.7.3 Calculating Path Differences

At the precise center of the array the path length to every loudspeaker is identical and all signals arrive perfectly in phase. However, from an off-center position the path lengths to each loudspeaker vary. This is true for the positions of a person's ears even if their head is centred within the array. As the total solid angle encompassing a collection loudspeakers increases so does the variance in their path length. This is shown in Fig. 4.12.

A signal that is output from multiple sources with varying path length is subject to comb filtering. The frequency at which this will begin to significantly affect a signal is a function of the difference in path length. By reducing the differences in path lengths the frequency at which destructive interference begins to present is increased.

Table 4.2 presents some typical values indicative of the frequencies at which destructive comb filtering will begin to severely impact various orders of Ambisonic reproduction. Results are shown for a loudspeaker array of radius (1m) and path lengths are calculated from a radial position 8cm away from the center of the array. This distance is typical of the radius of a human head and is therefore representative of the position of a listener's ear.

Table 4.2: The approximate spread of various Ambisonic order source representations and the resulting maximum path differences for a sampling point 8cm from the center of the array. The first frequency at which destructive interference will occur as a result of comb filtering is also presented.

order	Spread (approx.)	Maximum Path Difference	First out-of-phase frequency
1	180°	16.00cm	1071Hz
5	40°	5.46cm	3143Hz
36	5°	0.7cm	24652Hz

The table indicates the approximate width of the frontal lobe of basic 1^{st}, 5^{th} and 36^{th} order Ambisonic panpots, as shown in Fig. 4.10. These angles are representative of the primary spread of loudspeakers from which each source would be output. From this, the approximate (worst case) path differences of the repeated signals to the off-center listening position can be found. The path difference is then used to find the first frequency to encounter a phase difference of $\pi/2$. Here we assume a speed of sound of $343ms^{-1}$ in the case of dry air at 20°C.

$$f = \frac{c}{\lambda} = \frac{343}{\text{Path Difference} \times 2} \tag{4.63}$$

i.e. complete destructive interference.

Of course, these values do not represent the exact frequency at which the Ambisonic reproduction becomes inaccurate. This is, after all, a gradual deterioration that also depends on the number and position of loudspeakers. However, the results do indicate the approximate frequencies at which significant destructive interference will impact the reproductions. It is reassuring therefore to see that these values are of similar order to those presented in Table 4.3 (the results of another measure of Ambisonic reproduction accuracy). Similar results are also presented by Rafaely, 2005 who describes the spatial aliasing frequency in the form of the wave number, k, as a function of head radius, r, and Ambisonic order, N, as $kr \leq N$.

Poletti shows that for an Ambisonic array the minimum reconstruction error is found to be for the case of a regular layout where $L = K$ (Poletti, 1996). However, Daniel discusses that this is only the result of the mean error and is primarily due to the angles from which soundfield reconstruction is exact, i.e. the directions of the loudspeakers, rather than the maximum error (Daniel, Rault and Polack, 1998; Daniel, 2001). Typically, it would be necessary to avoid a severely oversampled case due to the significant destructive inference of high frequency sources outside of the sweet spot. It is common to see loudspeaker arrays implemented with $L > K$ although this should not be excessive. For example, rendering 1st order Ambisonics over a cube: 4 channels → 8 speakers or rendering 5th order Ambisonics over a 50 point Lebedev grid: 36 channels → 50 speakers.

4.8 Decoder Matrix Weightings

4.8.1 Overview

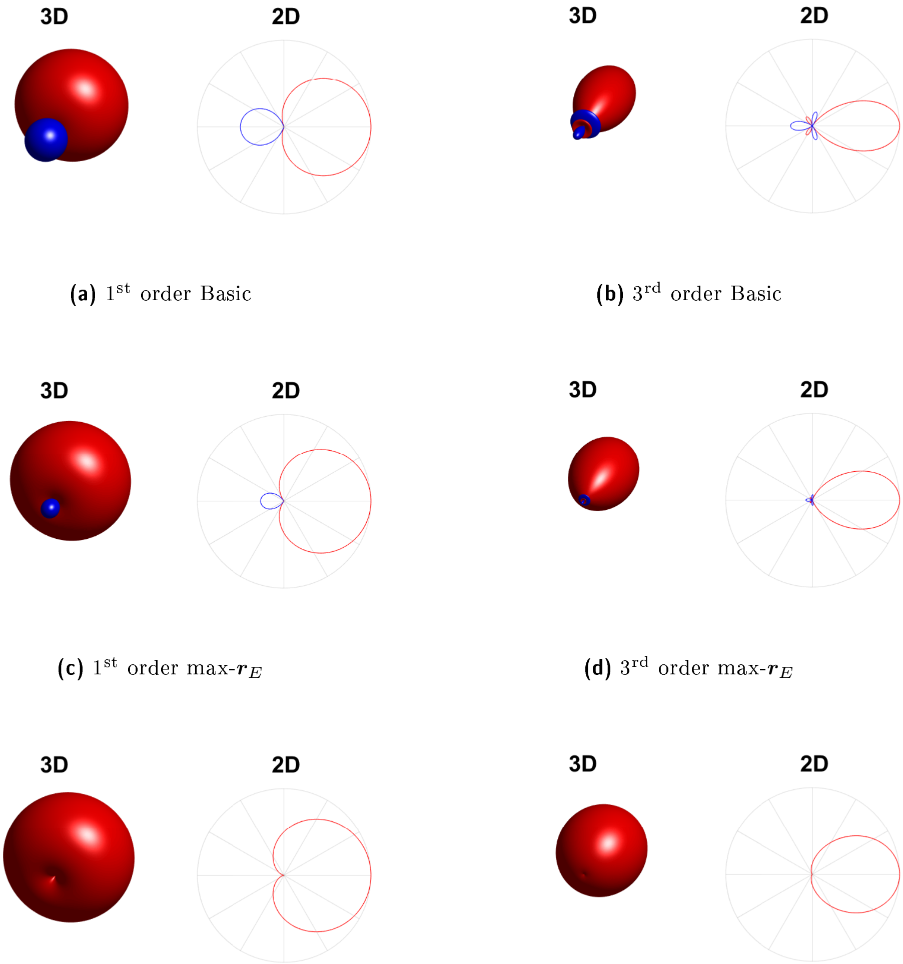
As previously discussed within this chapter, Ambisonic reproduction may be considered in terms of the virtual microphone pick-up patterns reproduced by each loudspeaker in a playback array. The shape of these patterns is entirely defined by the summation of spherical harmonic components. By applying an additional weighting, either prior to decoding or directly within the columns of the decoding matrix, it is possible to skew the shape of the pick-up patterns. By weighting the components within each degree of spherical harmonics by an equal value the orientation of the resultant pick-up patterns remains unaffected and remains solely defined by the decoding matrix.

It has been shown in Section 4.7 that beyond a certain frequency the accuracy of a standard Ambisonic decoder begins to deteriorate outside of the sweet spot. However, by manipulating the virtual pickup patterns of the loudspeakers it is possible to optimize the decoders based on alternative psychoacoustic parameters instead. In this way the decoders no longer attempt to accurately reconstruct an entire soundfield, but instead are designed to improve the reconstruction of perceptual cues such as ILD (Daniel, 2001; Gorzel, Kearney and Boland, 2014). Graphical comparisons of 2 popular schemes are presented in Fig. 4.13 for 1st and 3rd order decoders.

A set of weights, $a_m^{\{...\}}$, are defined as the weights that are individually calculated for each degree. They can also be presented in matrix form such that

$$\mathbf{D}^{\{...\}} = \mathbf{D}' \cdot \mathbf{\Gamma}_M^{\{...\}} \tag{4.64}$$

$$\mathbf{\Gamma}_M^{\{...\}} = \mathrm{diag}\{[a_0^{\{...\}}, ... \underbrace{a_m^{\{...\}}, ..., a_m^{\{...\}}}_{2m+1}, ... \underbrace{a_M^{\{...\}}, ..., a_M^{\{...\}}}_{2M+1}]\} \tag{4.65}$$



(a) 1$^{\text{st}}$ order Basic

(b) 3$^{\text{rd}}$ order Basic

(c) 1$^{\text{st}}$ order max-$\boldsymbol{r}_E$

(d) 3$^{\text{rd}}$ order max-$\boldsymbol{r}_E$

(e) 1$^{\text{st}}$ order In-Phase

(f) 3$^{\text{rd}}$ order In-Phase

Figure 4.13: Virtual microphone pick-up patterns rendered by 1$^{\text{st}}$ (left) and 3$^{\text{rd}}$ (right) order Basic, max-$\boldsymbol{r}_E$ and In-Phase weighted pseudo-inverse decoders. The 1$^{\text{st}}$ order decoder is designed for a regular octahedron layout. The 3$^{\text{rd}}$ order decoder is designed for an almost regular 26 point Lebedev Grid and the pick-up pattern is shown for the frontal loudspeaker. Note that the small out-of-phase component in Fig. 4.13f may be attributed to using a particular set of in-phase weights calculated from a generalised formula derived for perfectly regular loudspeaker arrays (Daniel, 2001). Positive values are shown in red. Negative (out of phase) values are shown in Blue.

Table 4.3: Example Crossover frequencies for Basic - max-$\boldsymbol{r}_E$ Dual-Band decoding (Daniel, Rault and Polack, 1998)

Order	1	2	3	4	5
Freq. (Hz)	700	1250	1850	2500	3000

4.8.2 Dual-Band Decoding

Dual-Band decoding is the commonly implemented technique to alter the decoder weighting scheme over frequency (Heller, Benjamin and Lee, 2012; Heller, Benjamin and Lee, 2010; Kearney and Doyle, 2015b; Heller, Lee and Benjamin, 2008).

Typically a basic weighting would be used at low frequencies with a max-$\boldsymbol{r}_E$ weighting at high frequencies. A crossover frequency is usually selected at the approximate frequency at which holophonic reproduction becomes invalid for a human listener as a result of a shrinking sweet spot with an increase in frequency (Yao, Collins and Jancovic, 2015). After this point, energy localisation is prioritised by the max-$\boldsymbol{r}_E$ weighting scheme to best preserve ILD localisation cues. Some example figures are shown in Table 4.3 but are by no means definitive (Daniel, Rault and Polack, 1998).

The technique may be applied either by computing two separate decode matrices and combining the individual outputs using a crossover filter or by applying a series of shelf filters to the Ambisonic channels (in essence creating a single frequency dependant weighting) (Gerzon, 1980; Gerzon and Barton, 1992; Kearney et al., 2012; Gerzon, 1992b). Each method is entirely equivalent.

4.8.3 Amplitude/Energy Preservation

Before weighting the Ambisonic channels, consideration must be given as to whether the overall set of weights should preserve either the amplitude or energy levels of the restored soundfield with respect to the original. Note that given the discrete sampling and therefore superposition of multiple loudspeaker sources these two metrics do not necessarily equate. By altering the shape of the pickup pattern the spread of a source between neighbouring loudspeakers is altered and therefore the distribution and total summation of energy. This may be accounted for with an additional normalization coefficient, a_0.

In the case of amplitude preservation:

$$\text{(Amp. Preserving)} \qquad a_0 = 1 \tag{4.66}$$

In the case of energy preservation, the change in the spread of a source amongst a number of loudspeaker must be accounted for. Daniel derives simple linear scaling factors in the case of non-minimal ($L > K$) regular arrays (Daniel, Rault and Polack, 1998). First, he defines the total energy gain as a result of the channel weights, a_m (Daniel, 2001, Eqn. A.62)

$$E_M(\boldsymbol{a}_M) = \frac{1}{L}\sum_{m=0}^{M}(2m+1)(a_m)^2 \tag{4.67}$$

For playback energy to be preserved unity gain is required

$$E_M(\boldsymbol{a}_M) = 1 \tag{4.68}$$

The normalisation factor, a_0, may be considered such that

$$a_m = a_0 \cdot a'_m \tag{4.69}$$

By factoring the term a_m^2 in Eqn. 4.67 it may be written

$$E_M(\boldsymbol{a}_M) = a_0^2 \cdot \frac{1}{L}\sum_{m=0}^{M}(2m+1)(a'_m)^2 \tag{4.70}$$

$$= a_0^2 \cdot E'_M \tag{4.71}$$

By defining the term

(Energy Preserving) $$a_0 = \sqrt{\frac{1}{E'_M}} \tag{4.72}$$

$$= \sqrt{L\frac{1}{\sum_{m=0}^{M}(2m+1)(a'_m)^2}} \tag{4.73}$$

Eqn. 4.68 is satisfied by a cancellation of terms in Eqns. 4.71 and 4.72.

In practice, such absolute normalization of playback levels is often frivolous. Many alternative, unavoidable and untraceable normalization steps are present throughout almost every signal chain from signal normalization in the saving/transmission of data to individual loudspeaker gains/sensitivity. Preserving the absolute amplitude/energy levels of an original soundfield is therefore best achieved through acoustic calibration.

The case where this is most relevant is when combining the output from multiple decoding matrices, for example, within a dual-band decoder. In the case of amplitude preservation the task is trivial. However, in the case of energy preservation each matrix must be considered individually as the different weighting schemes independently effect the total playback energy. One option would be to normalize each set of weights in order to preserve the absolute energy in each case

$$a_m^{\{\mathbf{A}\}} = a_0^{\{\mathbf{A}\}} \cdot a_m^{'\{\mathbf{A}\}} \tag{4.74}$$

$$a_m^{\{\mathbf{B}\}} = a_0^{\{\mathbf{B}\}} \cdot a_m^{'\{\mathbf{B}\}} \tag{4.75}$$

Where **A** and **B** are referring to the gains applied to individual decoding matrices within a dual-band decoder. However, as absolute normalization is often compromised regardless it may be more convenient to consider only the relative normalization

$$a_m^{\{\mathbf{A}\}} = a_m^{'\{\mathbf{A}\}} \tag{4.76}$$

$$a_m^{\{\mathbf{B}\}} = a_0^{\{\frac{\mathbf{B}}{\mathbf{A}}\}} \cdot a_m^{'\{\mathbf{B}\}} \tag{4.77}$$

This method matches the energy output of the second matrix to the first matrix without necessarily ensuring a match to the original source. The ratio

$$a_0^{\{\frac{\mathbf{B}}{\mathbf{A}}\}} = \frac{a_0^{\{\mathbf{B}\}}}{a_0^{\{\mathbf{A}\}}} \tag{4.78}$$

may therefore be considered. This expression may then be expanded

$$a_0^{\{\frac{\mathbf{B}}{\mathbf{A}}\}} = \frac{\sqrt{\frac{1}{E'_M(\boldsymbol{a}_M^{'\{\mathbf{B}\}})}}}{\sqrt{\frac{1}{E'_M(\boldsymbol{a}_M^{'\{\mathbf{A}\}})}}} \tag{4.79}$$

$$= \sqrt{\frac{E'_M(\boldsymbol{a}_M^{'\{\mathbf{A}\}})}{E'_M(\boldsymbol{a}_M^{'\{\mathbf{B}\}})}} \tag{4.80}$$

$$= \sqrt{\frac{\sum_{m=0}^{M}(2m+1)(a_m^{'\{\mathbf{A}\}})^2}{\sum_{m=0}^{M}(2m+1)(a_m^{'\{\mathbf{B}\}})^2}} \tag{4.81}$$

In the case that

$$a_m^{'\{\mathbf{A}\}} = 1 \qquad m = 0, 1, ..., M \tag{4.82}$$

(as is true for a basic decode matrix) this may be simplified further

$$a_0^{\{\frac{\mathbf{B}}{\mathbf{A}}\}} = \sqrt{\frac{\sum_{m=0}^{M}(2m+1)(1)^2}{\sum_{m=0}^{M}(2m+1)(a_m^{'\{\mathbf{B}\}})^2}} \tag{4.83}$$

$$= \sqrt{\frac{1}{\frac{\sum_{m=0}^{M}(2m+1)(a_m^{'\{\mathbf{B}\}})^2}{\sum_{m=0}^{M}(2m+1)}}} \tag{4.84}$$

By expanding the weights, $\boldsymbol{a}_M^{'\{\mathbf{B}\}}$, into a vector of appropriate length, Eqn. 4.84 may be re-imagined as

$$a_0^{\{\frac{\mathbf{B}}{\mathbf{A}}\}} = \sqrt{\frac{1}{\text{mean}([(a_0^{'\{\mathbf{B}\}})^2, ...\underbrace{(a_m^{'\{\mathbf{B}\}})^2, ..., (a_m^{'\{\mathbf{B}\}})^2}_{2m+1}, ...\underbrace{(a_M^{'\{\mathbf{B}\}})^2, ..., (a_M^{'\{\mathbf{B}\}})^2]}_{2M+1})}} \tag{4.85}$$

Where normalisation is performed with respect to the mean value of the weights applied to each spherical harmonic coefficient. This is the simplified max-$\boldsymbol{r}_E$ normalization step implemented in common applications of dual-band decoding such as Google Resonance[20].

4.8.4 Basic (Max $\boldsymbol{r}_V$)

Before alternative weighting schemes are derived, standard or basic weighting must be defined. Basic decoding refers to unity or no additional weighting across all channels. It is suited for holophonic/mode-matching decoders where one wishes to abide by the mathematically founded techniques. By nature, Basic weighting maximises the velocity vector and as such is occasionally referred to as Max- r_V decoding. This technique is optimally utilized at low-frequencies where soundfield reconstruction is at its most accurate and ITD cues may be preserved.

4.8.5 Max $\boldsymbol{r}_E$

The high frequency directional quality of a panned source may be somewhat measured by the energy vector, $\boldsymbol{r}_E$. Gerzon makes particular reference to its indication of quality in the region 700-4000Hz (Gerzon and Barton, 1992). It is beneficial to

[20] developers.google.com/resonance-audio

maximise this measure for high frequency sources where perceptual localisation is dominated by energy cues over timing cues. A general derivation for a set of weights, or gains, a_m, which maximise $\boldsymbol{r}_E$ over regular loudspeaker layouts is given by Daniel, 2001. It is common, however, to implement these weights for irregular arrays also for which they remain approximately correct.

The energy vector, r_E, is first derived, and then differentiated, with respect to a_m. The result is set to zero to find the maximum of the function (i.e. when the $\boldsymbol{r}_E$ vector is maximised). This may be shown to give us the following equation (3D) (Daniel, 2001).

$$(2m+1)\boldsymbol{r}_E a_m = (m+1)a_{m+1} + m a_{m-1} \tag{4.86}$$

The recurrence relationship bares significant resemblance to one of the Legendre functions, known as Bonnet's recursion formula (Morse and Ingard, 1968)

$$(2m+1)\eta P_m(\eta) = (m+1)Pm + 1(\eta) + mP_{m-1}(\eta) \tag{4.87}$$

if $a_m = P_m(\eta)$ and $\eta = \boldsymbol{r}_E$ is defined. The vector $\boldsymbol{r}_E$ is therefore maximised by the relationship

$$a_m = P_m(\boldsymbol{r}_E), \quad m = 0, 1...M \tag{4.88}$$

Eqn. 4.88 defines $a_0 = 1$ and $a_1 = \boldsymbol{r}_E$. As it is required that $a_{-1} = 0$ and $a_{M+1} = 0$, $\boldsymbol{r}_E$ may be defined as the largest root of P_{M+1} such that

$$a_{M+1} = P_{M+1}(\boldsymbol{r}_E) = 0 \tag{4.89}$$

is satisfied.

A max-$\boldsymbol{r}_E$ weighting scheme will reduce the presence of side lobes in the pickup pattern at the expense of slightly widening of the frontal lobe, as shown in Fig. 4.13. In practise, this reduces the contributions from loudspeakers located far from a panned source increasing the concentrating of energy in the direction of the source.

4.8.6 In-Phase

In-Phase decoding weights the Ambisonic channels such that the output from all loudspeakers remains in phase. Unlike the max-$\boldsymbol{r}_E$ weighting scheme there are multiple solutions to this stipulation.

A number of examples are presented up to second order by Monro, 2000. A general Ambisonic panning function, $g(\vec{v}_s)$, is derived and a set of weights, a_m, $m = 0, 1, ...\ M$ are subsequently introduced to the function. A multi-dimensional region may then be defined in terms of a_m such that the output to $g(\vec{v}_s)$ is positive for all $\vec{v}$.

Although any solutions within this region may be considered as In-Phase, the solutions that are of most interest lie on the boundary. In particular, Daniel adopted a solution referred to by Monro as the 'Smooth Solution' that presents a single forward facing lobe. He generalises the description in his thesis (Daniel, 2001) to say that the output of $g(\vec{v}_s)$ must fall over the range $\vec{v} = [0\pi]$ and is 0 when $\vec{v} = \pi$. The weights may be calculated

$$a_m = \frac{M!(M+1)!}{(M+m+1)!(M-m)!} \tag{4.90}$$

Smooth In-Phase decoding has shown particular promise in rendering Ambisonics for large audiences (Stitt, 2015; Kearney, 2010; Malham, 1992) by removing the output of potentially overpowering side lobes that may become more prominent due to an individual listener's proximity to a particular loudspeaker.

4.9 Binaural Rendering of Ambisonics

4.9.1 Overview

Complete spherical loudspeaker rigs suitable for Ambisonic reproduction are expensive, dominating and hard to come by. Securing an environment suitable for an array such as that in Fig. 4.14 is not an easy task. Neither is it applicable or even ideal in many situations. With a recent uptake in the field of virtual reality portable head mounted devices are becoming a popular source of content delivery. Whether at home or in a more remote location headphones provide a far more convenient and reliable means to deliver audio in such a circumstance. Even without the visual accompaniment, headphones are a flexible and accessible alternative to loudspeakers throughout a range of industries from sports to cinematic. It is therefore beneficial to consider binauralisation of an Ambisonic array.

There are two leading methods when it comes to rendering an Ambisonic signal binaurally: via a set of virtual loudspeakers or within the spherical harmonic domain.

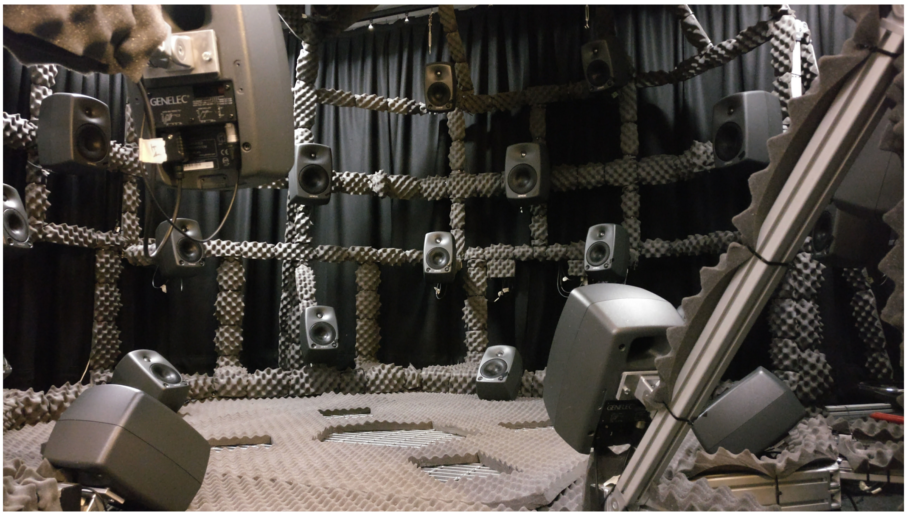


Figure 4.14: The 50 point spherical loudspeaker array housed in AudioLab, Genesis 6, University of York. The array has been treated with acoustic foam in an attempt to reduce reflections.

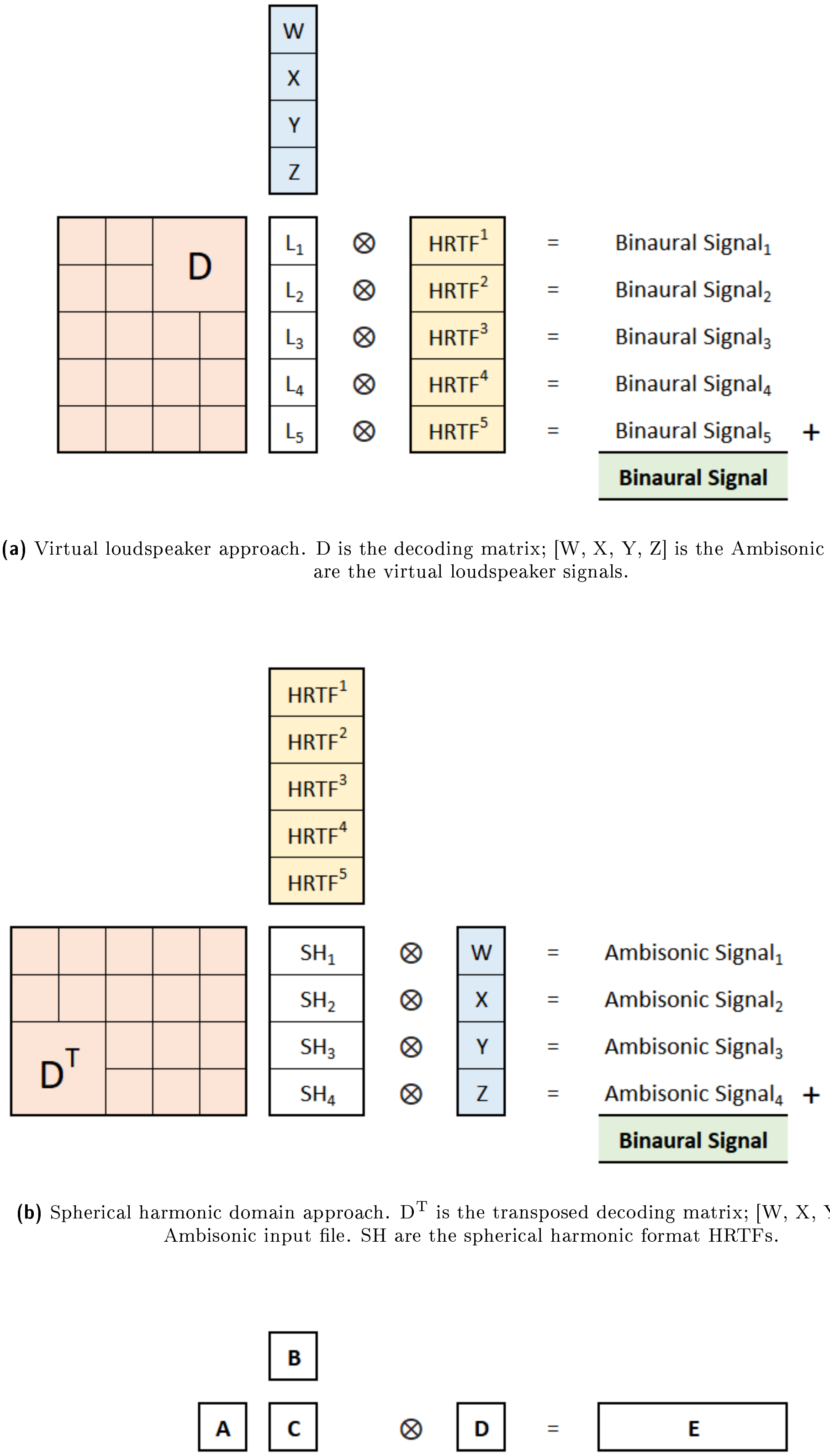
4.9.2 Virtual Loudspeakers

The virtual loudspeaker approach (Kearney and Doyle, 2015a; Smyth and Smyth, 2016; Zotter and Frank, 2012; Menzies and Al-Akaidi, 2007), see Fig. 4.15a, most directly replicates the real-world method for rendering Ambisonics over a loudspeaker array and is the simplest to understand. It is performed exactly as described in Section 2.6.3. The initial rendering process is identical to real-world reproduction. A loudspeaker configuration is chosen and loudspeaker signals are generating using a decoding matrix (e.g. Pseudo-Inverse, SN3D normalised). A set of HRTFs is measured that correspond to the positions of the loudspeakers. The loudspeaker signals are then convolved with the corresponding HRTFs and the results are summed together for each ear into a single stereo file.

4.9.3 The Spherical Harmonic Domain

The spherical harmonic domain approach (Avni et al., 2013), see Fig. 4.15b, differs from the virtual loudspeaker approach in that the convolutions are now undertaken in the spherical harmonic domain. Given the same loudspeaker configuration the two approaches may be shown to be numerically identical. However, by first encoding the HRTFs into spherical harmonics the number of convolutions required now depends on the number of spherical harmonics, and not on the number of loudspeakers. As Ambisonic decoding generally requires $L > K$ this reduces the number of convolutions required and therefore the complexity of the renderer.

The HRTFs are first encoded into spherical harmonics. However, this is *not* done



(a) Virtual loudspeaker approach. D is the decoding matrix; [W, X, Y, Z] is the Ambisonic input file. L are the virtual loudspeaker signals.

(b) Spherical harmonic domain approach. D^T is the transposed decoding matrix; [W, X, Y, Z] is the Ambisonic input file. SH are the spherical harmonic format HRTFs.

(c) Spherical harmonic domain approach

Figure 4.15: Binaural rendering workflows for Ambisonics. The figures should be read left to right and generally depict a matrix multiplication, a series of stereo convolutions and a final summation. They may be read with reference to Fig. 4.15c: **A** is multiplied by **B** to give **C** which is convolved with **D** to give **E**.

using a standard encoding matrix, but instead with a transposed decoding matrix. By doing this it is ensured that the same weights are being applied as would be applied during a virtual loudspeaker render. The result is a set of spherical harmonic components which may be convolved with corresponding Ambisonic input components. Again, the outputs of the convolutions are summed to get our binaural signal.

$$\sum_{k=1}^{K}\left(\left(\sum_{l=1}^{L} Y_k(\vec{v}_l)\cdot h_l\right) * \beta_k\right) \tag{4.91}$$

for each stereo channel where:

- K is the number of Ambisonic Channels,
- L is the number of loudspeakers,
- $Y_k(\vec{v}_l)$ is the Decoding matrix coefficient representative of the Ambisonic channel, k, for the loudspeaker l,
- h_l is the binaural filter measured from the position of that loudspeaker,
- β are the Ambisonic input channels.

One of the major advantages of this technique is that a dual-band decoder may be implemented by pre-processing the encoded HRTFs. This removes the need to implement parallel decoding matrices or perform real time filtering of the input Ambisonic signal. This is typically not possible as the matrix weightings depend on the degree of spherical harmonics. However, as the HRTFs are now stored in a compatible format, this operation becomes trivial. The spherical harmonic components of the HRTFs are simply weighted/filtered in exactly the same fashion as if they were the input Ambisonic signal.

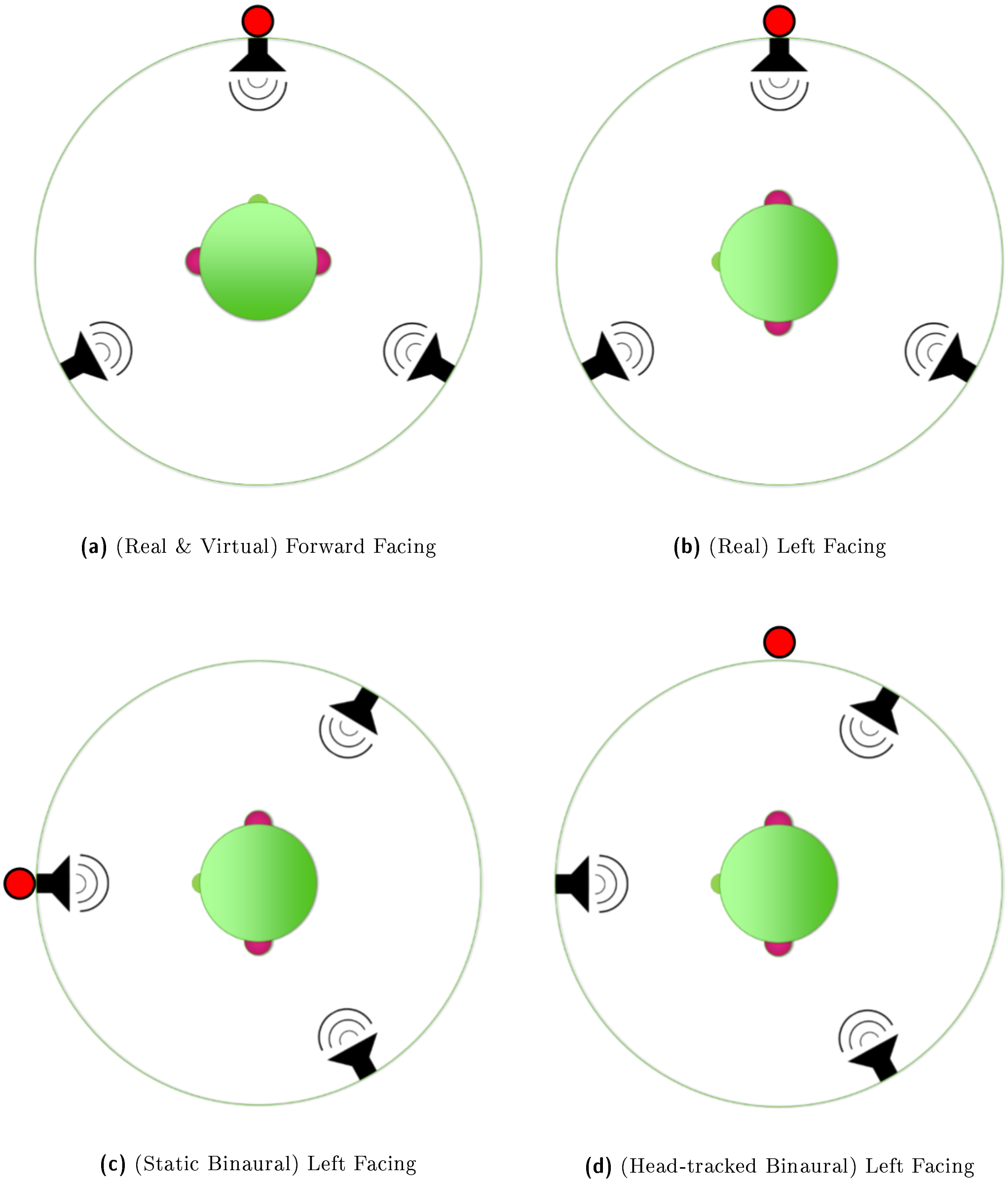
4.9.4 Head Tracked Binaural

A significant challenge within binaural Ambisonics relates to allowing a listener to ‘move’ or ‘rotate’ within the virtual array (however it is rendered). In real life this action is trivial. As a listener turns their head the surrounding loudspeakers remain stationary. Sources that have been rendered to the loudspeakers therefore also remain stationary and stable. However, for a loudspeaker array that has been rendered virtually over a pair of headphones this is not the case. As the listener moves, so do their headphones, and therefore the virtual array carrying with it the rendered sources.

An obvious solution would be to dynamically render the virtual loudspeakers such that as the listener turns their head the loudspeaker positions are redefined, updated and the loudspeaker signals re-rendered through new HRTFs. However, this is computationally troublesome. Large numbers of HRTFs would be required in combination with accurate interpolation to execute this effectively. Careful consideration must be given to the latency of such a system as a deterioration in perceptual quality is seen in systems with delays of approximately 70-80ms or more (Brungart et al., 2004; Brungart, Simpson and Kordik, 2005). Fortunately, there is an alternative method that once again takes advantage of the spherical harmonic format of the Ambisonic signals.

Head-tracking is utilized to monitor a subject's head movements. Simple linear transformation matrices are then applied in the spherical harmonic domain to counter-rotate the soundfield being reproduced. For example, as a user rotates their head to the left, the soundfield is counter-rotated to the right. The result is a constantly updated soundfield which appears to the listener to remain stable. Soundfield rotations are lossless and exhibit a relatively inexpensive computational load. Examples of how this is done are given by Kronlachner and Zotter, 2014. The basic principle is to generate a new set of spherical harmonic components based on the weighted summation of the previous set.

One consideration is that this technique does not exactly mimic a real-world situation. A comparison of the methods is given in Fig. 4.16. By rotating the soundfield it is accepted that the virtual loudspeakers are effectively locked to the geometry of the head. In Ambisonics the accuracy with which a source is reproduced may depend on its alignment with a loudspeaker (or sample point). For a source that is aligned with a loudspeaker in real life, its position never changes. However, for a source that is initially aligned with a loudspeaker and rendered binaurally, its position may shift as a user turns their head and the soundfield is rotated. The effects of this discrepancy are limited by opting for optimally regular loudspeaker configurations or dense arrays.



(a) (Real & Virtual) Forward Facing

(b) (Real) Left Facing

(c) (Static Binaural) Left Facing

(d) (Head-tracked Binaural) Left Facing

Figure 4.16: Differences between real-world, static binaural and head-tracked binaural reproductions of Ambisonic soundfields. Note the rotation of loudspeakers in Fig. 4.16c and 4.16d and counter-rotation of the source in Fig. 4.16d.

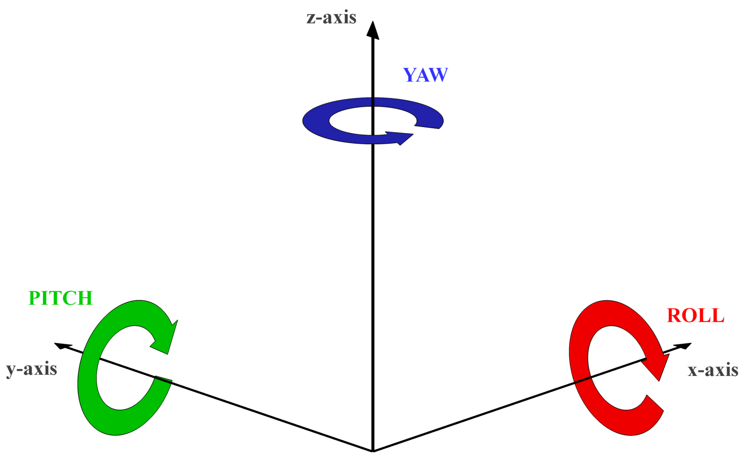


Figure 4.17: Yaw, Pitch and Roll Rotations about the z, y and x axes respectively.

4.9.5 Soundfield Rotations

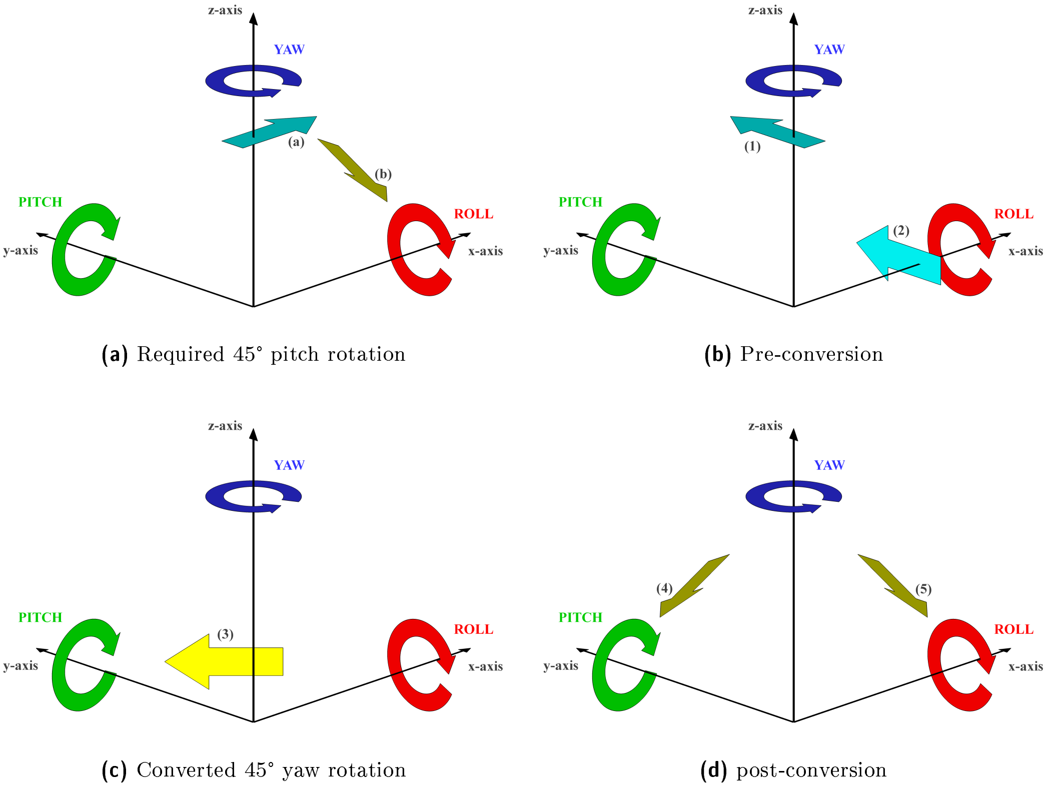
The counter rotation of a soundfield for head tracked applications may be done through simple rotation matrices applied to the Ambisonic signals.

$$\beta_{\text{out}} = \mathbf{T} \cdot \boldsymbol{\beta}_{\text{in}} \tag{4.92}$$

Note the order of matrix multiplication.

The calculation of the transform matrix, T, is trivial for rotations around the vertical axis and is described by Kronlachner and Zotter, 2014. However, rotations are generally defined around multiple axes, or at least in 3 dimensions. One method, the Euler rotations, define a *yaw*, *pitch* and *roll* as per Fig. 4.17. Note that the definition of the axes and rotation directions is somewhat arbitrary and varies between application. The order in which these rotations are defined and implemented is critical. For rotations around the x or y axes the calculation of the transform matrices is more complex. Although their derivation is defined by Zotter, 2009, he proposes a solution in which these rotations may in fact still be carried out around the z axis by using a simple set of fixed 90° conversion matrices which may be pre-calculated. An example of this transform is given in Fig. 4.18.

A 45° pitch is implemented as a low computation yaw by first performing a 90° yaw (1) followed by a 90° pitch, implementing the rotation, then inverting the process



(a) Required 45° pitch rotation

(b) Pre-conversion

(c) Converted 45° yaw rotation

(d) post-conversion

Figure 4.18: Implementation of a pitch rotation (a to b) by means of pre- and post- 90° conversion rotations and a variable yaw rotation. First a 90° yaw (1) is followed by a 90° pitch (2). The variable rotation is then performed as a yaw (3). The conversion is then inverted by a −90° pitch (4) and a −90° yaw (5).

with a −90° pitch followed by a −90° yaw. Each basic operation is therefore defined

$$\text{Yaw} = \mathbf{T}_z(\alpha)\boldsymbol{\beta} = T_z(\alpha) \cdot \boldsymbol{\beta} \tag{4.93}$$

$$\text{Pitch} = \mathbf{T}_y(\gamma)\boldsymbol{\beta} = \mathbf{T}_z(-90) \cdot \mathbf{T}_y(-90) \cdot T_z(\gamma) \cdot \mathbf{T}_y(90) \cdot \mathbf{T}_z(90) \cdot \boldsymbol{\beta} \tag{4.94}$$

$$\text{Roll} = \mathbf{T}_x(\zeta)\boldsymbol{\beta} = \mathbf{T}_y(90) \cdot T_z(\zeta) \cdot \mathbf{T}_y(-90) \tag{4.95}$$

The transformation matrices $\mathbf{T}_y(-90)$ are provided by Zotter up to order 21 online [21]. From these $\mathbf{T}_y(90)$ may be calculated by taking the inverse

$$\mathbf{T}_y(90) = (\mathbf{T}_y(-90))^{-1} \tag{4.96}$$

By defining the operations individually the most efficient conversion matrices may pre-calculated for any order of rotations.

4.10 Summary

An exhaustive review of Ambisonic principles, mathematical foundations, encoding and decoding strategies has been presented. Psychoacoustic based solutions to the spatial aliasing problem outside of the sweet spot (e.g. max-$\boldsymbol{r}_E$, in-phase) are shown to promote perceptually relevant spatial cues such as interaural time and level differences. It is shown that such manipulations must be considered carefully and that normalisation between decoding matrices based on amplitude or energy preservation must be properly implemented.

Methods for presenting Ambisonics binaurally over a virtual loudspeaker array have been explained and a solution to limited computational expenditure, spherical harmonic domain convolution, has been numerically shown to give exactly equivalent results. Examples of 3D soundfield transformation matrices commonly used to compensate for head movement in head tracked binaural applications are also given.

The information in this chapter provides a strong foundation from which to begin exploring optimizations in the decoding and rendering processes of binaural Ambisonics. The remainder of this thesis will go on to investigate the implications of manipulating the HRTFs within the binaural renderer to achieve a substantially improved objective and perceptual output.

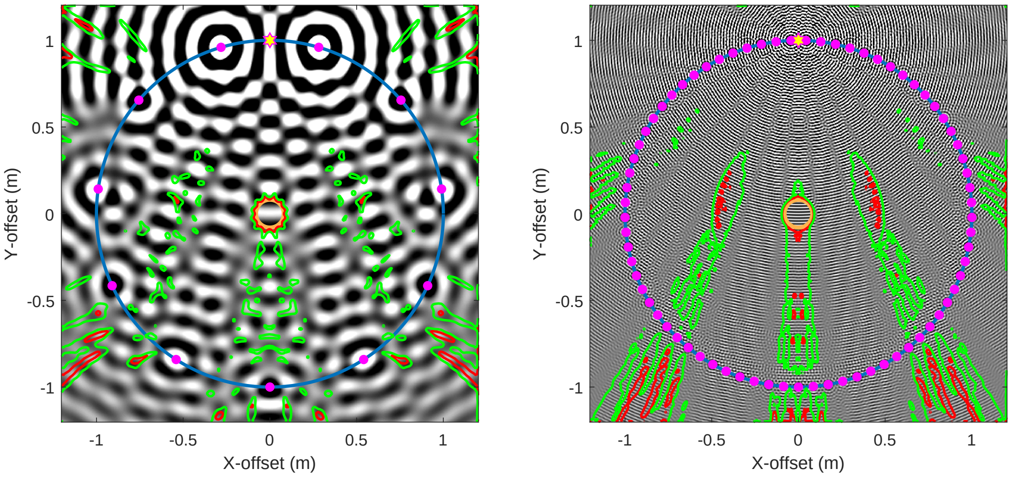
[21] iaem.at/ambisonics/xchange/fileformat/docs/spherical-harmonics-rotation

This Page Has
Intentionally Been
Left Blank

5 The Conceptual Development and Evaluation of BiRADIAL

Chapter Overview

This Chapter introduces two techniques for optimizing the binaural rendering of Ambisonics through manipulation of HRTFs and hence the reproduction of the sweet spot independently for each ear. The methods of both techniques are explained within a multi-frequency band approach. The reduced need for decoder matrix weighting schemes is discussed along with the effects of NFC filters for rendering source distance. A perceptual model for comparing binaural filters (the PSDM model) is also presented and validated through objective and subjective testing.



(a) 5th order: 2500Hz

(b) 36th order: 20000Hz

Figure 5.1: 2D Pseudo-Inverse Ambisonic reproductions of a sinusoidal point source ($\varphi = 0°, \vartheta = 0°, r = 1$) with a zero-to-peak amplitude of 1 at the center of the array. Loudspeakers (pink dots) have a radius of 1m. Amplitude is plotted on a capped colour scale: -1 to 1 black to white respectively. An orange ring of radius 8cm indicates the approximate size of a human head. Areas of accurate reproduction are highlighted by contours: $<$ 20% error (green); $<$ 10% error (red). The figure shows the similarly sized sweet spots of a frequency limited 5th order decoder reproducing a source of 2500Hz and a 36th order decoder reproducing a source of 20000Hz.

5.1 Introduction

In Chapter 4 it is shown that the soundfield reconstruction properties of Ambisonics deteriorate beyond the center of the array outside of the area known as the sweet spot. The size of the sweet spot depends on the frequency of the source, order of Ambisonics, source position(s) and loudspeaker configuration. Errors outside of the sweet spot are a result of a reduced spatial resolution and the spatial aliasing caused by the superposition of multiple loudspeaker signals of variable path length subject to comb filtering.

These errors become problematic as one considers the way in which humans interpret a soundfield. Humans sample a space from two points, the ears, which sit approximately 16cm apart for the average male (Plaga et al., 2005). This means that a sweet spot must be preserved of minimum diameter ≈16cm in order to accurately reproduce a source for a human subject. Often, this would require significant frequency limitations or the use of very high order Ambisonics ($>$ 36th order (Zaunschirm, Schörkhuber and Höldrich, 2018)). For example, Fig. 5.1 shows the similarly sized sweet spots of a 5th order decoder reproducing a source of just 2500Hz and a 36th order decoder reproducing a source of 20000Hz.

These limitations are almost unavoidable in a real world scenario. However, by exploiting the cross-talk exempt nature of binaural rendering (i.e. the signal sent to the left ear is completely isolated from the right and vice versa) the signals sent to either ear may be independently optimized. By doing this, higher frequency reproduction may be preserved at lower order Ambisonics. One such optimisation is to manipulate the central location of the sweet spot in order to shift it to the ideal sampling location, i.e. the position of each ear. This is referred to as Time-Alignment (Zaunschirm, Schörkhuber and Höldrich, 2018). An alternative workflow is to duplicate the loudspeaker feeds and generate a pair of independent virtual loudspeaker arrays each centered precisely around either ear. This is referred to as BiRADIAL. The two methods are outlined in Fig. 5.2 and are discussed further throughout this chapter.

The work presented in this chapter has been published by Armstrong, Murphy and Kearney, 2018; Armstrong et al., 2018b.

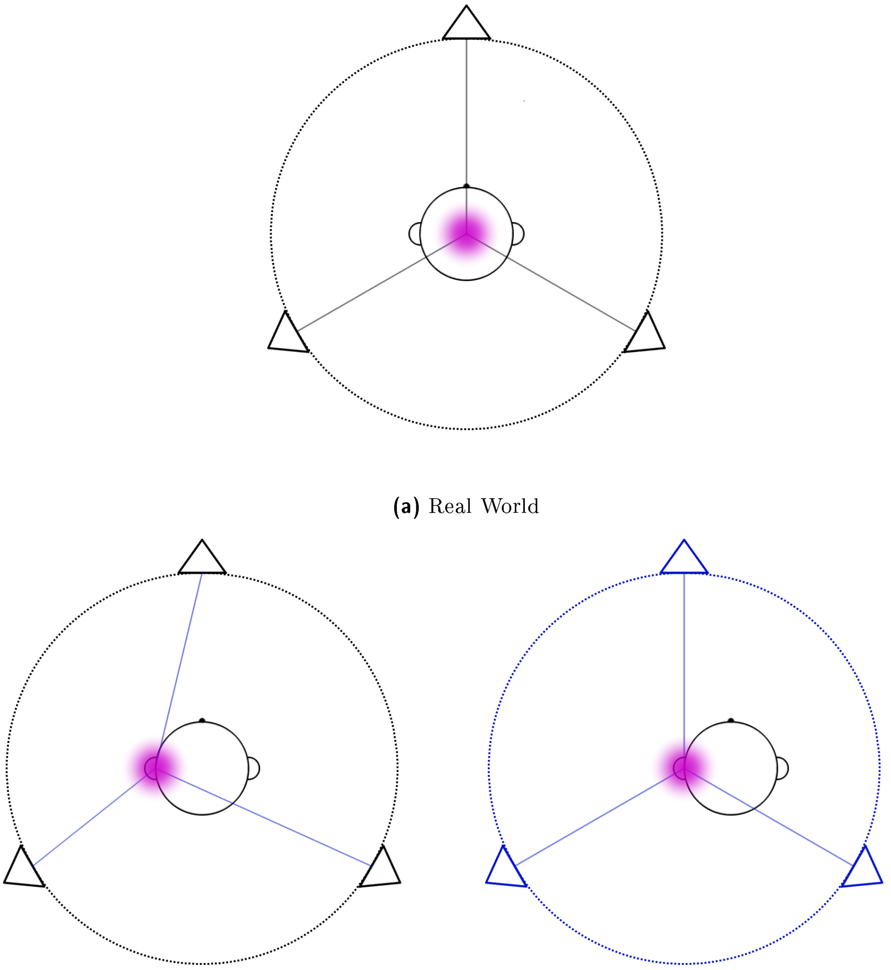
5.2 Manipulation of the Sweet Spot

5.2.1 General Considerations

Ambisonic reproduction is generally considered to be optimal at the precise center of a standard spherical playback array. However, this position may be considered more generally as the point in space at which every loudspeaker is temporally equidistant i.e. the point in space at which an impulse played simultaneously through each loudspeaker would meet. When the total time delay of each virtual loudspeaker is equal then there is no possibility of destructive comb filtering between the reproduced loudspeaker signals and hence there is no spatial aliasing.

By imposing variable delays onto each loudspeaker feed, the point in space at which the signals 'meet', and therefore the boundary of accurate reproduction i.e. the sweet spot, may be redefined.

Similar techniques have previously been proposed in various forms but usually within the context of time aligning HRTFs within a spherical harmonic representation. In 1998 Evan's showed how the removal or equalisation of the HRTF onset times caused by ITD before they were encoded to spherical harmonic format reduced the energy present in higher order spherical harmonic components (Evans, Angus and Tew, 1998). Richter describes the process as optimising the sound expansion for the most

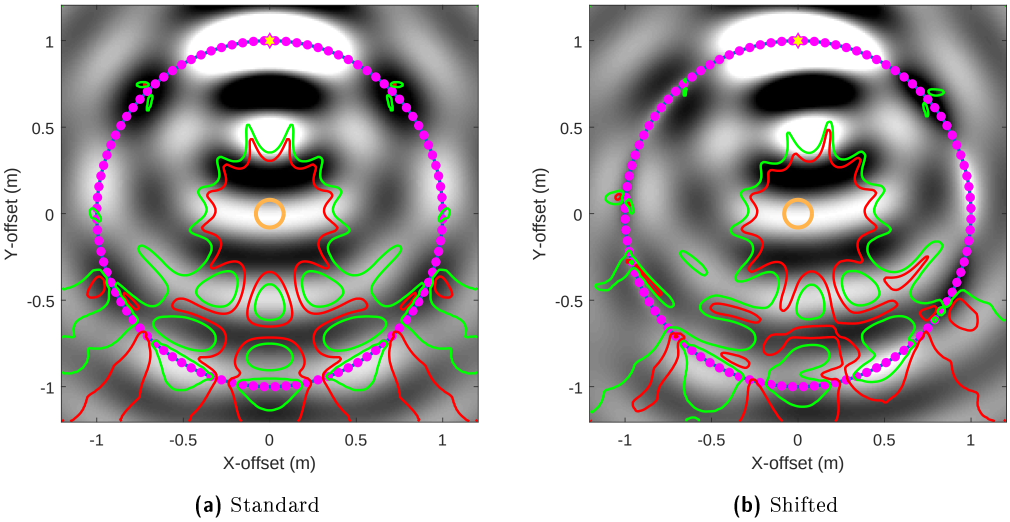


(a) Real World

(b) Time-Alignment

(c) BiRADIAL

Figure 5.2: Examples of a) real world, b) Time-Alignment and c) BiRADIAL Ambisonic rendering and the positioning and optimization the sweet spot for binaural reproduction. Note that the examples in Fig. 5.2b and 5.2c are only being shown for the left ear and would ordinarily also be performed separately for the right ear. The approximate location of the sweet spot is in each case highlighted in purple.



(a) Standard

(b) Shifted

Figure 5.3: 2D 5^{th} order Pseudo-Inverse Ambisonic reproductions of a 750Hz sinusoidal point source ($\varphi = 0°, \vartheta = 0°, r = 1$) with a zero-to-peak amplitude of 1 at the center of the array. Loudspeakers (pink dots) have a radius of 1m. Amplitude is plotted on a capped colour scale: -1 to 1 black to white respectively. An orange ring of radius 8cm indicates the approximate size of a human head. Areas of accurate reproduction are highlighted by contours: $<$ 20% error (green); $<$ 10% error (red). The ability to shift the area of accurate reproduction is shown by delaying the loudspeaker feeds such that the center of the sweet spot is the point in space at which the loudspeaker signals are temporally equidistant. In this case the loudspeaker feeds have been delayed with respect to the far-right edge of the orange circle (i.e. the right ear).

compact transformation (Richter et al., 2014). Similar findings were presented by Zaunschirm, Schörkhuber and Höldrich, 2018. They go on to present a binaural renderer that uses these time aligned HRTFs and shows significant improvements in spectral reproduction.

Considering this workflow it may be said that by time aligning the HRTFs they are better represented by the lower order spherical harmonic components and thus reconstruction of the soundfield is more accurate at higher frequencies given a set order. In a sense, the correct spectral reproduction may be managed by the lower order components, say 1^{st}-5^{th} order, but in order to correctly reproduce the spectrum over a broad area of space (i.e. with the inclusion of temporally spaced characteristics such as ITDs) higher order components must be utilized, say 5^{th}-36^{th} order.

In a real world scenario such a technique may be used to re-focus the area of reconstruction toward a particular listener situated away from the center of the array. However, as the technique has little effect on the size of the sweet spot such a system would still be limited to maximum reproduction frequencies at low orders.

The technique is demonstrated in Fig. 5.3. The delay imposed onto each loudspeaker feed may be calculated using Pythagorean triangles as the path difference between

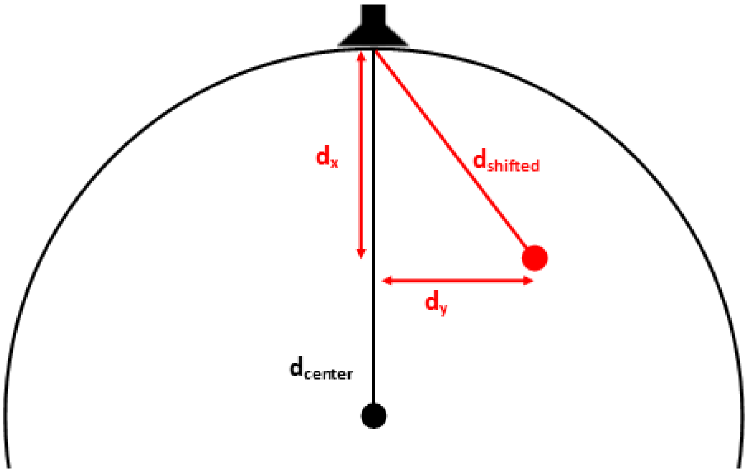


Figure 5.4: Calculation the path difference and hence delay of a loudspeaker to shift the sweet spot. Note this calculation must be made independently for each loudspeaker.

the loudspeaker and a) the center of the original sweet spot (the center of the array) and b) the center of the new sweet spot. With reference to Fig. 5.4

$$\Delta d = d_{\text{center}} - d_{\text{shifted}} \tag{5.1}$$

$$= d_{\text{center}} - \sqrt{d_x^2 + d_y^2} \tag{5.2}$$

$$\Delta t = \frac{\Delta d}{c} \tag{5.3}$$

where c is the speed of sound, d_{center} is the distance from the loudspeaker to the center of the array and d_{shifted} is the distance from the loudspeaker to the center of the new sweet spot.

Fig. 5.3 shows that the technique does not simply shift the placement of the original accurately reconstructed soundfield, but rather the bounded area within which the reconstruction is accurate. Note the subtle change in the angle of curvature within the green and red contours of the reconstructed waveforms and the direction of propagation tracking the wavefronts back to the modelled source. Both cases are consistent with the original source located straight ahead on the radius of the loudspeakers.

5.2.2 Binaural Reproduction

In binaural reproduction there is complete isolation between the left/right signals that are presented to a listener through either headphone. There is no restriction to present an identical set of virtual loudspeaker feeds to each ear. It is therefore possible to individualise the spatial rendering process for each of a listener's ears. Two modified sets of virtual loudspeaker feeds can therefore be presented that shift the sweet spot to center around the precise locations of each ear independently (Zaunschirm, Schörkhuber and Höldrich, 2018). A left-ear-centered sweet spot to the left ear and a right-ear-centered sweet spot to the right ear.

It is understandable to assume that such a technique would require twice the complexity but this is not the case. Rather than requiring l stereo convolutions, the technique simply requires $2 \times l$ mono convolutions. As it transpires, the manipulation is trivial and may be performed by simply time delaying/advancing the HRTF filters with fractional sample delays rather than adjusting of the loudspeaker signals themselves. This requires no alterations to be made to a standard decoding and rendering workflow other than to swap out the HRTFs. A simple geometric or spherical model may be used to calculate the correct time shift needed for each channel of each HRTF. It is dependant on the position at which the HRTF was measured and width of the person's head. Similar to Fig. 5.4, the distance between the HRTF source and the relevant ear is calculated and compared to the radius of the source (the distance between the source and the center of the array). The difference in distance is converted to a time delay and the HRTF channel is shifted appropriately.

This technique largely removes the issue of spatial aliasing by equalling the path distances to each loudspeaker from each ear. It is therefore possible to use dense sets of HRTFs, in the order of hundreds to > 10000 measurements, for even low Ambisonic order rendering without the overlapping virtual microphone patterns causing huge comb filtering. Advantages of using dense HRTF sets are discussed in Section 5.4. Whilst these numbers of measurements are currently available for dummy head recordings (e.g. the SADIIE database, see Section 3.3, or those measured by Bernschütz, 2013), it has so far not been possible to capture such high resolutions for human subjects due to comfort and time constraints. However, the reader is referred onwards to Chapter 6 in which a solution is proposed to the fast capture of dense HRTF sets of human subjects.

Similar methods of optimisation have been found independantly by means of error minimisation functions such as Magnitude Least Squares (Schörkhuber, Zaunschirm and Robert, 2018). This method takes the approach of iteratively adjusting the

HRTFs in order to find the optimal time and frequency domain filters that result in the most accurate binaural reproduction. The drawback of this technique is that it relies upon a dense set of reference HRTFs to compare the Ambisonically rendered signals to in order to minimise the error between the two. Nevertheless, confidence may be gained from the fact that for dummy head recordings this technique also found that approximately time aligning the HRTFs was optimal to improving spectral reproduction.

5.2.3 Consideration of ITDs

Time aligning HRTFs helps to preserve the spectral characteristics of the virtual soundfield. However, the technique introduces significant problems with regards to temporal based reconstruction, e.g. ITDs. By aligning the stereo channels of the HRTFs the time difference characteristics have been removed. Although this simplifies the measurements in a spherical harmonic sense, it means that a virtually rendered source will appear to arrive at both ears at exactly the same time irrespective of its angle with respect to the head.

ITDs may be preserved within a Time-Alignment renderer through the use of multi-band HRTFs that are comprised of standard measurements at low-frequencies and time aligned measurements at high frequencies only. This is the same approach as was taken by Zaunschirm, Schörkhuber and Höldrich, 2018. Ambisonics has been shown to accurately reproduce both spectral and ITD cues within a low frequency band, the width of which is dependant on order as presented in Table 4.3. Therefore, there is no drawback to using a standard HRTF decoding technique within this region and preserving the restoration of ITDs.

Fig. 5.5 summarises some of the relevant frequency bands. A decision must be made as to the most appropriate crossover frequency in the case of lower order Ambisonics (in particular $M < 3$) as the maximum frequency of accurate reproduction is below that at which ITD cues remain dominant. This is discussed further in Section 5.3. The multi-band HRTF filters are referred to in this thesis as Hybrid HRTFs to differentiate the decoding technique from the already established dual-band (basic/max-$\boldsymbol{r}_E$) decoding matrices, discussed in Section 4.8.2.

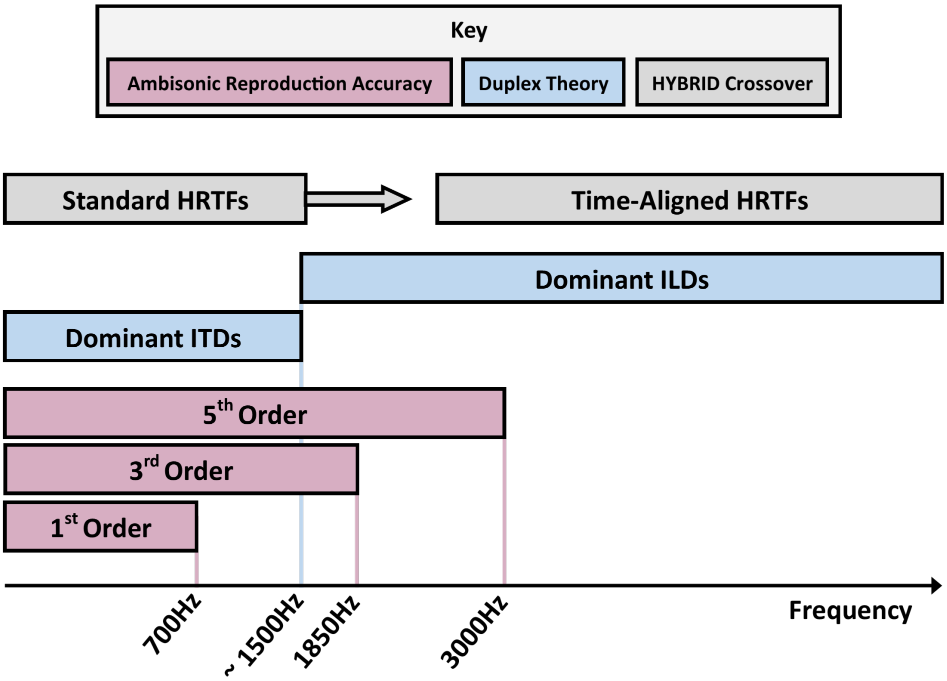


Figure 5.5: A summary of some of the relevant frequency bands for Hybrid HRTFs. The approximate regions of accurate reconstruction are shown for 1st and 5th order Ambisonics; the regions of ITD and ILD dominance; the approximate crossover region of the Hybrid HRTFs.

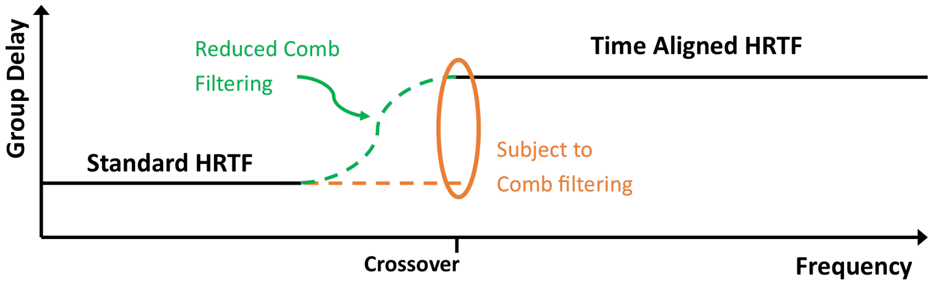


Figure 5.6: Implementation of a gradual group delay filter to smooth the transition between the standard and Time-Aligned HRTFs.

5.2.4 Applying a Smooth Group Delay

Caution must be taken in the crossover of standard to time aligned HRTFs. The process combines multiple time-delayed versions of the same filter. Any overlapping frequencies will therefore be subject to comb filtering and hence will be reduced in amplitude in the resulting filter.

To counter the issue, a gradual and incremental delay may be imposed in the lead up to the crossover frequency by means of an all-pass group delay filter. The principle of this crossover transition is shown in Fig. 5.6. The filter must be calculated individually for each HRTF as the difference in time delay is unique to each angle. By taking this approach, no single frequency is overlapped with a copy of itself

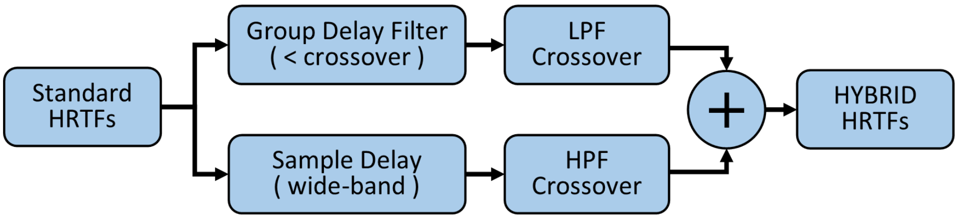

Standard HRTFs
Group Delay Filter (< crossover)
LPF Crossover
Sample Delay (wide-band)
HPF Crossover
+
HYBRID HRTFs

Figure 5.7: Block diagram outlining the practical 2-stage process of creating time aligned Hybrid HRTFs using a reduced-band group delay filter and fixed sample delay

that is of significantly different delay and hence destructive interference is therefore minimised.

The design of accurate wide-band group delay filters is not a trivial task. The following 2-stage approach is therefore implemented in practice. First, a narrower-band group delay filter is designed for the standard HRTFs up to the crossover frequency. This filter should preserve the original time delay of the HRTF for as long as possible before gradually increasing/decreasing the delay to that of the corresponding time aligned HRTF within a few hundred Hz of the crossover frequency. Beyond the crossover frequency region, the response of the group delay filter may be defined as a *don't-care* state. This simplifies the design of the filter, which may be computed using the Matlab function `iirgrpdelay()`.

A separate set of precise and absolutely time aligned versions of the HRTFs may then be computed using a simple sample delay function. At the crossover frequency, linear phase crossover filters are used to overlay the group delay filtered standard HRTFs and the sample delayed time aligned HRTFs. This is shown in Fig. 5.7.

5.2.5 Crossover Frequency

The time delays being implemented within the time alignment approach are short, less than ≈ 0.4ms. This is due to them compensating for the ITD and therefore being of the order $\pm$ half the ITD in each ear. The delays are therefore comparable to the periods of lower frequency components of the HRTFs. As such, it is beneficial to keep the crossover frequency as low as possible. By minimising the crossover frequency the wavelengths of the overlapping frequencies are maximised. The fixed time delays therefore result in a shorter phase delay and consequently less destructive interference.

Ideally the crossover frequency would be above the highest frequency at which a subject may still perceive dominant ITDs, but below the spatial aliasing frequency of the standard Ambisonic reproduction. With higher order Ambisonics (> 3) this is not a problem, however, at lower orders this is not always possible.

The preservation of ITD cues requires that time alignment of the HRTFs does not begin until after approximately 1.5-2kHz. However, the approximate spatial aliasing frequencies of 1^{st}-3^{rd} order Ambisonics fall below these frequencies for an average sized human head. The result is that for low order Ambisonic reproductions there is a small frequency gap within which spectral reproduction accuracy must be sacrificed for the preservation of ITD cues for localisation.

5.3 The Need for Decoder Matrix Weightings

The implementation of decoder matrix weightings was previously discussed in Section 4.8 to exploit perceptual localision cues beyond the frequency ranges of accurate spectral reproduction. However, by shifting the location of the sweet spot the spatial aliasing frequency is essentially boosted and therefore the need for such optimization is reduced.

Nevertheless, 2 frequency regions remain in which accurate spectral reconstruction is still not possible. The first is in the case of low order Ambisonic reproduction between the standard spatial aliasing frequency and the minimum frequency at which time alignment may be used whilst preserving dominant ITD cues (e.g. 700 - 1500Hz). The second is in a very high frequency band approximately $> 10 - 15$kHz. It is not immediately obvious as to the exact cause of the errors within this high frequency band. However, one explanation is that despite being perfectly centered around each ear, the sweet spot must still enclose a certain proportion of the pinnae in order to accurately reproduce high frequency features. At frequencies above 10kHz it is quite possible that the area of accurate reconstruction simply does not enclose the necessary physiological features required to reproduce the correct binaural cues.

As per Eqn. 4.63 a frequency of 10kHz corresponds to a path difference of 1.7cm which is of similar order to the radial distance that covers the main features within a person's ear. Further, shadowing of the loudspeaker signals by the head, torso and ear itself will impact the restored waveforms.

In addition, as the soundfield is being restored from multiple sources there is no guarantee that reflections off of an uneven surface (the person's ear) will exactly

Table 5.1: A description of a quad-band decoder for binaural Ambisonic rendering. Note that band 2 is optional and depends on the spatial aliasing frequency of the standard decoding being below 1.5-2kHz.

Band	Description	Approx. Freq	HRTFs	Matrix Weighting
1	Standard Reconstruction		Standard	Basic
2	> Standard Spatial Aliasing		Standard	max-$\boldsymbol{r}_E$
3	Time-Aligned Reconstruction	> 1.5-2kHz	Time-Aligned	Basic
4	Unavoidable Spectral Error	> 10-15kHz	Time-Aligned	max-$\boldsymbol{r}_E$

replicate those as if the soundfield had been generated from an ideal point source. Pinnae reflections are found to be significant at frequencies above 4kHz and so it is natural for us to expect some error beyond these frequencies.

In these cases, it may be appropriate to again resort to a perceptually motivated decoder such as max-$\boldsymbol{r}_E$ weighting. A tri- or even quad-banded decoder is then perceivable as described in Table 5.1. Conveniently, for orders of Ambisonics above 3$^{\text{rd}}$ the second band (standard HRTFs, max-$\boldsymbol{r}_E$ weighting) becomes redundant and the crossover frequencies may be fixed for all orders.

5.4 Compatible Decoders

In a standard Ambisonic rendering scenario significant destructive interference is encountered outside of the sweet spot that worsens with respect to the number of loudspeakers utilized, the reader is referred back to Fig. 4.11. As a result, high frequency sources are often poorly reproduced for human listening by oversampled arrays. However, by time aligning the virtual loudspeaker feeds the influence of the reconstructed soundfield from outside of the sweet spot is greatly reduced. Whilst the technique is still limited by Eqn. 4.62 ($L \geq K$) in order to avoid gaps in the soundfield there is no longer a restriction as to the maximum number of loudspeakers.

There are a few advantages of using a higher number of loudspeakers. Firstly, the timbral consistency of a panned source increases as the reconstructed soundfield becomes less dependant on the alignment of the source with a particular loudspeaker direction. This was shown graphically in Fig. 4.11 to improve with a higher number of speakers.

Secondly, the compatibility between decoders designed for different Ambisonic orders is increased. Individualising the loudspeaker count and layout for each order requires multiple decoding matrices/spherical harmonic HRTF files to be stored and switched depending on the input signal. However, by defining the same, dense, loudspeaker configuration that satisfies $L \geq K$ for all foreseeable orders only a single decoding matrix /spherical harmonic HRTF representation is required. Only the relevant data may then be extracted from that file up to the order required. For example, take an L by K (e.g. 1000 HRTFs by 36 spherical harmonic channels) decoding matrix compatible with up to 5^{th} order Ambisonic reproduction. If presented with a 1^{st} order input file only the first 4 columns would be utilised in the decoder convolutions (e.g. 1000 by 4).

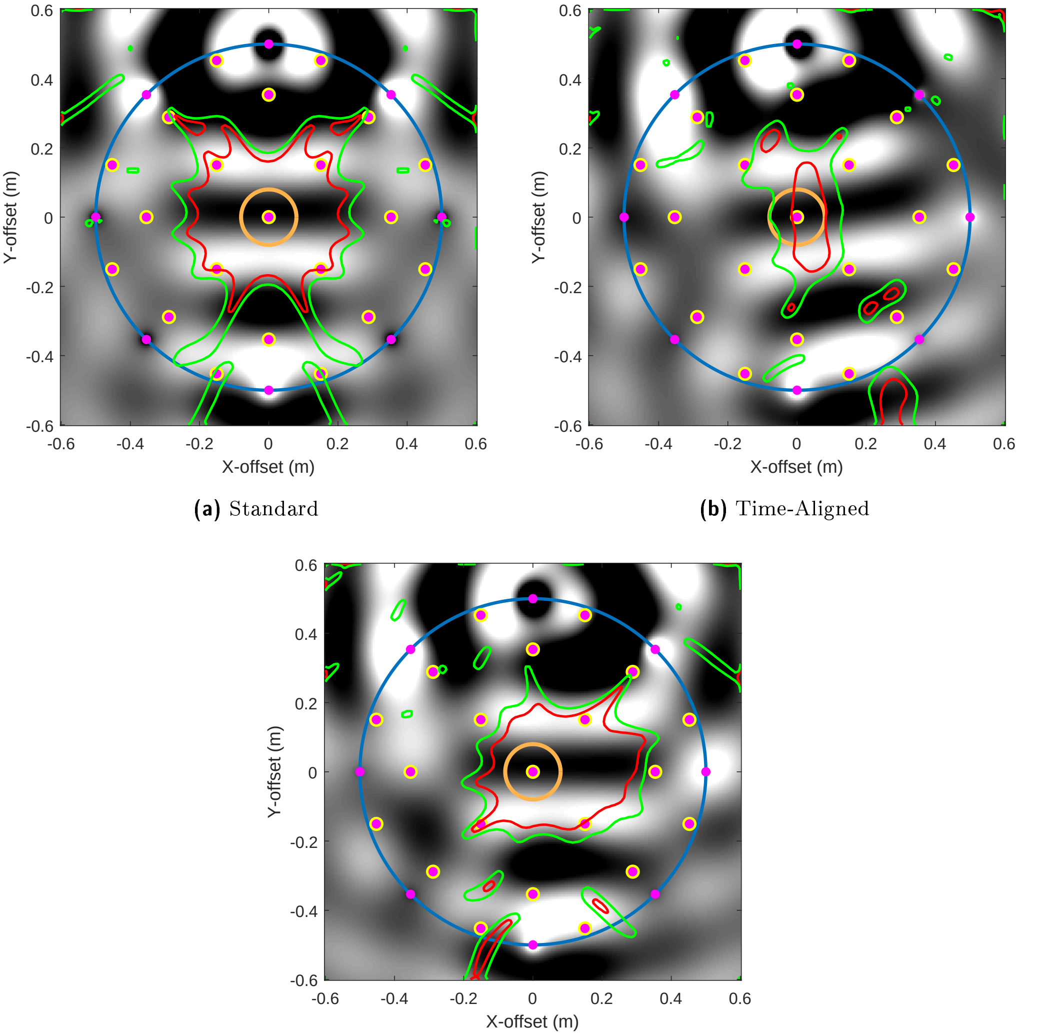
Whilst the technique would encounter significant computational (or indeed practical) costs in the case of a virtual (or real) loudspeaker render, there is no difference in complexity if the binaural signals are computed in the spherical harmonic domain. The number of convolutions, equal to the number of Ambisonic channels, remains the same whether 1 or 1000 loudspeaker sources are encoded.

The only drawback is in implementing a decoder matrix weighting scheme (e.g. max-$\boldsymbol{r}_E$) within the *one-size-fits-all* decoder. As the spherical harmonic component weightings are order dependant individual files would still be required in each case, or at least a mechanism for adjusting the weightings of each channel would need to be defined within the renderer.

5.5 Effects of Near-Field Control Filters

Other than the removal of the time-dependant characteristics of the reproduced soundfield there is one other potentially major drawback of implementing the Time-Alignment approach. The issue relates to how Near-Field Control (NFC) filters affect the wavefront reconstruction and how the effects are skewed by the Time-Alignment approach. First, consider how NFC filters, discussed in Section 4.6.11, achieve the effect of distancing a source. They manipulate the frequency dependant amplitude and phase response of the surrounding loudspeakers in order to reshape the wavefront curvature to simulate a source of distance other than that of the loudspeaker radius. However, what the filters are unable to change is the true position of the actual(/virtual) loudspeakers.

Fig. 5.8 demonstrates the result of applying NFC filters to standard and time aligned Ambisonic renderers. Fig. 5.8a shows the effect of NFC filters used to modify the



(a) Standard

(b) Time-Aligned

(c) Time-Aligned (re-panned)

Figure 5.8: 3D 5$^{\text{th}}$ order Pseudo-Inverse Ambisonic reproductions of a 1250Hz sinusoidal point source ($\varphi = 0°, \vartheta = 0°, r = 10$) with a zero-to-peak amplitude of 1 at the center of the array. Loudspeakers (pink dots for elevations, $\vartheta \geq 0°$, and larger yellow dots otherwise) have a radius of 0.5m. Amplitude is plotted on a capped colour scale: -1 to 1 black to white respectively. An orange ring of radius 8cm indicates the approximate size of a human head. Areas of accurate reproduction are highlighted by contours: $<$ 20% error (green); $<$ 10% error (red). It is shown that by time aligning the loudspeaker signals (comparing Fig. 5.8a and 5.8b) the angle of incidence of the wavefronts and therefore the simulated origin of the rendered source has changed. A correction technique is demonstrated in Fig. 5.8c by encoding the source at an angle of 9.2°to make it appear straight ahead with respect to the right ear.

wavefronts to appear more planar, as if the source were originating from a distance further than the loudspeaker radius. In this case the source has been rendered at 10m within a 0.5m radius loudspeaker array. Fig. 5.8b shows the same rendered source time aligned to the position of a person's right ear. It can be seen that although the shape of the wavefront has remained planar, the origin of the wave has not changed. It is fixed at the true source, the frontal loudspeaker. The NFC filters have only acted to modify the wavefront curvature, not its point of origin on the loudspeaker radius. The reconstructed source is hence simulated at a distant location that is no longer consistent with the direction of the original source. This effect holds for all frequencies.

One solution is to encode separate Ambisonic files to be decoded to each ear individually. In this sense, it is possible to control both the wavefront curvature and direction of incidence with respect to each ear independently. For example, a source may be encoded at such an angle that at a radius of 0.5m it would lie directly in front of the listener's right ear. This is calculated to be 9.2° in the current example. Fig. 5.8c demonstrates how this corrects for the skewed observation point of the time aligned approach.

However, such technical manipulation is hardly appropriate. Firstly it must be known, in advance, that a Time-Alignment decoder is being utilised and within which frequency bands it is being implemented. The loudspeaker radius, and ideally the width of the listener's head must also be known and further, at least double the number of Ambisonic channels would need to be transmitted between the encoding and decoding stages (separate soundfield representations for the left/right ear decoders). To avoid interference, minor discrepancies in the time delay of a signal now originating from multiple points on the surface of the sphere that could position it either closer to or further from each individual ear now also need to be accounted for. A selection of these issues are then complicated further by the introduction of head-tracking. A more compact and general solution is therefore proposed, particularly in the case of far-field (planar) sources, referred to as BiRADIAL.

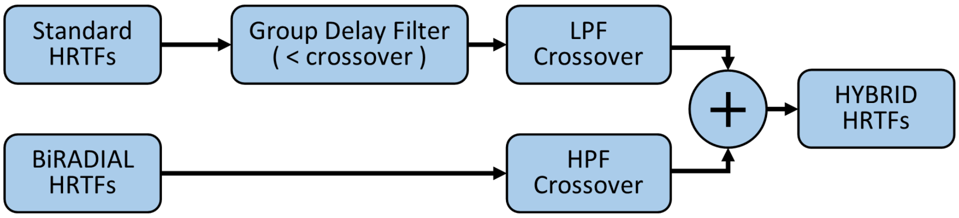


Figure 5.9: Block diagram outlining the practical 2-stage process of creating BiRADIAL Hybrid HRTFs using a reduced-band group delay filter

5.6 BiRADIAL Rendering

5.6.1 Overview

To counter the inaccuracies of distance rendering using NFC filters and a Time-Alignment approach, an alternative and more general strategy has been developed - BiRADIAL. Rather than to time align a set of head-centered HRTFs, it is proposed to re-measure two sets of ear centered HRTFs that are, by their nature, also time aligned. Again, the method may be utilised within a Hybrid HRTF as described in Fig. 5.9.

A set of left-channel only HRTFs are measured at positions centered around the left ear and a set of right-channel only HRTFs are measured at the same set of positions around the right ear. The two mono sets of measurements are then combined into a single stereo set to be used in the renderer. Conceptually, two independent virtual loudspeaker arrays are now rendered (one centered around each ear) all without any increase in complexity (increase in the number of channels). The approach is demonstrated in Fig. 5.10

The method is particularly applicable to rendering far-field/planar sources. In this case the direction of incidence of a source with respect to each ear is the same, or at least very similar. It is therefore appropriate to decode the same Ambisonic file to each virtual array. The technique is demonstrated in Fig. 5.11. In the graphic the rendered locations of three sources (ideally positioned to the front, left and rear of the listener) are shown for both the left and right virtual loudspeaker arrays. The left graphic shows the initial rendered location of each source. The right graphic shows the rendered location of each source after a 90° anti-clockwise rotation of the listener and corresponding 90° counter-rotation of each Ambisonic soundfield. Sources are inherently presented to the listener as plane waves due to the lack of

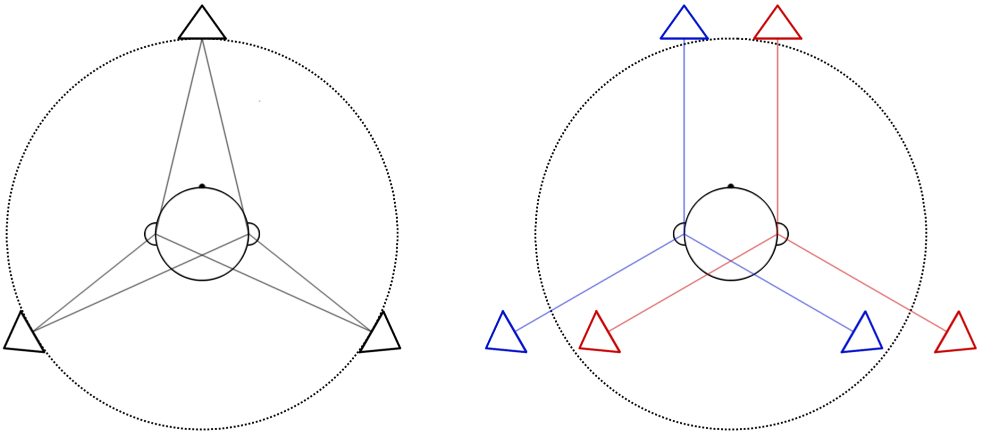


Figure 5.10: Comparison of the virtual loudspeaker arrays rendered for standard Ambisonic decoding (left) and Bi-RADIAL Ambisonic decoding (right). Note the same number of straight lines in each case indicating the same level of complexity. 3 x stereo = 6 x mono.

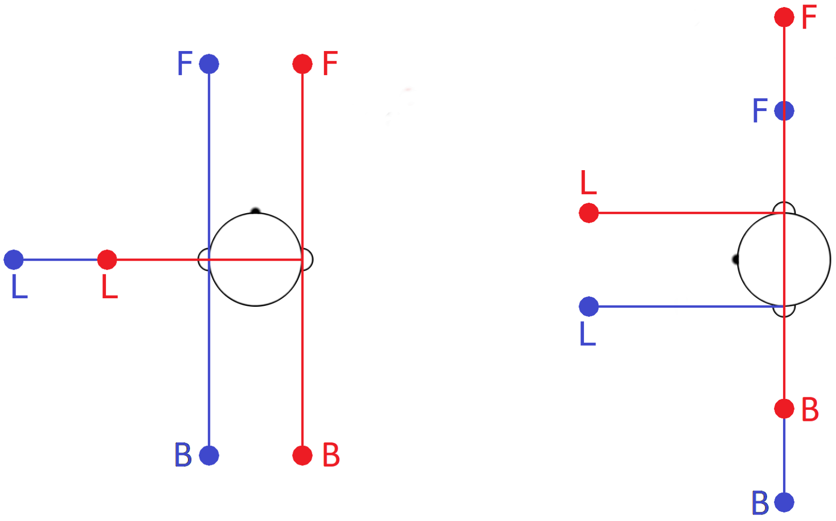
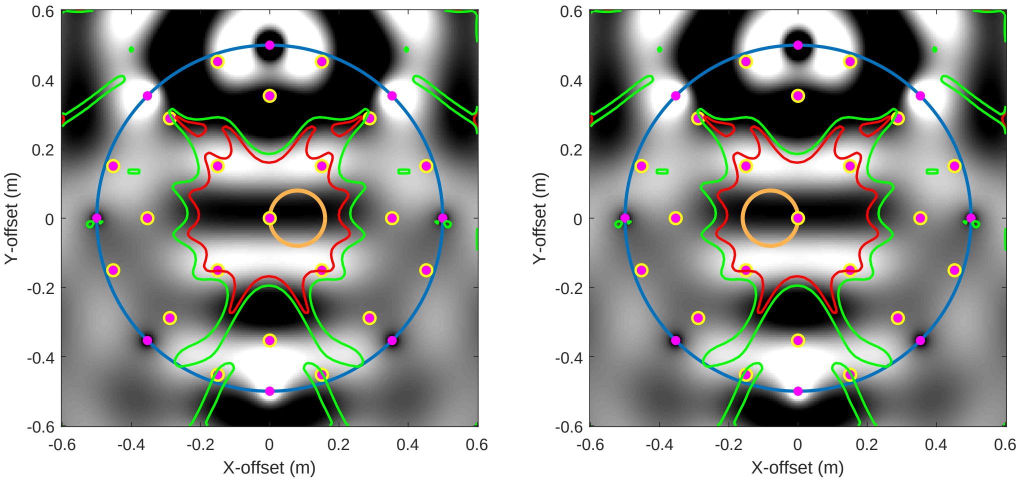


Figure 5.11: The rendered locations of three sources (ideally positioned to the Front (F), Left (L) and Back (B) of a listener) for left (blue) and right (red) Bi-RADIAL loudspeaker arrays. Locations are shown for two listener orientations. Note the corresponding 90° counter-rotation of the soundfield.



(a) Left-Ear Aligned

(b) Right-Ear Aligned

Figure 5.12: 3D 5^{th} order Pseudo-Inverse Ambisonic reproductions of a 1250Hz sinusoidal point source ($\varphi = 0°, \vartheta = 0°, r = 10$) with a zero-to-peak amplitude of 1 at the center of the array. Loudspeakers (pink dots for elevations, $\vartheta \geq 0°$, and larger yellow dots otherwise) have a radius of 0.5m. Amplitude is plotted on a capped colour scale: -1 to 1 black to white respectively. An orange ring of radius 8cm indicates the approximate size of a human head. Areas of accurate reproduction are highlighted by contours: $< 20\%$ error (green); $< 10\%$ error (red). Identical loudspeaker feeds are output from two independent virtual loudspeaker arrays. However, one has been precisely measured and reconstructed around the left ear, and the other around the right. By doing this the placement of the sweet spot is optimised on the ear in each case, whilst maintaining the correct directions of incidence and wavefront curvature for a far-field source.

any parallax but appropriate NFC filters may also be used to simulate a planar wavefront curvature. This is shown through soundfield simulations in Fig. 5.12.

5.6.2 Adjusting for Non-Planar Sources

Although a greater sense of distance and externalisation is often a desirable outcome of spatial audio there are cases in which one may wish to render a source close to the head. In that case this method may still be applicable albeit with some drawbacks. To render a source at a finite distance NFC filters may be altered or even left out entirely in order to retain some curvature in the wavefront, as is typical of a near-field source. Unfortunately, this will have the effect of spatially widening or simulating a discontinuity of frontal/rear sources. This would be equivalent to rendering two separated cross-talk exempt near-field sources - one to either ear. Consider the multiple virtual loudspeaker arrays, the duplicate rendering of each source and the positions of each source relative to the head in Fig. 5.11.

Alternatively, multiple Ambisonic files may be encoded in a similar but more robust way than in the case of adjusting source direction for time aligned decoders, as in Fig. 5.8c. The reason for the increased robustness being that the radius of the

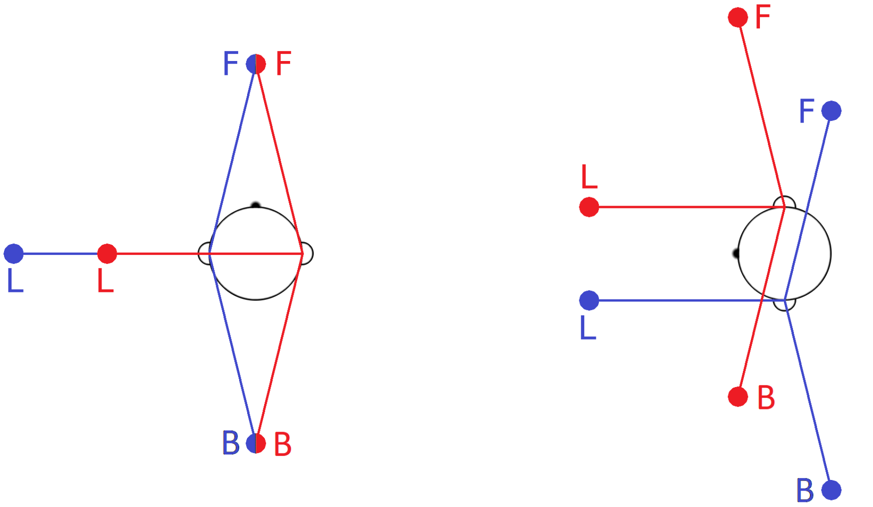


Figure 5.13: The rendered locations of three sources (ideally positioned to the Front (F), Left (L) and Back (B) of a listener) for left (blue) and right (red) Bi-RADIAL loudspeaker arrays. Locations are shown for two listener orientations. Note the corresponding 90° counter-rotation of the soundfield.

decoding loudspeaker array, head radius, and time delays are no longer critical. The BiRADIAL approach samples the loudspeaker arrays from the center, therefore the angle of incidence of a source is fixed regardless of the loudspeaker radius.

However, issues remain regarding the feasibility of head-tracking in this specific case. The example in Fig. 5.11 is adjusted for near-field sources in Fig. 5.13. Ideally, the location of each source would be coherent between the left and right ear i.e. the sources with respect to each ear should lie on top of one other. Although the sources have been directly encoded to complimentary angles for the listener's initial orientation, the accuracy of the renderer quickly deteriorates as the listener rotates. The particular method for rendering near-field sources is therefore most applicable for non head-tracked applications or scenarios in which the head movement of a listener may be restricted or limited, for example when looking at a screen.

5.7 Methods of Objective Spectral Evaluation

5.7.1 Overview

In attempting to evaluate the performance of BiRADIAL rendering objectively special care must be given to the quantitative performance of the spectral output. Human perception is complex and varies significantly from an analytical analysis of a waveform. Humans are sensitive to both frequency and amplitude variations over time with bias towards certain parts of the spectrum. However, it is possible to quantify and compensate for such sensitivities and critical listening resolution across the spectrum.

ASD, calculated from the difference between the FFTs of two audio signals, is not an accurate metric for human auditory perception. Consider an extreme example: a 0 → -10 dB difference at 1KHz is far more perceptually relevant than a -100 → -110dB difference at 10Hz, despite both being an absolute difference of 10dB.

Improving upon ASD calculations leads us towards perceptual loudness models which may be used to approximate the loudness of a particular sound taking into account the sensitivity of the auditory system. Fletcher and Munson, 1933 presented work that explored the varying perceived loudness of different frequencies despite being of equal intensity. Stevens, 1936 (Stevens, 1955) published work on the non linear link between sound intensity and perceived loudness and Zwicker, 1961; Zwicker and Scharf, 1965 developed models for approximating the summation of loudness across frequency. These models have since been revised by Moore and Glasberg, 1995. Such models can incorporate single or multi-band analysis, however, their usual application (e.g. broadcast, music production) tends to require the output of a single loudness figure only that describes the entire wide-band stimulus as a whole. This tells us nothing about the perceptual spectral differences between two stimuli other than their overall perceived magnitude, which would likely be normalized within a reproduction stage anyway.

5.7.2 Proposed Model

A new model is therefore proposed, referred to as the PSDM, that derives from the standardized ITU-T recomendation P.862: PESQ (Rix et al., 2001) and is inspired by the likes of Pulkki, Karjalainen and Huopaniemi, 1999 and Moore, Glasberg and Baer, 1997. Simply, it is a model that utilizes multi-band loudness model weightings to analyse the perceptual relevance of frequency components. The raw results are then used within a spectral comparison algorithm *before* the difference is reduced to a single representative figure. PESQ follows very similar processes to those used in PEAQ (Thiede et al., 2000), PSQM (Beerends and Stemerdink, 1994) and (Wang, Sekey and Gersho, 1992). The model outputs a single difference figure for spectral similarity which is shown in Section 5.7.4 to vary more in line with human perception than absolute difference. A flowchart of the model is shown in Fig. 5.14. Three key features are taken into account: the varying sensitivity of the ear

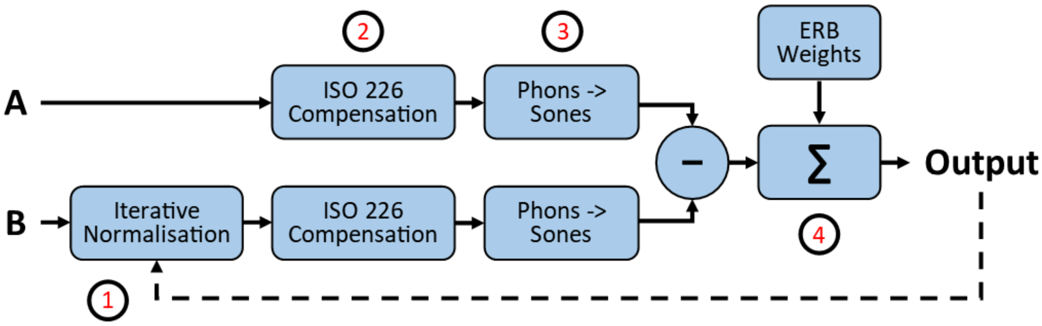


Figure 5.14: Flowchart of the PSDM showing the comparison of input spectra A and B

to different frequencies, the subjective loudness scale and frequency scale warping. PSDM is particularly applicable to binaural signals as the subjective loudness scaling is triggered by interaural level differences and hence the model weights the louder ipsilateral channels with greater relevance. A potential criticism of the model is that these sensitivities are in general defined for non-spatial signals. Ideally, a perceptual model such as this should take into account the improved clarity of sources due to their spatial separation. For example, several studies have demonstrated this in the context of speech (Glyde et al., 2013; Best et al., 2013; Brungart and Simpson, 2002; Freyman, Balakrishnan and Helfer, 2001). However, in defining a generic model that accepts only time-invariant frequency spectra as its input this is difficult to replicate. The proposed model therefore compensates only for the non-spatial theory.

5.7.3 Method

Normalisation

First, the normalisation of each input spectra (A and B) is considered relative to the other (Fig. 5.14 Step 1). An iterative process is undertaken to find the optimal level matching of the inputs to minimise the PSDM. The mean value of each input is first normalised to the other and the PSDM is calculated. An approximate course *normalisation error* is then calculated as the difference in the mean values between the perceptually weighted inputs. The amplitude of input B is then shifted by this amount, the PSDM is re-calculated and a fine normalisation procedure is then initiated.

The fine normalisation procedure begins by applying a small amplitude shift to input B (+0.2dB) and then re-calculating the PSDM. If the PSDM reduces, then the amplitude shift is repeated. If the PSDM increases, then the amplitude shift is reversed. At each reversal the magnitude of the amplitude shift is reduced by a

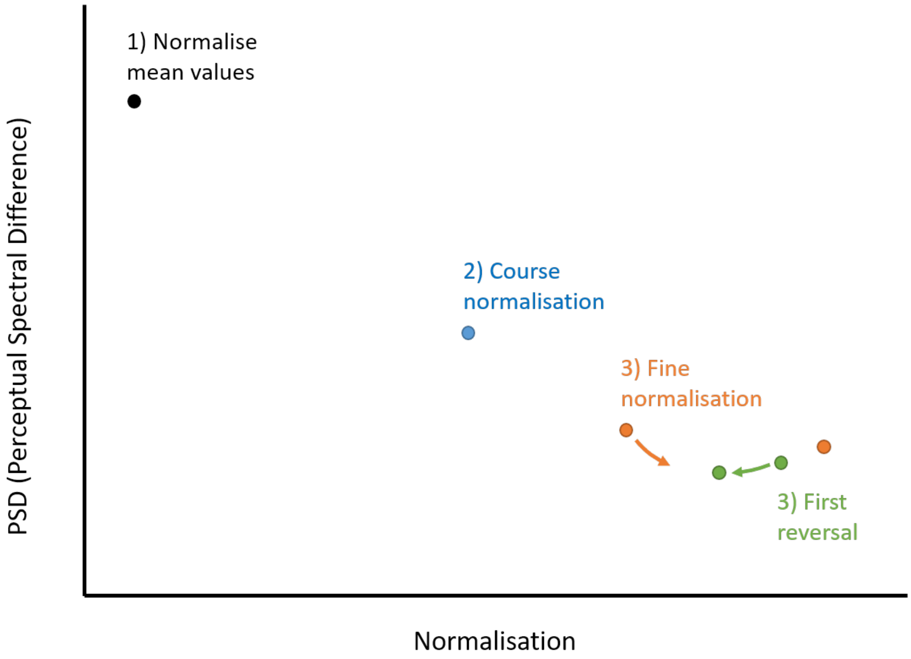
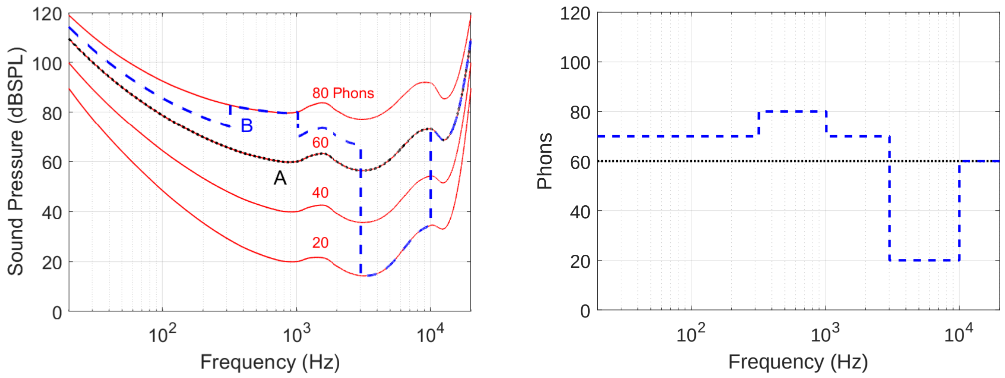


Figure 5.15: PSDM normalisation procedure including 1) the initial performance test 2) the coarse normalisation based on the difference in mean perceptually weighted values 3) the initial step in the iterative normalisation process 4) an example of the first reversal and subsequent decrease in amplitude shift.

factor of 0.4, empirically chosen. The process continues for either a set number of iterations or until the amplitude shifts fall under a certain threshold. In this theses the process continues until the amplitude shifts fall within 0.1dB. The process is described in Fig. 5.15.

Typically this normalisation procedure will offset the inputs by a couple of dB compared to, for example, simply equating the mean RMS values of the signals directly. In the case that the model is analysing multiple binaural signals a solid angle weighting option is also provided.

The solid angle weighting option attributes a weighting to a particular angle within a group of angles based on its relative position. Simply, it assigns a value to each angle such that it is proportional to the area on a sphere within which it is the closest angle. The Voronoi method is used to calculate this weighting. The result is that angles that are grouped closely together will each be assigned a low value, and angles spaced far apart will each be assigned a high value. Weighting measurements in this way prevents certain spatial locations biasing an spherically average result simply due to a high number of measurements being taken in that location.



(a) Spectra (dB)

(b) Spectra (Phons)

Figure 5.16: Conversion of 2 input spectra (A and B shown in black and blue respectively) from a dB scale to a Phons (Equal loudness) scale. Equal loudness contours are shown on the left in red and labelled in Phons.

Frequency Sensitivity

The second step considers a person's varying sensitivity to different frequencies (Fig. 5.14 Step 2). In particular this step accounts for increased sensitivity in the region between 1 kHz and 5 kHz. Frequency sensitivity has been well researched and is summarized by the ISO 226 standard (ISO, 2003). It is often referred to as the Outer-Middle-Ear (OME) filter in literature and is typically used in the pre-processing stages of auditory filter banks (Pfluger, 1997). ISO 226 defines a series of equal loudness curves that vary according to absolute reference volume as well as frequency (see Fig. 5.16). Each curve defines the dBSPL levels of different frequencies that are perceived to be the same volume as a 1kHz tone. Each curve represents a perceptual loudness level which can be measured in Phons. Unlike previous models which use a single averaged equal loudness contour filter based on the threshold of hearing (Moore and Glasberg, 1995; Moore, Glasberg and Baer, 1997), this model utilizes 90 magnitude dependent equal loudness contours in 1 dB increments from 0 to 90 dB SPL. The contours are calculated and saved within a lookup table.

The input signals are converted to the frequency domain using an FFT. It is assumed that the signals would be heard/played back at a comfortable listening level of 75dB SPL (Best, McRoberts and Sithole, 1988). As such the frequency spectra are magnitude shifted accordingly so that the average value is 75dB. The magnitude value of each frequency bin is then rounded to the nearest integer. Each value is then searched for within the lookup table of equal loudness contours. The FFT samples are replaced with the corresponding Equal loudness contour value. By doing this, the data is converted from a dB scale to a Phons scale. This process is shown in

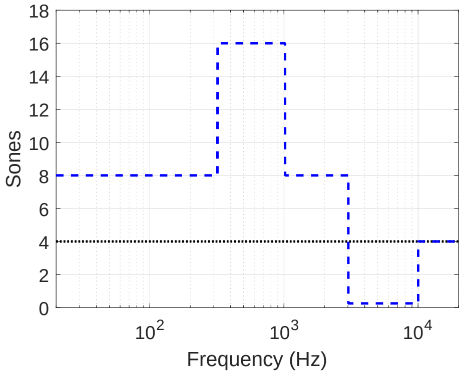


Figure 5.17: Conversion of the data in Fig. 5.16b from a Phons scale to a Sones (perceptual loudness) scale

Fig. 5.16.

Subjective Loudness Scale

Next, the magnitude value of each frequency bin is converted from a Phons scale to a Sones scale (Fig. 5.14 Step 3) using the equation

$$S = 2^{\left(\frac{P-40}{10}\right)} \tag{5.4}$$

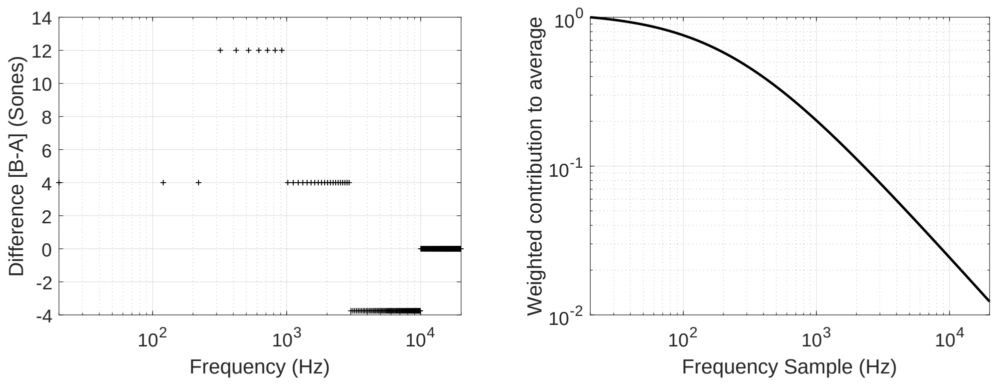
where:

- P denotes the value in Phons,
- S the value in Sones.

This conversion is shown in Fig. 5.17. The Sones scale is based on the human perception of loudness. It is well cited that at normal listening levels a drop of 10 Phons (10dB at 1KHz) is roughly equal to a 50% drop in perceived loudness (Wang, Sekey and Gersho, 1992; Stevens, 1955; Bauer and Torick, 1966; Fletcher and Munson, 1937). It is proposed in this model that perceptual error is more proportional to loudness than to amplitude, i.e. a difference of -10 → -12 Phons is only half as important as a difference of 0 → -2 Phons. This also supports the knowledge that spectral peaks are more perceptually significant than notches (Bücklein, 1981). For example, at a baseline listening level of 10dB a 10dB peak (10dB → 20dB) will result in twice the shift in loudness to a 10dB notch (10dB → 0dB).

Cochlea Frequency Sensitivity

The linearly spaced samples of an FFT do not fairly represent the approximately logarithmic sensitivity of the inner ear (cochlea). This type of finding is well doc-



(a) Linear sampling points **(b)** ERB Scale

Figure 5.18: Depiction of the linear sampling of an FFT shown on a log scale. This is not representative of the sensitivity of the cochlea and so when averaging the samples the contribution of each data point is weighted by the inverse ERB value.

umented in literature surrounding critical bands e.g. (Smith and Abel, 1999). To do this, the ERB is calculated for each frequency bin and the contribution of each sample to the average is weighted by the corresponding inverse value (Moore and Glasberg, 1995) (Fig. 5.14 Step 4). The ERB scale (Fig. 5.18b) is an approximate mapping of the ears' reduction in analytical sensitivity with respect to increasing frequency. The weighted sampling is demonstrated in Fig. 5.18.

The output of the model is the weighted average and is representative of the average perceptual difference in Sones between the two input spectra.

5.7.4 Validation

The PSDM is validated by comparing its results to an absolute spectral difference calculation for multiple test stimuli and judging the performance against known human perception. The stimuli can be divided into two categories. The first category is a group of 8 simple test signals that were specifically designed to relate to the auditory features accounted for by the model. The second category is a set of listening test stimuli taken from a study conducted by Mckenzie, Murphy and Kearney, 2019.

Test Signals

Four test scenarios were considered to ratify the separate features of the model. In each scenario, 2 complimentary test signals were generated by passing a Dirac pulse through individual band pass filters (Zölzer, 2011). Each test signal was then compared to an unfiltered copy of the signal (flat EQ) using both the PSDM and an ASD calculation. The results were compared to show which model better evaluates

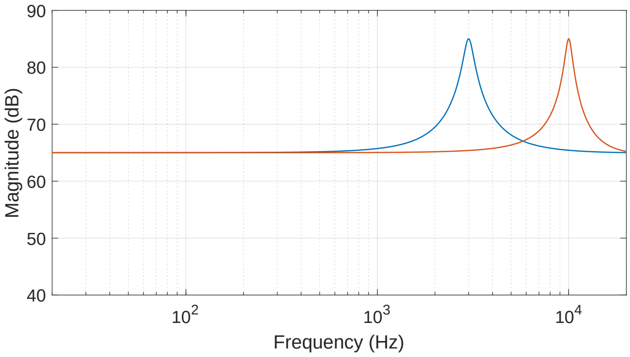


Figure 5.19: Frequency response plots of 3 kHz (blue) and 10 kHz (red) +20 dB peak filters with each bandwidth equal to the ERB of those frequencies.

Table 5.2: Results of comparing the 3 kHz and 10 kHz +20 dB peak filtered signals with filter bandwidths equal to the ERBs of those frequencies at 65 dB SPL to flat response reference signals at the same level.

	3kHz	10kHz
ASD (dB)	1.87	3.90
PSD (sones)	1.53	0.57

the differences in the test signals with regards to well researched human auditory perception.

To test the implementation of the ISO 226 equal loudness curve compensation, two signals with +20 dB peaks at 3 kHz and 10 kHz were compared to a flat 65dBSPL reference signal, see Figure 5.19. The results are given in Table 5.2.

The ASD calculation produced a higher spectral difference value for the peak at 10kHz. It is known that this should not be the case as the ISO 226 standard tells us that our ears are far less sensitive in this range. The PSD model reflected this with a higher perceptual difference value at 3kHz.

To test the handling of data on the Sones scale two scenarios were considered. The first presents the case that there are equally sized peaks in the spectrum, but at 2 different absolute reference levels. Two different reference signals, one with a flat response at 65dBSPL and one with a flat response at 45dBSPL were compared to test signals with a +20dB peak at 1kHz, see Figure 5.20. The results are given in Table 5.20.

The second scenario tested whether a peak would result in a greater spectral difference than a notch, as it should (Bücklein, 1981). Two test signals, each with either a ± +20 dB peak/notch at 1kHz were compared to a flat 65dBSPL reference signal, see Figure 5.21. The results are given in Table 5.4.

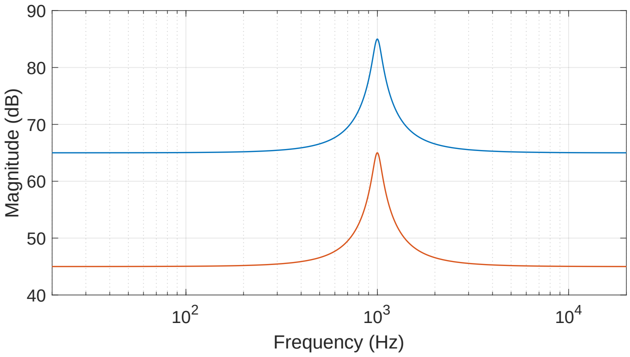


Figure 5.20: Frequency response plots of 1 kHz +20 dB peak filters at 65dBSPL (blue) and 45dBSPL (red) baseline amplitudes.

Table 5.3: Results of comparing the 1 kHz +20 dB peak filtered signals at 65dBSPL and 45dBSPL to flat response reference signals of the same respective levels.

	65 dBSPL	45 dBSPL
ASD (dB)	0.59	0.59
PSD (sones)	1.06	0.26

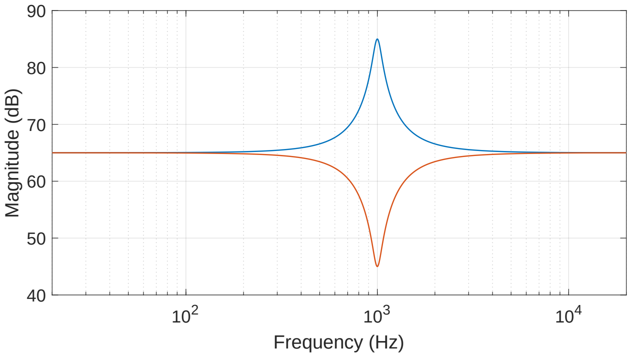


Figure 5.21: Frequency response plots of 1 kHz +20 dB peak (blue) and -20 dB notch (red) filters.

Table 5.4: Results of comparing the 1 kHz +20 dB peak and -20 dB notch filtered signals at 65 dB SPL to flat response reference signals of the same level.

	+20 dB	-20 dB
ASD (dB)	0.59	0.59
PSD (sones)	1.06	0.53

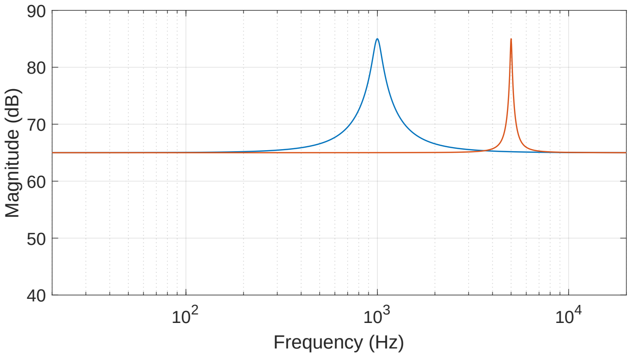


Figure 5.22: Frequency response plots of 1 kHz (blue) and 5 kHz (red) +20 dB peak filters with 100 Hz -3 dB bandwidths.

Table 5.5: Results of comparing the 1 kHz and 5 kHz +20 dB peak filtered signals with 100 Hz -3 dB filter bandwidth at 65 dB SPL to flat response reference signals of the same level.

	1 kHz	5 kHz
ASD (dB)	0.59	0.55
PSD (sones)	1.06	0.23

In each case the ASD calculation produced the same value of spectral difference for each test signal, whereas the PSDM produced a higher result for the peak and for the louder signals. Generally, the model predicts that spectra of lower absolute amplitude will result in lower perceptual differences.

The final test evaluated the use of ERB weighting to compensate for the unnatural linear frequency interval sampling of an FFT. Two signals with +20dB peaks with 100Hz -3dB bandwidths at 1 kHz and 5 kHz respectively were compared to a flat 65dBSPL reference signal, see Figure 5.22. These frequencies were chosen as they are they present with similar sensitivities on the ISO 226 equal loudness curves. The results are given in Table 5.5.

The theoretical perceptual relevance of the peak at 5 kHz should be less than 1 at 1 kHz as the peak is spread over fewer critical bands (Smith and Abel, 1999). The ASD calculation showed little difference between the two signals due to the linear frequency interval sampling, whereas the PSD predicted a much greater, and correct, difference between the two.

Perceptual Listening Test Results

The model was further tested by inputting the stimuli of a prior listening test conducted entirely separately from this thesis by McKenzie, Murphy and Kearney, 2018

Table 5.6: The 8 source directions tested as part of the PSDM's validation during comparison to human test results. They represent the faces of a dodecahedron.

	1	2	3	4	5	6	7	8
Azimuth (°)	180	50	118	0	180	62	130	0
Elevation (°)	64	46	16	0	0	-16	-46	-64

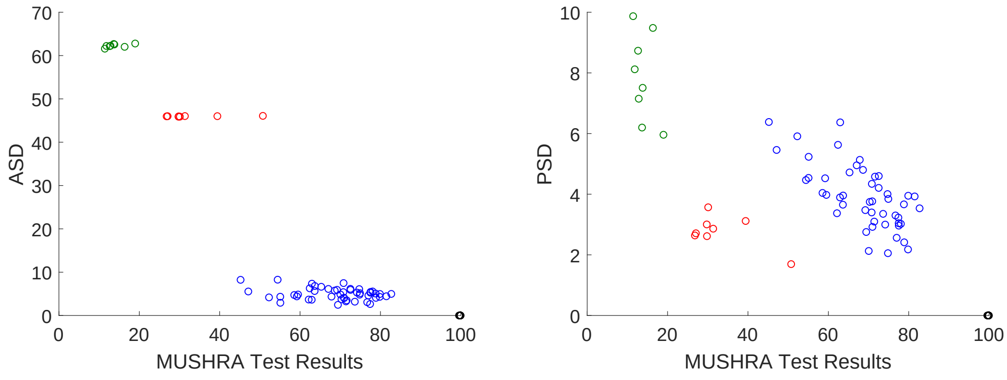
and comparing the model's output to the results of the real participants. The listening test in question was a MUSHRA style test with hidden reference, low and mid anchors (ITU, 2003b). Each subject was asked to compare a selection of statically rendered (non head-tracked) binaural stimuli to a reference and rate their timbral similarity out of 100.

6 types of test stimuli were considered: 1^{st}, 3^{rd} and 5^{th} order binaural Ambisonic reproductions of a point source with and without diffuse-field equalisation. An explanation and the implications of finite order Ambisonic reproduction are discussed in Chapter 4. For now, it is sufficient to understand that this test was chosen as it was representative of comparing a selection of timbrally similar binaural signals. Where spectral differences between the stimuli did exist they were primarily above the spatial aliasing frequency (see Section 4.7) of the Ambisonic order: approximately 700Hz, 1850Hz and 3000Hz for 1^{st}, 3^{rd} and 5^{th} order Ambisonics respectively. In the case of 5_{th} Order Ambisonic reproduction in particular the perceptual differences between the stimuli were slight.

The reference was direct non-individualized HRTF convolution. The stimuli were 1s pink noise bursts rendered at 8 different source directions, shown in Table 5.6, were tested. Each comparison was repeated twice and there were 20 participants (Aged 20-38, 17 Male, 2 non-binary, 1 female). Further details of the exact procedure are given by McKenzie, Murphy and Kearney, 2018.

The results were averaged across participants (20) and repetitions (2) to give 72 data points, 48 of which were from test stimuli (6 stimuli x 8 directions), 16 of which were from anchor stimuli (low/mid-anchor x 8 directions) and 8 of which were from hidden reference stimuli (1 reference x 8 directions). Each stimulus was compared to the true reference using the PSDM and an ASD calculation. The results from the models are presented against those from the human listening test on scatter graphs in Fig. 5.23.

The ASD calculation shows 4 very distinct groupings based on the type of stimuli presented. This is primarily due to the low pass filters used to generate the anchor



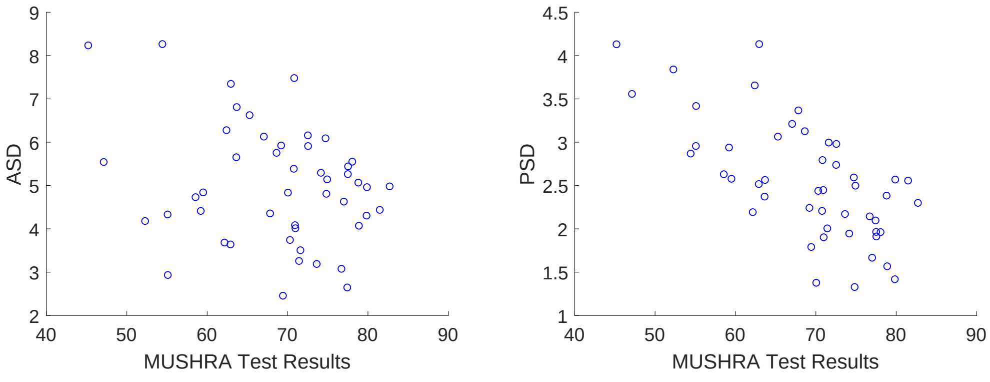
(a) ASD **(b)** PSDM

Figure 5.23: Comparison of the Absolute and Perceptual Spectral Difference calculations plotted against real MUSHRA listening test data. Data includes true reference and low pass anchors. Note that whilst the ASD is representative of an average difference between stimuli in dB, PSDM is better described as a perceptually weighted average difference between stimuli in Sones. Black: Hidden reference, Red: Mid-Anchor, Green: Low Anchor, Blue: Test stimuli.

stimuli resulting in very large changes to the spectra in the high frequencies. Similarly, the perfect spectral match of the reference stimuli is quite different to that of the general test stimuli. These large differences in the spectral responses between groups completely dwarf any more subtle intra-group differences. This highlights a significant drawback of the ASD calculation in determining perceptual similarity.

The PSDM results are more evenly spread. A strong negative correlation is evident throughout the low anchor, test and hidden reference stimuli, although the vertical spread does increase slightly within the low anchor group. What is interesting to note is the separation of the mid-anchor stimuli. The separation suggests that the PSDM is poorly predicting the perceptual change within this group. One reason for this could be the nature of the MUSHRA test. The ratings given by a MUSHRA test are self calibrated i.e. a subject first defines a scale upon which to rate the test stimuli by first rating the reference and low/mid anchors. It is possible therefore that the low and mid anchors may be rated more poorly in order to improve the resolution of the ratings given to the test stimuli. For example, a participant may first identify the two low anchors and rate them both below 10% in order to rate the remaining stimuli between 10-100% (assuming the reference is placed at 100%. A more accurate representation may be to rate the low-anchor at 10% and the mid-anchor at 40%, however, this would reduce the anticipated range within which to rate the test stimuli (now 40-100%). It is therefore possible that the mid anchor stimuli may be being biased negatively in the human listening test.

That being said, the comparison methods presented here were not designed for



(a) ASD **(b)** PSDM

Figure 5.24: Comparison of the Absolute and Perceptual Spectral Difference calculations plotted against real MUSHRA listening test data. Data includes only test material.

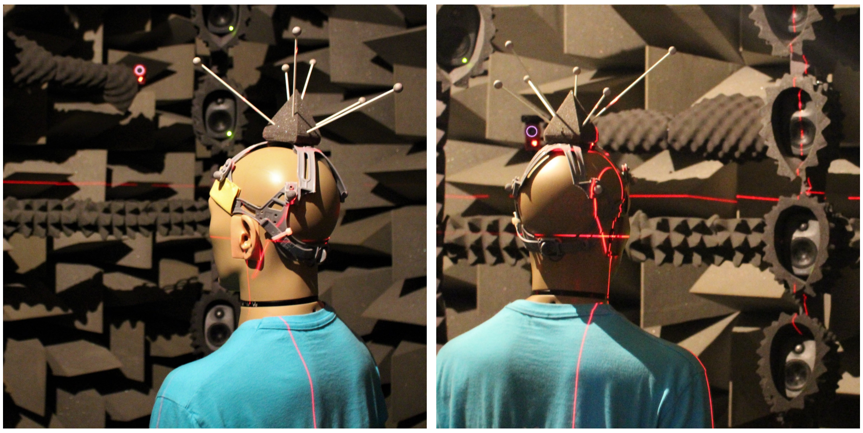
Table 5.7: Correlation coefficients (r) and p-values of the ASD calculation and PSDM results compared to real listening test data for stimuli excluding anchors and references. A p-value below 5% is deemed significant.

	r	p-Value (%)
ASD	-0.27	6.77
PSD	-0.67	0.000014

comparing stimuli that have been heavily filtered in this way (wide-band low/high pass filters). It is clear in these cases that the spectral responses will be very different. The point of the ASD calculation and PSDM is to quantify a difference between two similar spectral responses of similar shape and overall amplitude. Therefore, just the test stimuli are replotted in Fig. 5.24 in order to concentrate on analysing the better suited results.

These graphs give a much clearer representation of the realistic performance of the two techniques. It is now also appropriate to calculate the correlation coefficients of the two data sets. This data is given in Table 5.7. The data shows a significant negative correlation in the results of the PSDM confirming that the model output is indeed indicative of perceptual results. The same cannot be said for the non-significant result of the ASD calculation.

It is therefore possible to conclude that although reliable quantitative evaluation of very different frequency spectra is still a significant challenge, in comparing similar spectra the PSDM outputs results that are far more closely aligned to human perception than an ASD calculation.



(a) Left ear aligned (b) Right ear aligned

Figure 5.25: Alignment of the KEMAR mannequin during a) left and b) right ear aligned HRTF measurement as part of the SADIIE database. Note the position of the rear central laser aligned with either ear of KEMAR visible down the back on the shirt

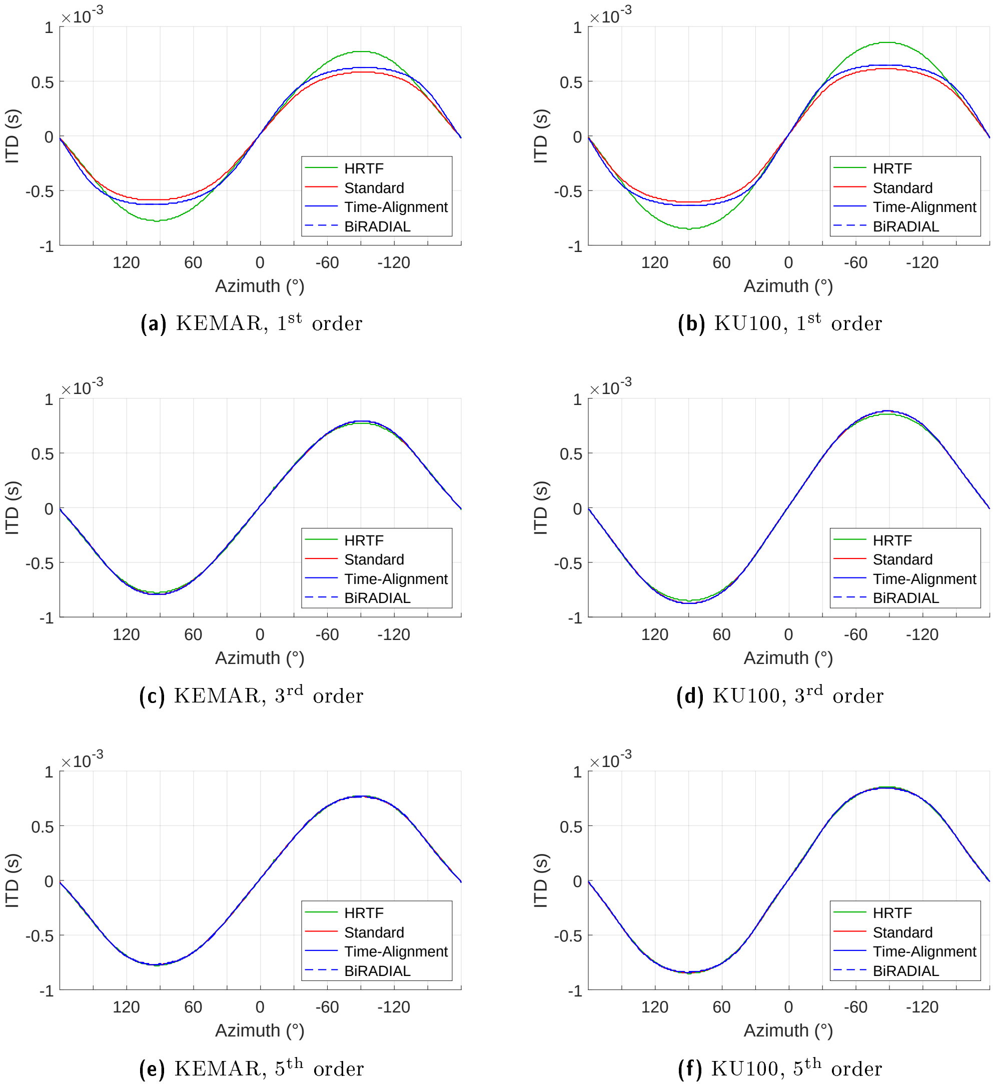
5.8 Evaluation of Time-Alignment and BiRADIAL Rendering Methods

5.8.1 Overview

During the measurement of the SADIIE database, discussed in Section 3.3, ear centered HRTFs were also captured of the KU100 and KEMAR mannequins. Examples of the set up and laser alignment are shown in Fig. 5.25. With these measurements improvements in binaural Ambisonic reproduction accuracy using Time-Alignment and BiRADIAL rendering methods may be evaluated. 3 binaural Ambisonic renderers are compared utilizing: standard dual-band (basic/max-$\boldsymbol{R}_E$), Hybrid Time-Alignment and Hybrid BiRADIAL HRTFs. Hybrid HRTFs are generated with a crossover frequency of 2100Hz. Binaurally rendered IRs are directly compared to head centred HRTFs measured about the sphere in each case. The evaluation is performed using the techniques outlined in Chapter 2 to compare ITDs, ILDs and spectral responses using the PSDM.

5.8.2 Objective Evaluation

The reproduction of ITDs is compared around the horizontal axis for 1^{st}, 3^{rd} and 5^{th} order Ambisonic reproduction over octahedral, 26 point Lebedev grid or 50 point Lebedev grid loudspeaker arrays respectively. HRTFs from both the KEMAR and KU100 mannequins are compared and the analysis is given for a 1° resolution. The results are presented in Fig. 5.26. It is observed that the responses are similar



(a) KEMAR, 1st order

(b) KU100, 1st order

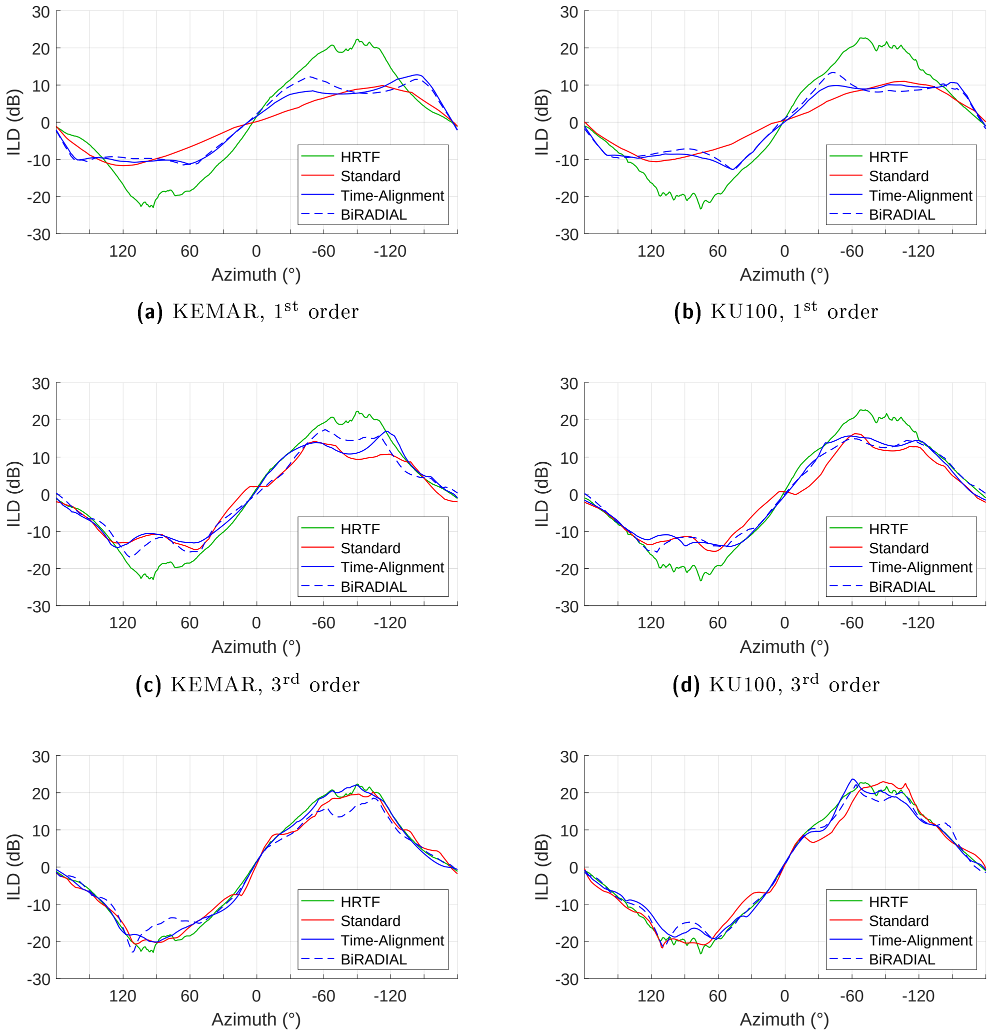
(c) KEMAR, 3rd order

(d) KU100, 3rd order

(e) KEMAR, 5th order

(f) KU100, 5th order

Figure 5.26: ITD comparison of standard, Time-Alignment and BiRADIAL binaural Ambisonic rendering methods using 1.2m KU100 and KEMAR HRTFs from the SADIIE Database. Data is taken in 1° increments about the horizontal axis. ITD is calculated from band-limited signals below 1250Hz.



(a) KEMAR, 1st order

(b) KU100, 1st order

(c) KEMAR, 3rd order

(d) KU100, 3rd order

(e) KEMAR, 5th order

(f) KU100, 5th order

Figure 5.27: ILD comparison of standard, Time-Alignment and BiRADIAL binaural Ambisonic rendering methods using 1.2m KU100 and KEMAR HRTFs from the SADIIE Database. Data is taken in 1° increments about the horizontal axis. ILD is calculated from band-limited signals above 1500Hz.

across all rendering techniques and are most accurate from 3rd order. This is due to the fact that each technique implements the same HRTF signals below 2100Hz and ITDs are calculated from band-limited signals below 1250Hz as stated in 2.4.3. The inaccuracies at 1st order are a result of the poor Ambisonic reproduction above 700Hz.

Fig. 5.27 compares ILDs around the same axis. In general the Ambisonic renderers under-reproduce the ILDs. There is greater variation in the three techniques with Time-Alignment and BiRADIAL generally approximating the HRTF response more closely. Whilst the Time-Alignment technique performs consistently across both subjects the BiRADIAL method seems to struggle slightly for the KEMAR subject at 5th order. It is not clear at this stage exactly what the cause of this result is but

Table 5.8: Average PSDM error (ERB weighted Sones) of horizontal sources rendered through single-banded standard, Time-Alignment and BiRADIAL 5$^{\mathrm{th}}$ order Ambisonic decoders

	Standard	Time-Alignment	BiRADIAL
KEMAR	1.50	0.87	0.94
KU100	1.53	0.82	0.81

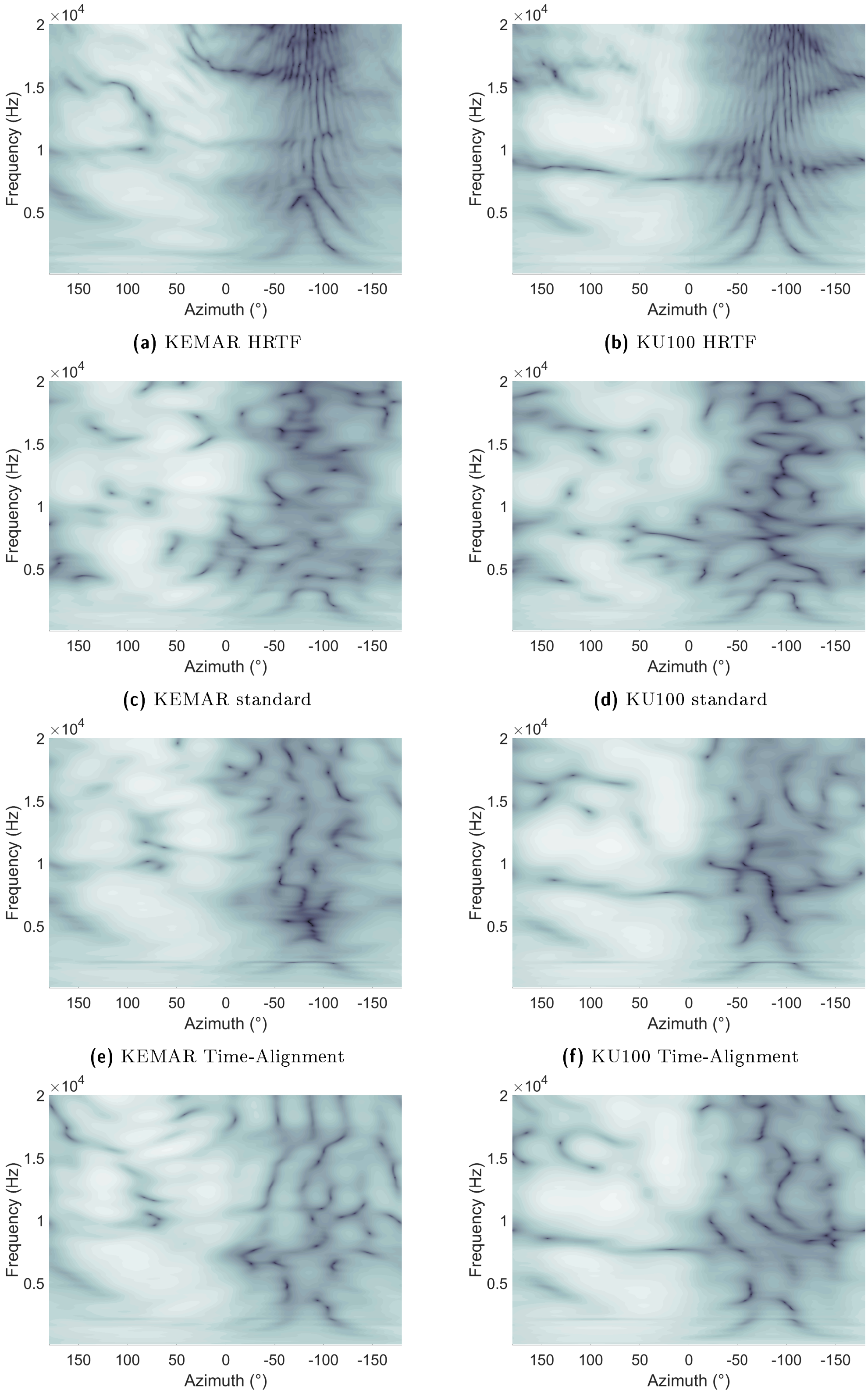
it may be explained somewhat by analysis of the spectral responses.

Fig. 5.28 shows the spectral responses of the three rendering techniques on the horizontal axis for the left ear at 5$^{\mathrm{th}}$ order. The mannequins are relatively symmetrical and so the response of the left ear is representative of both channels. The differences between the responses and the true HRTFs based on the output of the PSDM are shown in Fig. 5.29.

A table of the representative average errors of the renderers in Sones across angle and ERB weighted frequency bands in given in Table 5.8. It is quite obvious that the Time-Alignment and BiRADIAL techniques result in a more accurate spectral reproduction. This is particularly true for the frequency range 2-10kHz in which individual notches may be seen better reconstructed.

Despite still offering significant improvements over standard rendering, slightly more high frequency errors (particularly on the contralateral/right hand side) are visible in the BiRADIAL renderer, Fig. 5.29e, compared to 5.29c. These errors may have contributed to the low ILDs. One reason for this may be the due to the analysis method. When the Time-Alignment renderer is compared to the original HRTFs the exact same HRTF measurements are used in both cases. However, in order to capture the ear centered HRTFs for the BiRADIAL approach the dummy head must be repositioned and an entire new set of HRTF measurements captured. Research has shown that repeated HRTF measurements of even the same dummy head can result in dramatically different spectral responses in the high frequencies (Andreopoulou, Begault and Katz, 2015). It is therefore expected that there may be high frequency discrepancies found for the BiRADIAL renderer, but that these errors should be accounted for under repeated measurement error rather than true renderer error.

The presence of a torso must also be considered in the KEMAR measurements. The repositioning of the dummy maniquin to capture the ear centered HRTFs not only shifts the location of the ear but also affects the path lengths of any shoulder reflections. The calculation of these errors is beyond the scope of this work but may



(a) KEMAR HRTF

(b) KU100 HRTF

(c) KEMAR standard

(d) KU100 standard

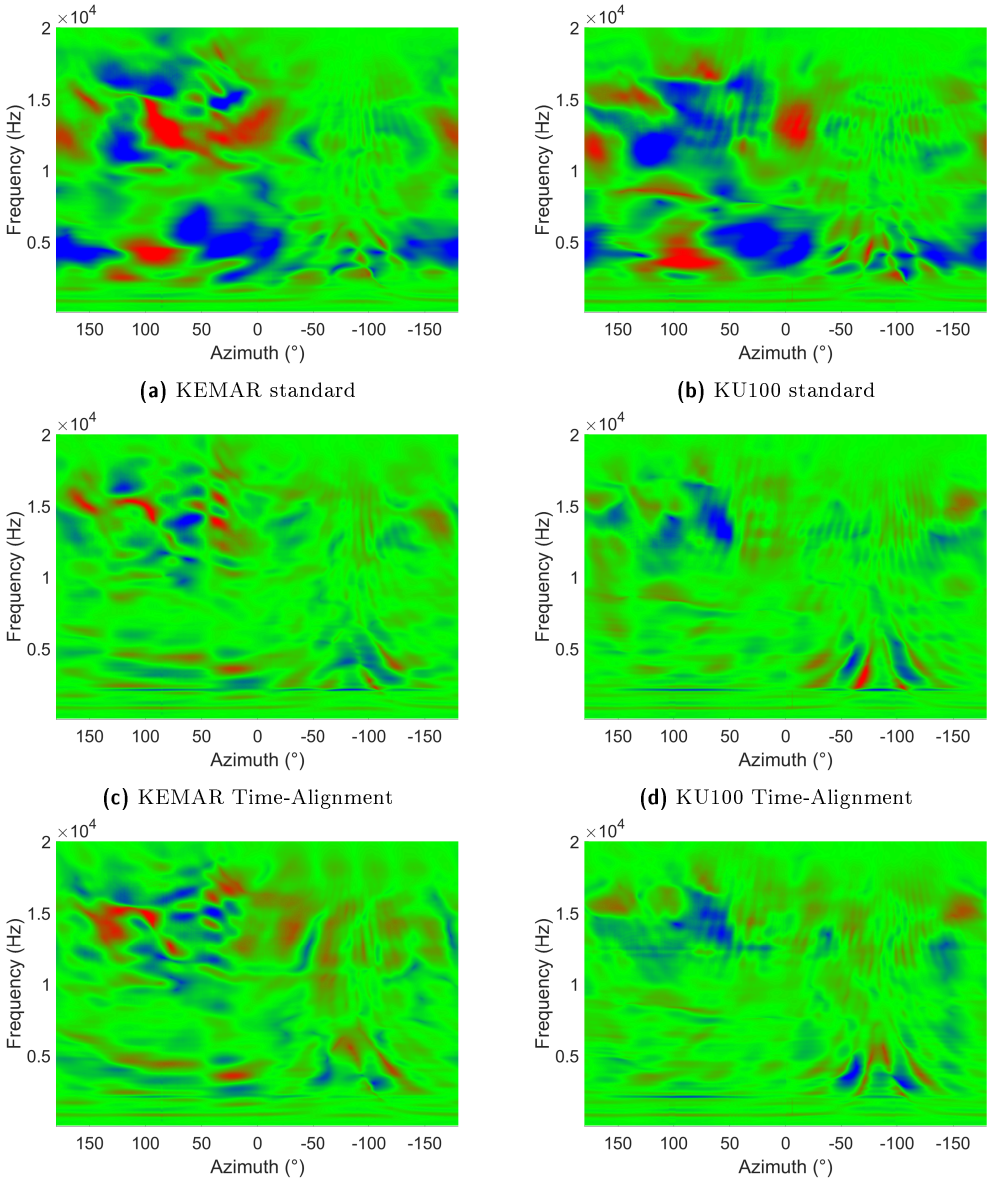
(e) KEMAR Time-Alignment

(f) KU100 Time-Alignment

(g) KEMAR BiRADIAL

(h) KU100 BiRADIAL

Figure 5.28: Spectral comparison of standard (dual-band), Time-Alignment and BiRADIAL binaural 5^{th} order Ambisonic rendering methods using 1.2m KU100 and KEMAR HRTFs from the SADIIE Database. Data is taken in 1° increments about the horizontal axis. Amplitude is shown by colour. Black: -60dB, White: +20dB.



(a) KEMAR standard

(b) KU100 standard

(c) KEMAR Time-Alignment

(d) KU100 Time-Alignment

(e) KEMAR BiRADIAL

(f) KU100 BiRADIAL

Figure 5.29: Spectral comparison of standard (dual-band), Time-Alignment and BiRADIAL binaural 5^{th} order Ambisonic rendering methods using 1.2m KU100 and KEMAR HRTFs from the SADIIE Database. Data is taken in 1° increments about the horizontal axis. The perceptual difference between the Ambisonic renders and HRTFs are shown by the blue to red colour map. Solid red indicates a positive amplitude (renderer > original HRTFs) difference of 8 Sones. Solid Blue indicates a negative amplitude difference of -8 Sones. Green indicates accurate reproduction.

explain some of the lower frequency errors around 4-5kHz that exist in the KEMAR reproductions but not the KU100 reproductions.

5.9 Summary

Two techniques for optimizing the sampling and rendering of the sweet spot in binaural Ambisonics have been discussed. The first, Time-Alignment, already exists in literature but has been re-conceptualised in this chapter as the manipulation of the position of the Ambisonic sweet spot. It is achieved by imposing a carefully calculated and unique time delay (±) onto each loudspeaker feed (practically implemented by time-shifting the HRTFs) to adjust the point in space at which the loudspeakers are temporally equidistant. Alternatively, the technique may be considered as a reduction of the energy of HRTFs within higher order spherical harmonic components by simplifying the temporal characteristics of the measurements. It is therefore possible to achieve a more accurate spectral image using the lower order spherical harmonic components only at the expense of the reproduction of temporal cues such as ITD.

Whilst the Time-Alignment technique is highly accurate for rendering sources on the loudspeaker radius, it is unable to satisfactorily render the planar wavefront curvature of a far-field source rendered using NFC filters. By manipulating the sampling position of the restored soundfield the angles at which the sources are rendered relative to the sampling position are skewed.

A second technique, BiRADIAL, has therefore been developed to solve these issues. The BiRADIAL method aims to virtualize two independent virtual loudspeaker arrays each centered about one of a person's ears. By doing this, the sampling position within either array remains central and therefore the reproduced angles of sources are preserved. The loudspeaker feeds are naturally temporally aligned by virtue of the central sampling position and NFC filters may be used to accurately render far-field sources with correct and consistent wavefront curvature.

A perceptual model is presented with an intended use for comparing the perceived similarities of binaural filters. It analyses the differences between the frequency spectra of the signals in line with the manner in which the human auditory system is sensitive to different frequencies and amplitudes. The model has been evaluated through objective measures that demonstrate the perceptually driven motivation of the model against a standard ASD calculation. The model has also been validated by identifying a significant correlation between its output and the results of a human listening test. Use of the model is therefore justified as a tool for comparing binaural filters and is hence used throughout this thesis as a form of objective analysis.

Objective evaluation has shown that both Time-Alignment and BiRADIAL rendering techniques improve upon the binaural rendering accuracy of a previously state of the art decoder. Methods in Chapter 6 will go on to show how these techniques can be used within a real-world near-field HRTF measurement rig in order to synthesise infinite high quality far-field HRTFs within a fast and convenient workflow. Alternatively, the techniques may be applied directly to current binaural Ambisonic renderers by replacement of the default HRTFs with Hybrid BiRADIAL/Time-Alignment HRTFs to improve the timbral and spatial quality of the output.

This Page Has

Intentionally Been

Left Blank

6

Feasible Near-Field HRTF measurement and Post-Processing Techniques

Chapter Overview

This chapter covers the development of MARC, a compact fast-capture HRTF measurement system. The physical construction, measurement procedure and post processing steps are laid out in full. Multiple options are discussed with regards to the synthesis of far-field HRTFs using Time-Alignment and BiRADIAL workflows. Objective analysis shows that measurements taken in MARC are of similar composition to those from the SADIIE database.

6.1 Introduction

Objective evaluations in Chapter 5 indicated the high performance levels of Hybrid HRTFs measured from the KEMAR and KU100 dummy heads. However, generic HRTFs of this kind have specific limitations. Whilst the more general benefits and drawbacks of individual measurements are still up for debate, they are the only genuine solution in seeking a truly authentic sound experience. Consider an Augmented Reality (AR) application where the aim is not to present necessarily the best sounding experience, but to match as closely as possible the virtual sources to any real sources.

As such, individual HRTF measurement has, in recent years, become subject to interest from both academic and commercial groups. However, the acquisition of high quality filters remains a challenge due to the requirement of specialist venues, large loudspeaker rigs and time consuming measurement processes. Consider the SADIIE Database; this was a 9 month project reliant on university laboratories, significant funding and 3 years of prior research, not to mention the uncomfortable seating arrangements for the participants.

Image based solutions such as that proposed by the joint business venture of Genelec and IDA Audio[22] aim to solve the problem through simulation. Common image capture devices (such as mobile phones) are used to generate a 3D model of a person's ears. Computer simulations are then run to compute the approximate HRTFs of the individual. Although convenient, the method relies heavily on the accuracy of both the 3D scan, and the simulation. For example, approximations must be made with regards to the acoustic absorbancy of the person's head/skin. Slight errors in calibration could lead to the mis-orientation of the person's pinnae.

A solution is proposed here in the form of a compact, fast HRTF capture booth. It is referred to as MARC. 50 unique HRTF measurements may be taken in under 7 seconds maintaining a high (>50dB) SNR. The device is designed to be housed within a typically reverberant research lab and still achieve pseudo-anechoic results. The primary concern with this type of rig is the proximity of the loudspeakers to the head and therefore the near-field nature of the measurements. However, by employing a binaural Ambisonic workflow the radial distances of the loudspeakers

[22] https://auralid.genelec.com/

may be accounted for using post-processing techniques and NFC filters. Following this technique, it is then possible to synthesise any number of far-field HRTFs.

The work presented in this chapter has been published by Armstrong and Kearney, 2020; Young et al., 2019.

6.2 Considering Near-Field Measurements

6.2.1 Overview

Binaural renderers too often rely on a plane wave assumption of HRTFs, that is that the measurements represent an approximately planar wave approaching the head. In this case the spectral response of the filter is no longer distance dependant (Brungart and Rabinowitz, 1999) and is said to be measured in the far-field. Previously, this has been considered to be anywhere beyond a 1-1.5m radius (Brungart and Rabinowitz, 1999; Spagnol, 2015). The majority of available databases contain HRTFs measured at a fixed radius within this region, for example (Armstrong et al., 2018a; Jin et al., 2014; Algazi et al., 2001; Gupta et al., 2010; Bernschütz, 2013), see Table 3.1.

Closer to the head is the near-field, in which the natural wavefront curvature of the source is highly relevant, and spectral alterations due to the proximity of the shoulders and pinnae are seen (Otani, Hirahara and Ise, 2009; Yu, Xie and Rao, 2010). Increased ILDs are also produced due to the acoustic shadowing of the head (Brungart and Rabinowitz, 1999). The acoustic parallax effect should also be considered. A source which lies centrally in front of the listener in fact lies to the left with respect to the right ear, and to the right with respect to the left ear. The shift in perceived angle results in a lateral shift of HRTF features with greater effect at smaller radii (Brungart, 1999). The perceptual impact of such shifts has not yet been fully explored but must be considered within a near-field HRTF measurement system.

At a radius of 1m the maximum angle subtended from a source by the centre of an average head and the ear is approximately 4°. However, at 0.5m this angle is closer to 15°. This shift is demonstrated in Fig. 6.1. The perceptual impact of the angular discrepancy is of interest above the minimum audible angle of a sound source, commonly cited as 1° (Mills, 1958).

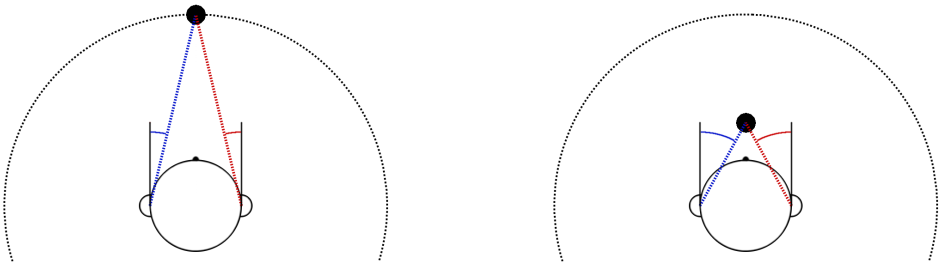


Figure 6.1: The parallax angles of far-field sources (left) are smaller than for the near-field (right).

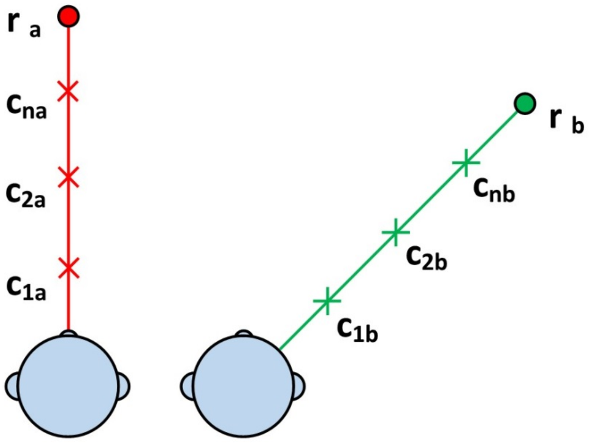


Figure 6.2: Radially-spaced HRTFs: at each angle **(a, b, ...)** a series of HRTFs $(\mathbf{c}_{1,\,2,\,\ldots\,n})$ are each compared to a reference of the same angle, but a 10m radius **(r)**.

6.2.2 Simulations and Comparisons

Analysis is undertaken into the perceptual differences of HRTFs to assess the requirement of specialist near-field considerations in measuring HRTFs for binaural Ambisonic reproduction. It is important to identify any meaningful differences between the HRTFs that may be available within a given dataset and those that are required for a given application. Two different variations are considered. Firstly, HRTFs of decreasing radii are compared to a far-field reference. These comparisons are referred to as radially spaced HRTFs and are demonstrated in Fig. 6.2. Secondly, HRTFs of equal radii, but centred about either a person's head or ears are compared. These comparisons are referred to as laterally spaced HRTFs and are demonstrated in Fig. 6.3. In each case the left channel of the head centred HRTF is compared to the left channel of the left ear centred HRTF. This is repeated for the right ear.

In order to obtain the number of HRTFs required for a detailed analysis BEM simulations were calculated for a 3D mesh of the KEMAR manikin previously validated up to 16kHz (Young, Tew and Kearney, 2016; Young, Kearney and Tew, 2018a;

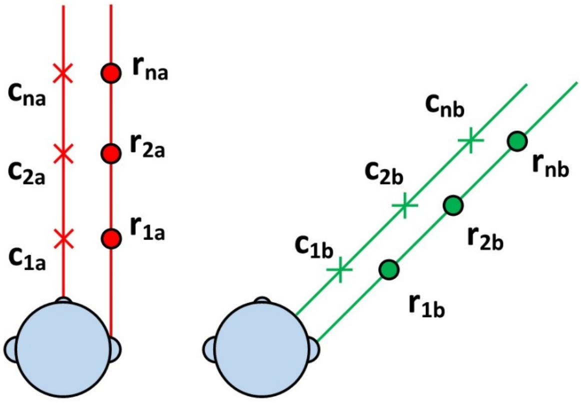


Figure 6.3: Laterally spaced HRTFs: at each angle (**a, b, ...**) a series of head centred HRTFs **c**$_{1,\ 2,\ ...\ n}$ are compared to a series of ear centred HRTFs (**r**$_{1,\ 2,\ ...\ n}$) such that the left channel of the head centred HRTF is compared to the left channel or the left ear centred HRTF and vice versa for the right. The diagram shows the comparison for the right ear only

Table 6.1: Radius distances used in the head-centred simulation, totalling 62012 simulation points. The ear-centred simulations use the same resolutions starting at 0.4m, therefore totalling 60828 points.

	Limits	Resolution
Radius (r)	0.32–3m	0.01m
	3.02–5m	0.02m
	5.1–10m	0.1m

Young, Kearney and Tew, 2018b). 148 angles were simulated at 419 (411 for ear-centred simulations) radial distances described in Table 6.1. The minimum radial distances simulated were 0.32m (head centred) and 0.4m (ear centred). These were limited by invalid source locations existing within the mesh. The angles simulated include a range of typical Ambisonic configurations. Such configurations are evenly distributed over the sphere and therefore representative of fair 3D sampling. A higher resolution distribution was also simulated on the horizontal axis.

The PSDM, introduced in Section 5.7, was used to compare each pair of HRTFs. The minimal radial distance at which the difference between the HRTFs fell within 1 Just Noticeable Difference (JND) was calculated for each angle. The JND of spectral amplitude is commonly quoted as 1dB across a range of frequencies and amplitudes (Yost, 2000; Mills, 1960). The PSDM normalises its input to 75dBSPL and gives an output in Sones. A difference of ±1dB at 75dBSPL is therefore considered for a 1kHz signal to define a JND in Sones. This is calculated by

$$S_{76} - S_{75} = 0.81 \tag{6.1}$$

$$S_{74} - S_{75} = -0.76 \tag{6.2}$$

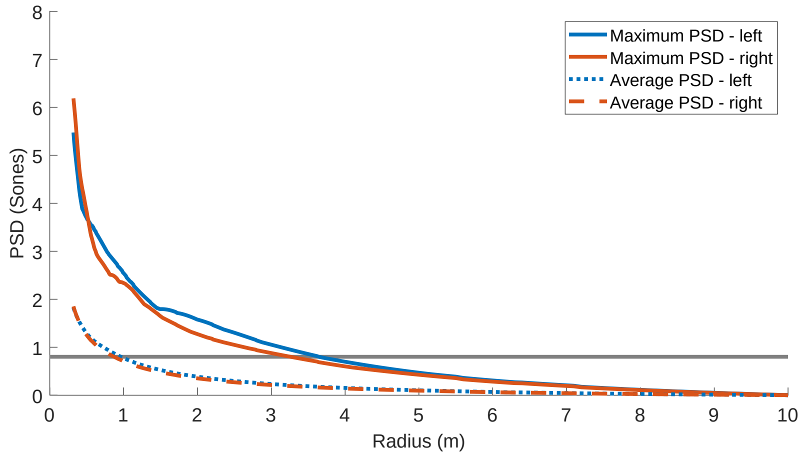


Figure 6.4: Average and maximum PSD between radially-spaced HRTFs and the 10m reference for an incident angle of ($\varphi = 0°, \vartheta = 0°$). The horizontal line represents the JND of 0.8 Sones.

where

$$S_p = 2^{(\frac{p-40}{10})} \tag{6.3}$$

and

- S_p is the value in Sones,
- p is the value in phones (equal to the value in dBSPL for a 1kHz tone).

For convenience, an approximate value was defined (rounded to 1 decimal place):

$$1 \text{ JND} = 0.8 \text{ Sones} \tag{6.4}$$

6.2.3 Results

Two cases are defined in which the output of the PSDM falls within 1 JND. The first is the point at which the *average* difference across all frequency bands falls within 1 JND, D_{avg}. The second is the point at which *all* differences across all frequency bands fall within 1 JND, D_{max}. Results are presented for both cases. The first is indicative of the point at which the measurements may begin to sound similar, the later is indicative of the point at which the measurements should sound identical.

A decreasing PSDM output for radially spaced HRTFs (Fig. 6.2) with an increasing radius for the angle ($\varphi = 0°, \vartheta = 0°$) is demonstrated in Fig. 6.4. D_{avg} is approximately 0.92m and D_{max} is approximately 3.47m. Both vary slightly between the left/right channel. The general shape of the graph is typical in every case.

It is required that the comparisons of both left and right channels individually fall within 1 JND for the comparison of the HRTFs to fall within 1 JND. In each case the larger of the two PSDM values from either the left of right channel is prioritised and referred to as the binaural PSDM. The radial distances at which this single value falls within 1 JND may then be plotted over 2 dimensions to represent the values over the sphere. The results for radially spaced HRTFs are presented in Fig. 6.5. The maximum value of D_{avg} is 2.7m and occurs at ($\varphi = -26°, \vartheta = -15.5°$). The maximum value of D_{max} is 9.3m and occurs at ($\varphi = 72°, \vartheta = 26.6°$). For reference D_{max} is also plotted for the left channel only in Fig. 6.5c. Note the smaller values on the ipsilateral side indicate greater differences in the radially spaced HRTFs in general.

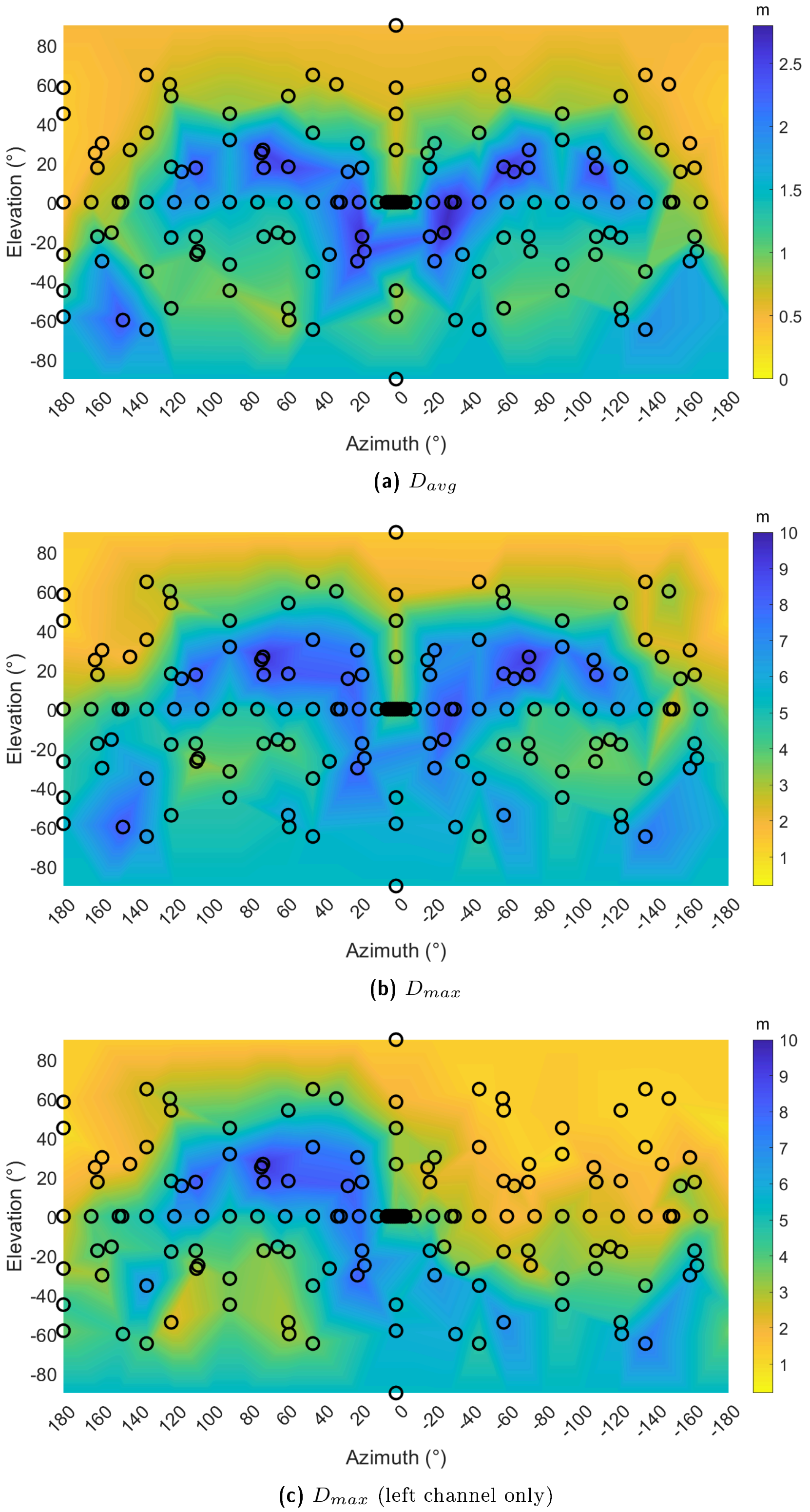
The results for laterally spaced HRTFs are presented in Fig. 6.6. The maximum value of D_{avg} is 1.53m and occurs at ($\varphi = -64°, \vartheta = 15.5°$). The white space in Fig. 6.6b indicates that D_{max} was not reached within the simulated 10m radius.

6.2.4 Analysis

Generally, D_{max} exceeds D_{avg} in all cases which is to be expected due to increased high frequency errors. Fig. 6.5a supports the findings of Otani, Hirahara and Ise, 2009 which state that radially spaced spherical and planar HRTFs are the same by 3m, however, Fig. 6.5b identifies some angles with abnormally high maximum PSD values across the spectrum e.g. $\varphi = \pm 45°, \vartheta = 35.3°$. The results are explained by a shift in high frequency features with an increase in radial distance, as shown in Fig. 6.7. It is suspected that at specific angles particular pinnae cues are excited in such a way that differences are still perceivable in the high frequencies up to distances of 9m.

Fig. 6.5c demonstrates the dependency of channel based PSD on whether a source is considered from the contralateral or ipsilateral side of the head. It is found that the angles of greatest error are those that most directly face the front of the pinnae $\pm 20° < \varphi < 110°$, $-20° < \vartheta < 40°$ (or are shadowed by the torso $\vartheta = -60°$). However, it should be noted that for general binaural reproduction the worst case scenario (be that the left or right ear) should always be considered.

In Fig. 6.6a it is seen that the PSD increases for laterally spaced HRTFs as they approach the angle of the interaural axis ($\vartheta = \pm 90°$) despite there being little change in the angle of incidence to either ear in these regions. This may at least partially be explained by significant ILD changes in the near-field. Magnitude offsets between the

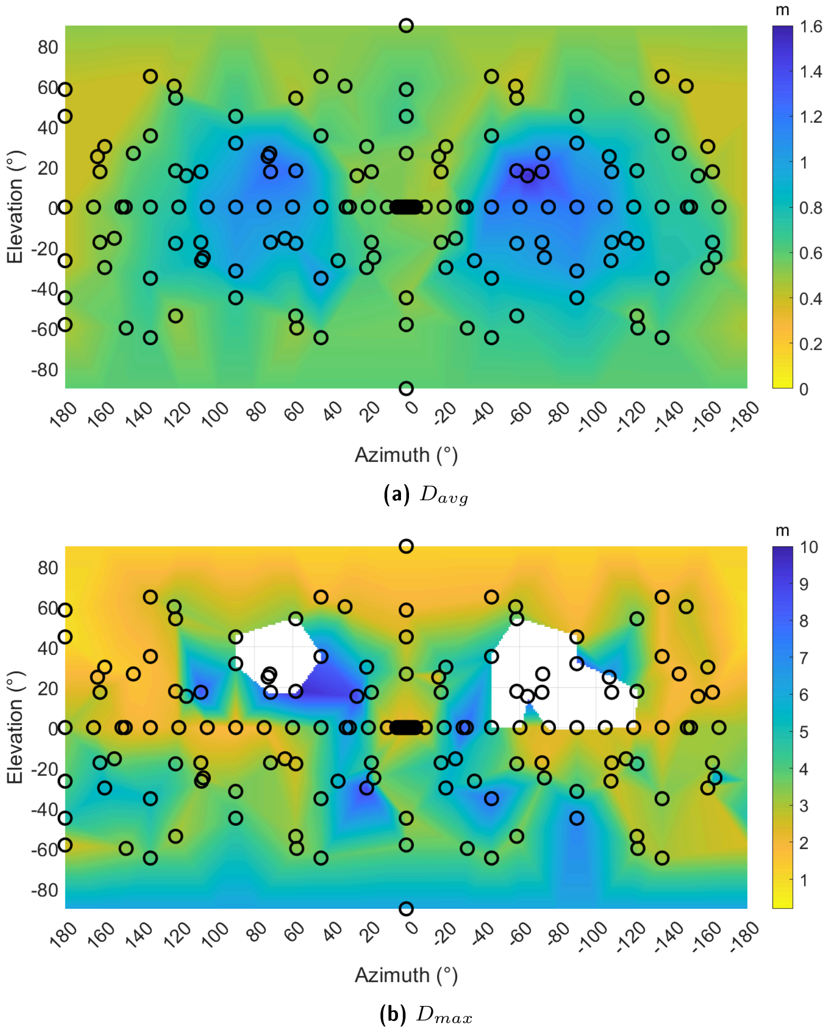


(a) D_{avg}

(b) D_{max}

(c) D_{max} (left channel only)

Figure 6.5: Binaural PSD between radially-spaced HRTFs. Values at simulated angles are identified by circular markers; results are interpolated between scattered data to aid in visualisation.



(a) D_{avg}

(b) D_{max}

Figure 6.6: Binaural PSD between laterally-spaced HRTFs. Values at simulated angles are identified by circular markers; results are interpolated between scattered data to aid in visualisation. White space indicates that D_{max} was not reached within the simulated distance.

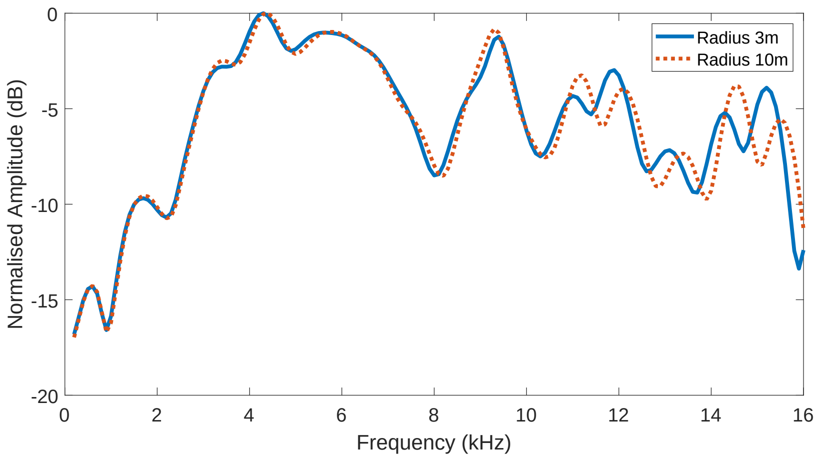


Figure 6.7: Left channel of HRTF for $\vartheta = \pm 45°$, $\varphi = 35.3°$, simulated at 3m and 10m, normalised such the the maximum value of each HRTF equals 0dB. Note the misalignment of high frequency notches resulting in a high maximum PSD.

HRTFs due to the lateral shifts are highlighted by the averaging comparison technique. However, considerable feature shifts are also identified within these HRTFs at close proximity to the head. Fig. 6.8 plots the frequency response of 3 HRTFs simulated at $\vartheta = 90°$, $\varphi = 0°$ and at radial distances of 0.4m, 0.7m and 1.5m. These results could suggest a secondary parallax effect with respect to the visible pinnae folds of the ear as discussed by Spagnol, 2015.

Two primary regions are identified within which there are perceptually significant changes to the frequency spectra of both radially and laterally spaced HRTFs up to distances exceeding 1-1.5m. The first region is near the horizontal axis within the frontal hemisphere and extending out to either side of the head. The second is at elevations below approximately $-60°$. Within VR applications these regions are often populated by important sources. It is therefore critical to consider the perceptual impact of radial and lateral HRTF distortion in these regions. Although it is shown that average HRTF variation reaches perceptual limits within approximately 2-3m from the head, differences in individual spectral features prevail up to and beyond 10m. These findings could therefore remain critical for narrow band sources especially.

6.3 Development of MARC

6.3.1 Overview

It is shown throughout Section 6.2 that HRTFs vary significantly with radius within 2-3m of a person's head. However, practical limitations generally demand that measurements of any real human must be taken within this range. Considerations must therefore be taken in exactly how these measurements are made and what processing can be applied to the filters to account for their radial distance. To that end a new measurement tool has been realised that prioritises convenience and feasibility over maximising the radial distance of the measurements.

6.3.2 Design

MARC is a miniaturised HRTF measurement rig that depends on both hardware and software to measure and account for the proximity of the HRTFs to the head. The original concept design is shown in Fig. 6.9. The equipment was designed in an attempt to meet the following criteria. To be:

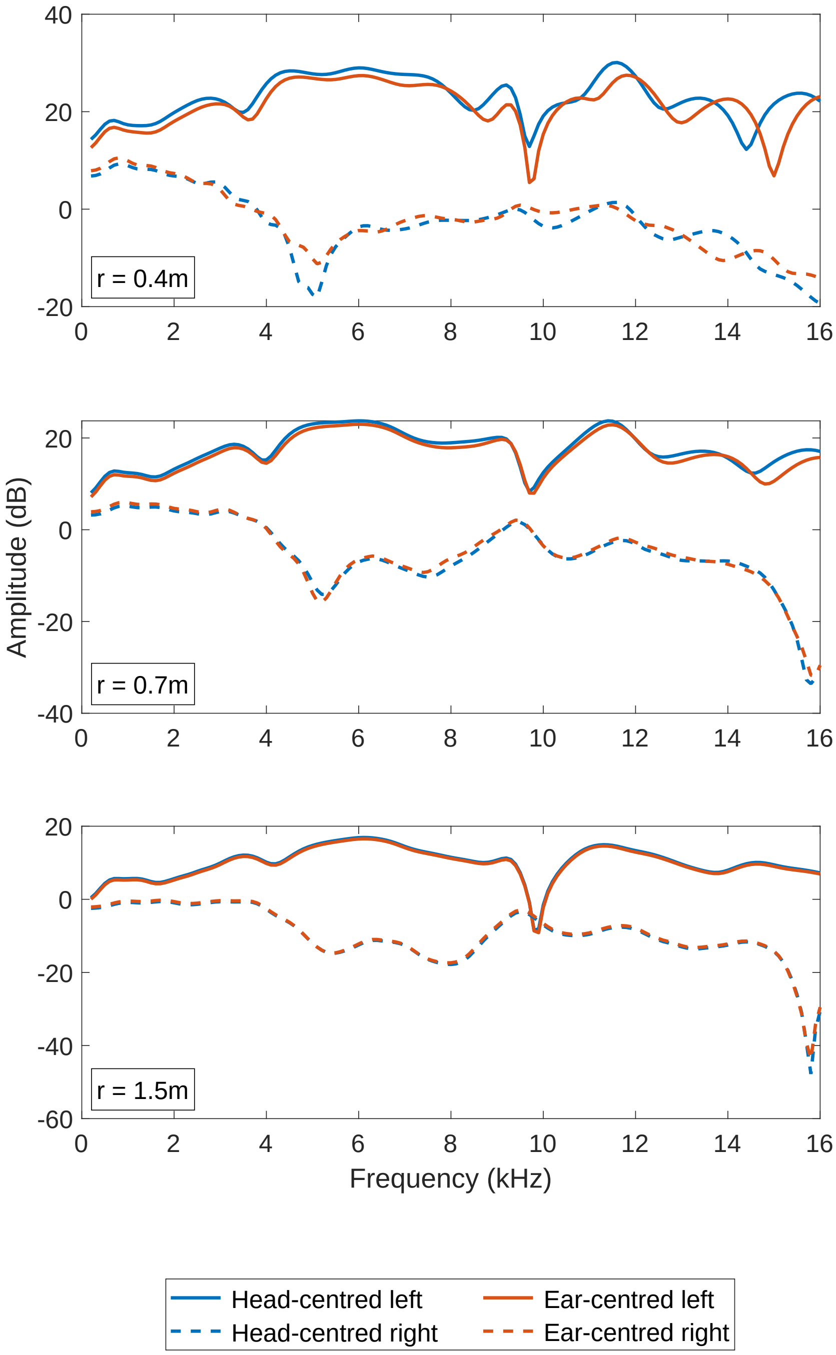


Figure 6.8: Laterally spaced HRTFs simulated from an incident angle of $\vartheta = 90°$, $\varphi = 0°$ at the radial distances of: 0.4m (top), 0.7m (middle) and 1.5m (bottom). The settling of spectral features with an increase in distance can be seen.

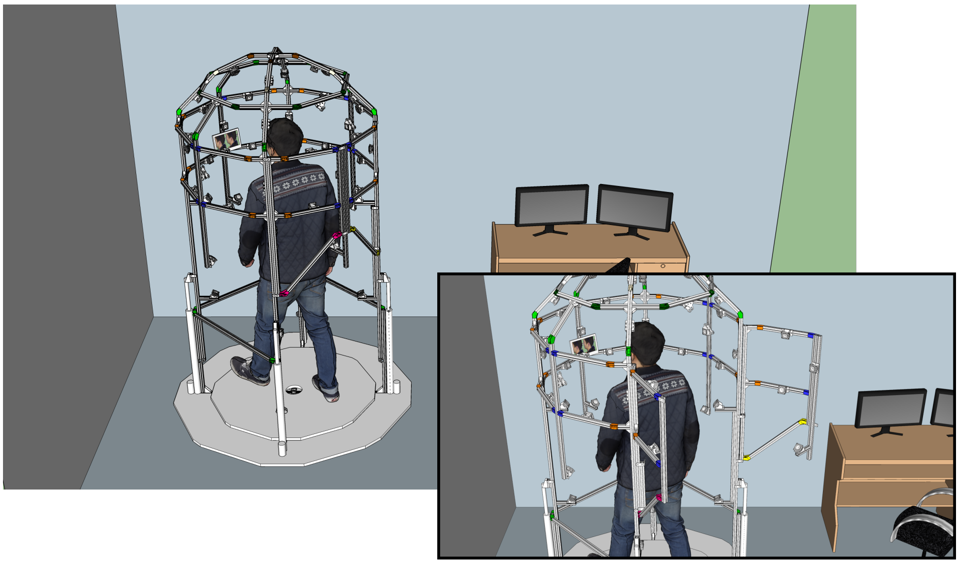


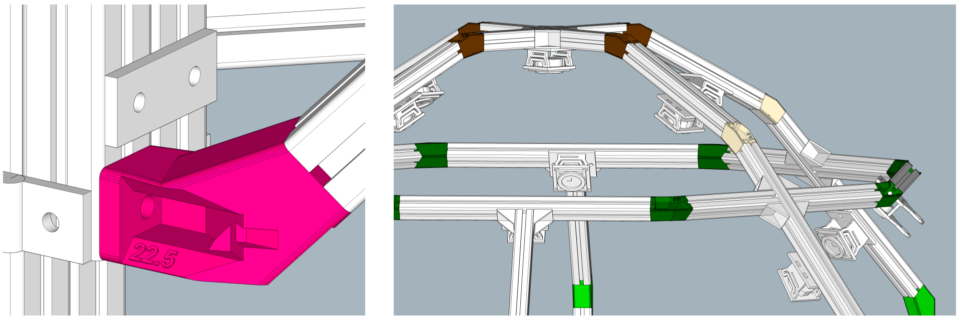
Figure 6.9: Original SketchUp[23] concept design of MARC, a walk-in near-field HRTF measurement rig.

- free standing,
- of minimum size,
- portable,
- optimal for rapid HRTF measurement,
- suitable for multiple people,
- capable of capturing HRTFs at many locations.

The hardware is composed of a rigid frame built from 20mm aluminium extrusion[24] upon which individual loudspeakers are mounted. A walk-in/walk-out approach is taken. Two doors on the front of the frame open allowing a person to walk inside. They are closed behind them to provide a full 360° coverage of loudspeakers. The entire rig is built onto a free-standing wooden base with a small lip around the edge to mount wiring. Analogue and digital input/output leads are then run from the wooden base to the relevant control systems, as discussed within the remainder of this section.

A 50 point Lebedev grid distribution of loudspeakers was chosen. The configuration is an even distribution of points over the sphere and is particularly optimal for Ambisonics up to 5th order (Lecomte et al., 2015; Lecomte et al., 2016). Further, the distribution includes convenient measurement locations ie. directly in front, to the side, above and below. It is also one of the higher Ambisonic order configurations

[24] valuframe.co.uk/



(a) Multi-angle bracket

(b) Collection of brackets

Figure 6.10: Construction of the frame using 3D printed angle brackets to sit perpendicular the loudspeaker angles.

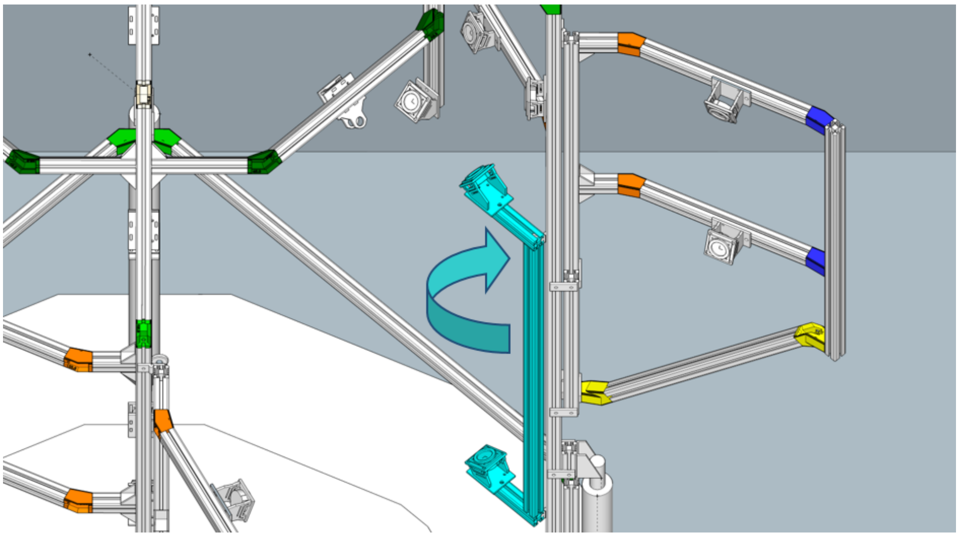
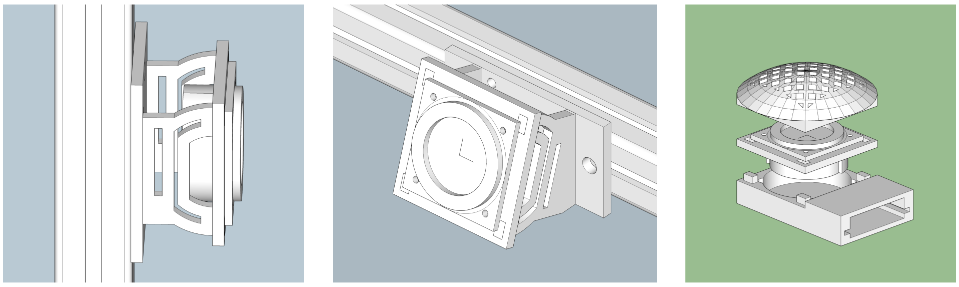


Figure 6.11: Movement of the internal mechanism that allows the loudspeakers next to the door to swing out of the way to prevent accidental contact.

measured as part of the SADIIE database and is therefore ideal for comparative studies.

Custom designed 3D printed angle brackets are used to connect and shape the aluminium extrusion. Where possible, the frame is designed to sit behind each loudspeaker and align perpendicularly with the desired angle of the speaker, see Fig. 6.10. Hinges and a simple acrylic latch allow the doors to open and lock in to place. Loudspeakers mounted immediately next to the door are also fixed to an internal mobile mechanism that allows them to swing inwards, shown in Fig. 6.11, minimising the risk of them getting bumped as people walk in/out of the rig.

The loudspeakers are 1.3" 2W RMS 8Ω Visaton BF32-8OHM full range mini speakers. They have a quoted frequency response of 150Hz-20kHz and a sensitivity (@ 1W/1m) of 78dBSPL. They are installed within laser cut acrylic mounting blocks that are designed to be able to angle the loudspeakers away from the frame where



(a) Loudspeaker mounter perpendicular to frame

(b) Loudspeaker mounted at angle to frame

(c) −90° elevation loudspeaker in 3D printed casing with protective grill

Figure 6.12: Laser cut acrylic speaker mounts designed to give minimum distance between the loudspeaker diaphragm and back screw plate.

necessary with minimum distance between the loudspeaker diaphragm and back screw plate which attaches them to the rig, see Fig. 6.12.

Ensuring a flush placement of the loudspeakers reduces obvious and distinct reflections off the frame from rearward-emitted sound waves. A special casing, shown in Fig. 6.12c, was 3D printed for the loudspeaker located at −90° elevation and included a grill to prevent it being accidentally stood on. The speaker is installed within a hollowed out slot in the wooden base.

A mono 3.1W unity gain differential amplifier (built around the Texas Instruments TPA6211A1DGN chip[25]), described in Fig. 6.13, is placed within the back of each speaker block. Its small size allows it to be printed onto a 36 × 26mm PCB with room to spare and is powered by a 5V DC connection. The amplifier accepts a balanced audio input facilitating low noise data transfer.

The signal and power lines are fed to each loudspeaker along the contours of the frame. Care was taken to evenly distribute the 50+ cables between the 4 legs so as not to unevenly load the array. The power supply lines were wired in parallel to reduce the number of cables entering the rig. Branches and sub-branches were tapped from 2 primary power lines using T-Tap quick slice insulated wire terminals. Remaining connections were spliced together using heat shrink butt connectors.

The height of the array is adjustable in order to center the loudspeakers about the person's head/ears. A moving internal frame is lifted up and down within a fixed basket section that sits on the floor and is atached to the frame by 4 pairs of roller wheels. The frame is lifted from the ground by 4 equally spread 50mm

[25] ti.com/store/ti/en/p/product/?p=TPA6211A1DGN

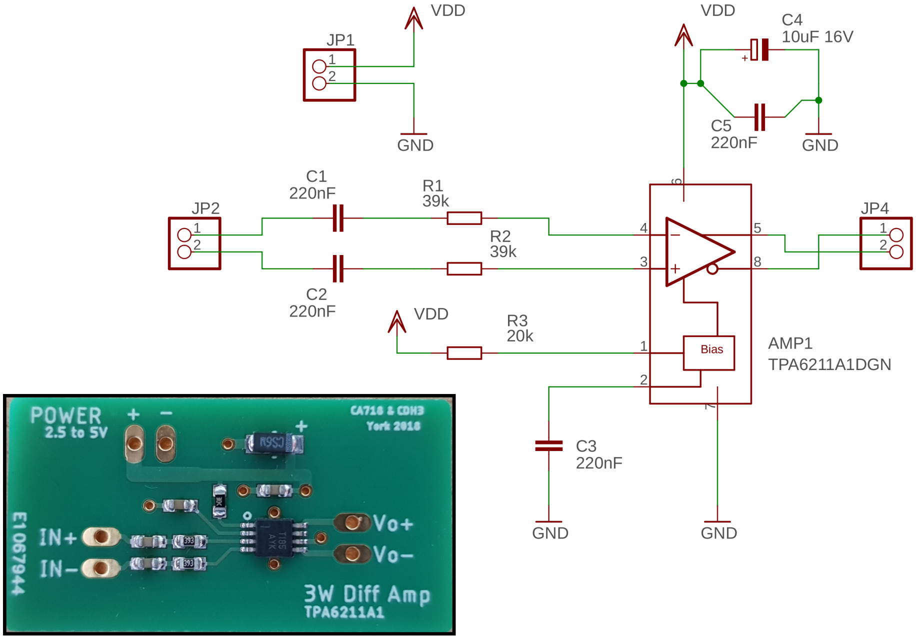


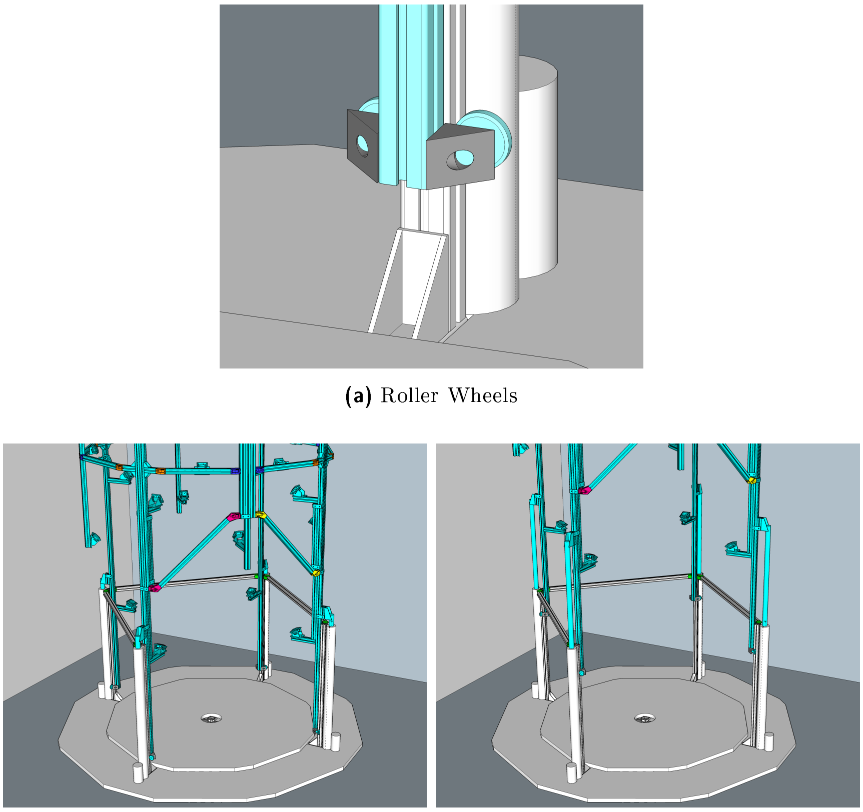
Figure 6.13: Circuit Diagram of the 3.1W mono unity gain differential amplifier built around the TPA6211A1DGN chip designed by the technical support staff of the University of York Electronic Engineering department.

linear actuators that are controlled in parallel by an operator and that are capable of millimetre precision. The repositioning of the internal frame is shown in Fig. 6.14.

Aligning a person to the loudspeakers by sitting them on an adjustable stool or chair (as was previously done, e.g. the SADIIE database) is a painstaking process that is prone to error. Often, the height adjustment mechanism of the seat is not particularly robust or precise and further the person's posture (whether or not they slouch or round their back) has a large impact of their ear height. It is therefore likely that their position will change throughout the measurement process. Other considerations such as the acoustic impact of the stool or potential thigh reflections must also be taken into account.

Variable-height flooring was also considered, however, this was an impractical solution. It would have been harder to adjust with the client in situ, would have been greatly effected by their weight and would have required the loudspeaker frame itself to be permanently fixed at a 'full height'. This would have made working on/inspection of the array difficult.

A self alignment system consisting of low powered lasers, cameras and a screen was developed. A cross is projected toward the centre of the array from either side (±

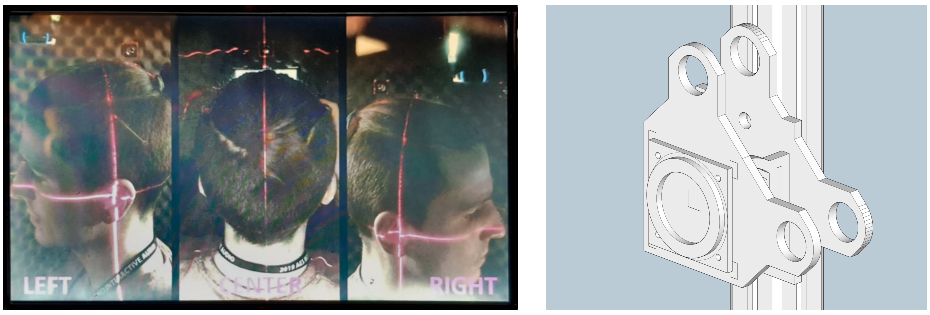


(a) Roller Wheels

(b) Minimum Height

(c) Maximum Height

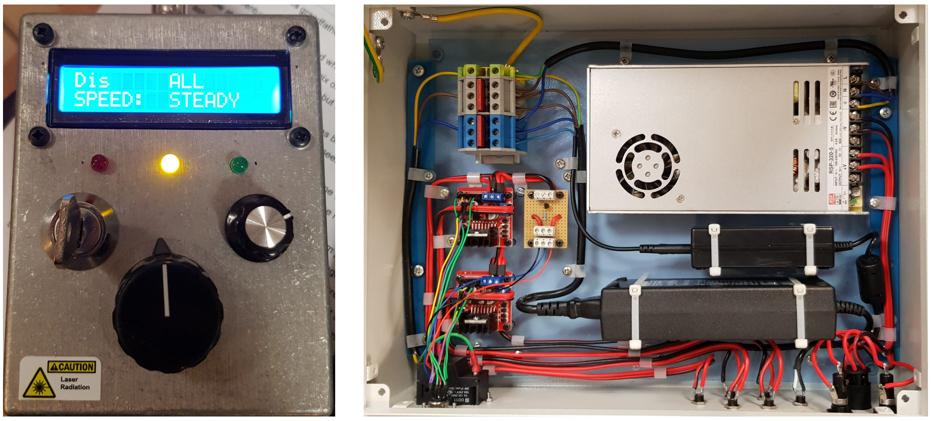
Figure 6.14: Adjustable height operation of MARC via 4 linear actuators mounted between the floor and an internal frame (highlighted blue). Attachment of the internal frame to the fixed basket by pairs of roller wheels is shown in Fig. 6.14a.



(a) Self alignment using lasers

(b) Extended speaker mounting plate including laser mounting holes

Figure 6.15: Laser self-alignment system developed for MARC.



(a) Arduino based control interface

(b) Multi-purpose power supply

Figure 6.16: Control interfaces and power supplies for MARC

90°) in addition to a vertical line projected from the rear (180°). The lines are indicative of the saggital, frontal and horizontal planes whose intersection defined the centre of the array. Images from 3 webcams are shown on a small TV screen mounted at the front of the array facing inward. They show the back of the person's head and each of their ears. The system allows a person to align themselves in the center of the array by moving until they are able to see a cross focused on either ear and a single line down the back of their head as in Fig. 6.15a. The laser modules are installed within extended loudspeaker mounting plates shown in Fig. 6.15b. Due to the placement of the loudspeakers, a pair of line modules (one vertical and one horizontal) are required to construct a cross pattern originating from the same location as the laterally placed speakers.

The linear actuators and power to the laser alignment system are operated and controlled via an Arduino based custom control interface, shown in Fig. 6.16a. The laser modules are wired to an isolated 5V supply which is gated by a simple transistor switch. The lasers must be switched on using a key to prevent accidental exposure.

A potentiometer dial allows an operator to raise or lower the linear actuators with variable speed. A rotary encoder allows selection between control of the actuators as a set, or individually for calibration. To operate the actuators, digital signals are sent from the arduino board to a pair of 2-Channel STMicroelectronics L298N PWM H-Bridge motor control units[26] which in turn feed analogue signals to the actuators.

Individual power supplies for the control interface, loudspeakers, actuators, TV screen and lasers are housed within a single mains powered supply unit, shown in Fig. 6.16b. The unit has a custom 12-pin DIN connector into which the control interface plugs via an extended cable.

Finally, the frame was treated with acoustic foam to minimise internal reflections. Gaps between struts are also filled to prevent an abundance of room reflections. Particular attention was paid to treating the solid wooden base. Small holes were cut in the foam for the loudspeakers to poke though, ensuring that no part of the foam directly shadowed any diaphragm. Fig. 6.17 presents the final system.

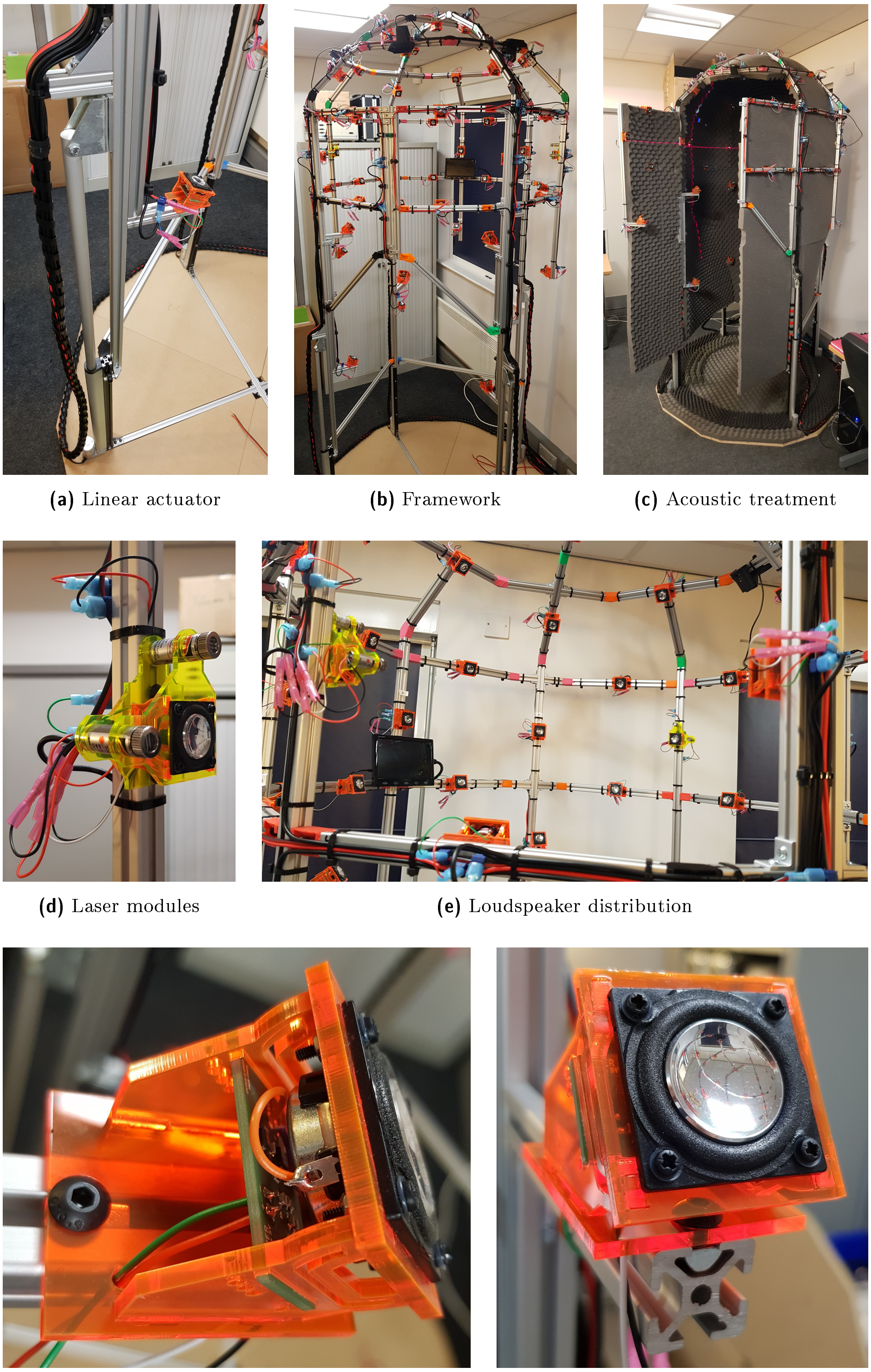
6.4 Control Software

6.4.1 Overview

The entire measurement process from calibration to post-processing is handled by a dedicated application designed in the MATLAB *appdesigner* environment on PC. Key features include:

- **Audio Interface:** Selection, Channel Mapping, Sampling Frequency
- **Output Sweep Parameters:** Number of measurements, Initial/End delay, Overlap/Interleave delays, Length of sweep, Frequency range, Amplitude (per sweep), EQ Filter (per sweep)
- **Calculation of compensation filters:** Smoothing Parameters, Frequency Range, Windowing
- **Post-Processing Parameters:** Free-Field Equalisation, Low Frequency Extension, Diffuse-Field Equalisation

[26] st.com/en/motor-drivers/l298.html

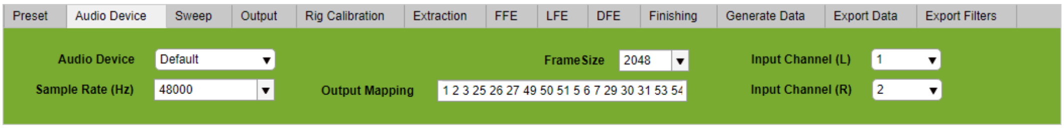


(a) Linear actuator **(b)** Framework **(c)** Acoustic treatment

(d) Laser modules **(e)** Loudspeaker distribution

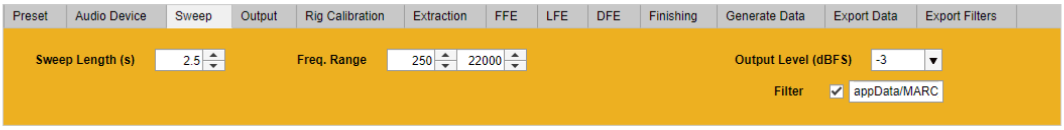
(f) Amplifier board installation **(g)** Speaker mounting block

Figure 6.17: Finished prototype of MARC showing some of the key features.



Audio Device: Default
Sample Rate (Hz): 48000
Frame Size: 2048
Output Mapping: 1 2 3 25 26 27 49 50 51 5 6 7 29 30 31 53 54
Input Channel (L): 1
Input Channel (R): 2

Figure 6.18: MARC Workflow: Audio Device settings.



Sweep Length (s): 2.5
Freq. Range: 250 22000
Output Level (dBFS): -3
Filter: ☑ appData/MARC

Figure 6.19: MARC Workflow: Sweep settings.

- **Output Format:** RAW Data, Processed Data, Distance simulation, spherical harmonic encoding of HRTFs

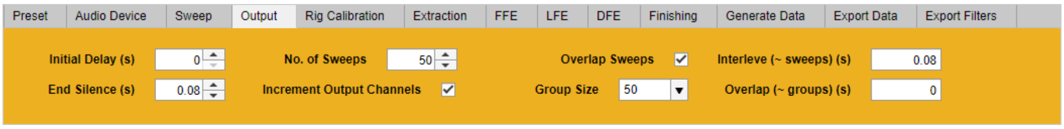
Whilst the software was a bespoke design for MARC, it is applicable generally to all HRTF measurement facilities. Section 6.4.2 details a complete and linear workflow for HRTF Measurement.

6.4.2 MARC Workflow

Setup Audio Device: With reference to Fig. 6.18, The *Audio Device* and desired *Sampling Rate* must first be selected from the drop down menus. The *Frame Size* relates to the input/output buffer of Matlab. This should be increased if there are audio glitches in the output. The *Output Mapping* option allows the order of the playback output channels to be remapped to the physical outputs of the audio device. The *Input Channels* specify which inputs on the audio device the microphones are plugged into.

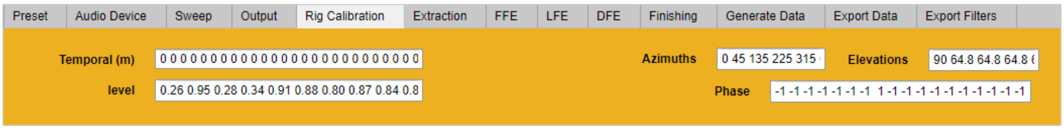
Define the Sweep: With reference to Fig. 6.19, The exponential sweep for the IR measurements must then be defined. The application will generate a logarithmic sine sweep based on the programmed values. *Sweep Length* should be specified in seconds and the *Frequency Range* in Hz (min to max). The parameters should depend on the desired SNR, time under test, loudspeaker quality, room effects etc. These may only be confirmed once all construction/soundproofing has been completed and the apparatus is in its final location.

The *Output Level* is defined in the digital domain in dBFS. The actual output volume will depend on the audio device, the amplifiers and the loudspeaker sensitivities. It is advised to first try this value very low and work up towards a sensible volume



Initial Delay (s) 0
End Silence (s) 0.08
No. of Sweeps 50
Increment Output Channels ☑
Overlap Sweeps ☑
Group Size 50
Interleve (~ sweeps) (s) 0.08
Overlap (~ groups) (s) 0

Figure 6.20: MARC Workflow: Output settings.



Temporal (m) 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
level 0.26 0.95 0.28 0.34 0.91 0.88 0.80 0.87 0.84 0.8
Azimuths 0 45 135 225 315
Elevations 90 64.8 64.8 64.8
Phase -1 -1 -1 -1 -1 -1 -1 1 -1 -1 -1 -1 -1 -1 -1 -1 -1 -1

Figure 6.21: MARC Workflow: Rig Calibration settings.

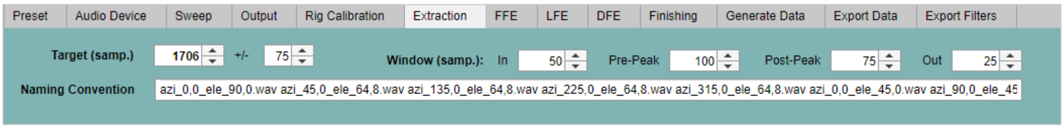
depending on the number of overlapping sweeps. The *Filter* option allows an individual frequency equalisation FIR filter to be applied to each sweep. This should be left unchecked at this stage (no EQ).

Define Output: With reference to Fig. 6.20, the output settings define how multiple sweeps are output during a single measurement phase. The *Initial Delay* specifies a length of silence at the start of the measurement. This is useful if, for example, the operator must start the measurement process and then exit the space before any sweeps are played. The *End Silence* specifies a length of silence at the end of the sweeps to ensure a complete reverberation tail is captured in the recording. This value should be of similar order/greater than the RT60 time of the measurement space. The *Number of Sweeps* to be output should also be selected.

The *Increment Output Channels* option is what defines whether the sweeps are output to a single output channel, or each to their own individual channel. Normally this would be the difference between repeated measurements from the same loudspeaker, or measurements from an array of different loudspeakers.

An option to *Overlap Sweeps* is also presented. This should be used checked if a multiple swept sine measurement approach is being used (Majdak, Balazs and Laback, 2007). If this is selected, the *Group Size* should be defined as well as the *Interleave* and *Overlap* delays. The interleave delay is the time delay between individual sweeps within a group and the overlap delay is the delay between each group. To optimize the time under test these delays should be minimised. However, they depend on the length or the HRTFs/reverberation time of the room and the presence of harmonic distortions that may interfere with SNR requirements.

Initial Rig Calibration: With reference to Fig. 6.21, the *Temporal* option is a



Target (samp.) 1706 +/- 75

Window (samp.): In 50 Pre-Peak 100 Post-Peak 75 Out 25

Naming Convention azi_0,0_ele_90,0.wav azi_45,0_ele_64,8.wav azi_135,0_ele_64,8.wav azi_225,0_ele_64,8.wav azi_315,0_ele_64,8.wav azi_0,0_ele_45,0.wav azi_90,0_ele_45

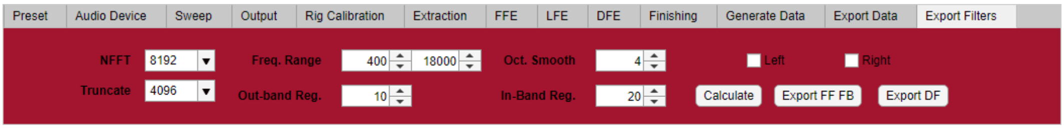
Figure 6.22: MARC Workflow: Extraction settings.

string of numbers (one for each loudspeaker) that defines any differences in radial distances between the loudspeakers relative to the center of the array. This information is used to align the sweeps in the time domain at the center of the array. For example, if the third loudspeaker is defined to be +0.2m then the third sweep will be compensated forward in the time domain to arrive at the center of the array at the same time as a loudspeaker specified as +0m.

Corresponding *Azimuths* and *Elevations* of the HRTF locations to be measured should also be input. This data should be listed in the same order as the output channel mapping. The data is used during diffuse-field equalisation processing to fairly weight the measurements based on their distribution. There is also an option to adjust the *Phase* of each loudspeaker (±) to compensate for any physical inversions (e.g. wiring a loudspeaker the wrong way around). The output *Level* of each loudspeaker can also also be calibrated but should be left blank at this stage.

Extraction: With reference to Fig. 6.22, the extraction settings are used to isolate and window the individual HRTFs after the recorded sweeps have been deconvolved but before any post-processing. This is a fairly precise operation and requires some manual input. The *Target* value defines at which sample the program searches for peaks in the recorded waveforms that would represent an HRTF. It is dependant on the delay of the system, input/output buffers, time of flight (distance between the loudspeakers and the microphones), among other factors. The easiest way to calculate the appropriate value for an individual set up is to take a set of measurements using a single mono microphone placed in the centre of the array and look at the sample number of the first peak in the RAW deconvolved output. A window of error is required to account for temporal discrepancies such as ITD and so the option is given to specify both a target and a range.

The *Window* settings are defined by 4 lengths (given in samples). The *In* and *Out* parameters represent half-Hanning windows that are applied at the start and end of the window. *Pre-Peak* and *Post-Peak* define the number of samples that are left



NFFT 8192 ▼
Truncate 4096 ▼
Freq. Range 400 18000
Out-band Reg. 10
Oct. Smooth 4
In-Band Reg. 20
□ Left
□ Right

Figure 6.23: MARC Workflow: Export Filter settings.

un-altered either side of the HRTF peak. The filenames that the files are saved under must also be defined under *Naming Convention.*

Equalise Loudspeaker Amplitude: At this point it is appropriate to take the first calibration measurement and equalise the amplitudes of each amplifier/loudspeaker. The measurement should be taken with a flat response measurement microphone placed in the center of the array. The program performs an average RMS analysis of the amplitude of each measurement after deconvolution and extraction and outputs the result to the console. This data can be copied and pasted into the *Level* option in Fig. 6.21: *Rig Calibration* settings.

The software uses this data to compensate the amplitude of each sweep. To prevent the chance of digital distortion this equalisation is implemented by reducing the amplitude of sweeps output to the louder loudspeakers. Generally this results in an overall reduction in the volume of the audible sweeps so it is common to boost the absolute sweep amplitude under the *Sweep* settings at this stage.

Design Preliminary Pre-Processing Filters: With reference to Fig. 6.23, the second step in calibration is to adjust the frequency response of the sweeps. This is done using the *Export Filters* tab. Frequency responses are calculated for the left and right channels of each isolated measurement. Inverse filters are then generated from the parameters described below:

NFFT: length of inverse filter during calculation

Truncate: length of final inverse filter truncated by hanning window

Frequency Range: defines the *In-Band* frequency range

In-Band Regularisation: maximum frequency response compensation (dB) within the in-band frequency range

Out-Band Regularisation: maximum frequency response compensation (dB) outside of the in-band frequency range

Octave Smoothing: smoothing applied to the inverse filter in $1/x$ octave bands

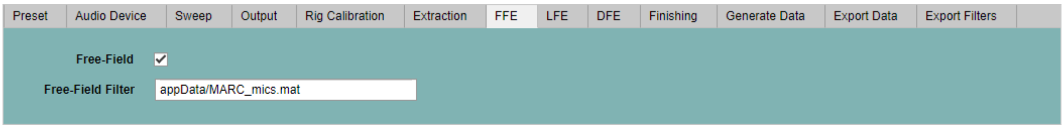
It is important at this stage not to overcompensate bass frequencies as this will result in over driving the loudspeakers and will lead to significant harmonic distortion. As such it is better to respect the natural low end frequency response of the speakers (and correct for this in the later post-processing stages instead). Compensation should therefore be limited to approximately $>$400Hz, depending on the loudspeaker. There is no strict requirement to be overly precise. The main objective at this stage is to reduce any obvious peaks in the frequency response in order to increase the overall amplitude of the sweep without amplifying certain frequencies beyond a comfortable listening level. This will in general increase the SNR across a wider range of frequencies.

The filters can be computed and saved for either the left or right channels singularly, or as a stereo pair. The program is capable of outputting either a bank of free-field equalisation filters (one for each measurement) or a single diffuse-field equalisation filter (one that best represents the entire set of measurements). In this case a single channel bank of free-field filters should be exported (the responses have only been measured through a single measurement microphone).

Once the filters have been saved, they can be applied to the sweeps by directing the program to the appropriate file in the *Sweep* settings tab. Once again, the application of the filters results in an overall reduction in audible amplitude and so the absolute sweep amplitude may need to be adjusted.

Final Calibration: Once the sweeps have been equalised with respect to frequency it is advised, although not strictly essential, to recalibrate the loudspeaker levels. The original level compensation values should be erased and another measurement should be taken with a flat response microphone. Once the final calibration values are calculated and input the absolute sweep amplitude may be adjusted to the most appropriate value.

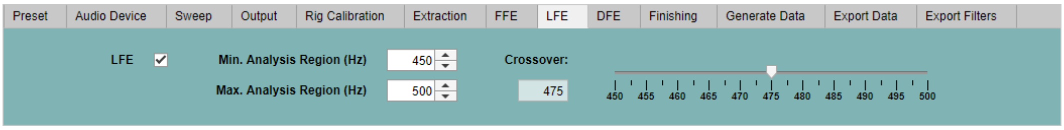
At this point the system is calibrated to output an amplitude matched, equalised flat response exponential swept sine sweep from each loudspeaker and is ready to begin measurements.



Free-Field ☑

Free-Field Filter appData/MARC_mics.mat

Figure 6.24: MARC Workflow: FFE settings.



LFE ☑

Min. Analysis Region (Hz) 450

Max. Analysis Region (Hz) 500

Crossover: 475

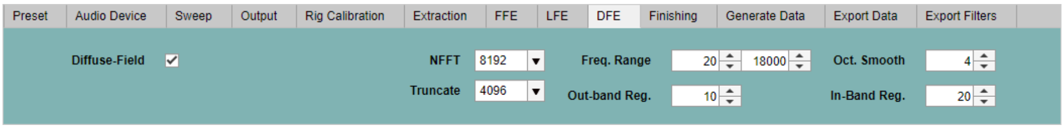
Figure 6.25: MARC Workflow: LFE settings.

Free-Field Equalisation: With reference to Fig. 6.24, the next step is to prepare the free-field equalisation filters to fully compensate for both the remaining loudspeaker frequency responses and the responses of the binaural microphones to be used in the HRTF measurements. A set of measurements should be taken with the binaural microphones placed in the center of the array. Ideally there should be minimal acoustic occlusion from any apparatus. It may be easiest to hang the microphones using their own cables/string. Although the microphones are ideally omnidirectional, in reality this will not be exactly the case. It is therefore beneficial to position the microphones such that the right microphone points roughly towards the right side of the array and vice versa for the left microphone. This ensures that the best matched compensation filters are calculated for the perceptually dominant ipsilateral sources.

Again, the filters may be calculated in the *Export Filters* tab. It is now appropriate to equalise the lower frequencies, <200Hz. This need not be done excessively, as the next post-processing step is to apply a low frequency model. However, it will help to flatten the response to at least the crossover frequency to ensure the model is being applied at the correct amplitude. This time a stereo bank of free-field filters should be output. The filters should be applied within the *FFE* tab.

Low Frequency Extension: With reference to Fig. 6.25, the next step is to apply a low frequency model to the measurements to compensate for the inadequate frequency responses of the miniature loudspeaker drivers. First developed by Xie, 2013, this model is very similar to the one used in the SADIIE database, as described in Section 3.3.3, but with some minor optimizations.

Generally speaking, a low-pass filtered Dirac pulse is delay and amplitude matched to the crossover frequency of the original High-pass filtered HRTF. The pulse is



Diffuse-Field ☑

NFFT 8192

Truncate 4096

Freq. Range 20 18000

Out-band Reg. 10

Oct. Smooth 4

In-Band Reg. 20

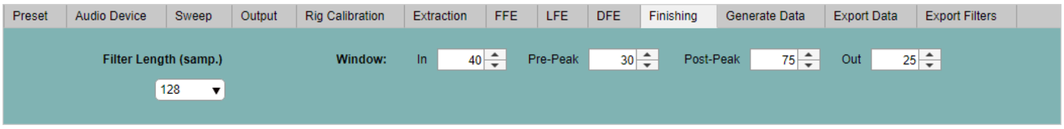
Figure 6.26: MARC Workflow: DFE settings.

then overlaid onto the HRTF signal using crossover filters. In the SADIIE database, approximations of the HRTF's group delay used to delay match the Dirac pulse were made by taking the mean group delay calculated over the frequency range ± 25Hz either side of the crossover frequency. The small frequency range resulted in erroneous group delay estimations that caused the Dirac model to lag the HRTF signal. This was compensated for by ensuring the Dirac pulse was only ever shifted forward in time to phase match the signals in the next stage of the model.

The new method calculates the median group delay over the range 450Hz to 1000Hz. It was noted that the average group delay hardly changed within this extended frequency range and the result is a much closer estimation to the actual group delay of the filter (in general, and at the crossover frequency). The requirement that the Dirac model was then shifted *forward* in time to align the phase with the HRTF signal (as in SADIIE) was therefore removed. Hence the Dirac spike is now shifted either forward or backward in time by the minimum amount ($-\pi <$ phase shift $< \pi$) so as to align the phase with the HRTF signal. The amplitude of the Dirac spike is still matched to the mean amplitude around the desired crossover frequency, in this case ±50Hz.

The frequency region over which the correct delay and amplitude is calculated is defined by the *Minimum* and *Maximum Analysis Region* options. These should be within a few hundred Hz of the crossover frequency and should not exceed the accurate response of the measurement system compensated for by the Free-Field Equalisation filters. The exact crossover frequency is defined by the sliding scale.

Diffuse-Field Equalisation: With reference to Fig. 6.26, the final post-processing step is the diffuse-field equalisation of the measurements. The filter is calculated using the same parameters as are defined in the *Export Filters* tab. It may in fact be useful the use the *Export Filters* tab to visualise the filter shape and then transfer across the appropriate setting values. A solid angle weighting is applied to the measurements based on their angle to ensure a fair calculation of the diffuse-field given a non-regular distribution of samples.



Filter Length (samp.) 128

Window: In 40 Pre-Peak 30 Post-Peak 75 Out 25

Figure 6.27: MARC Workflow: Finishing HRTF settings.

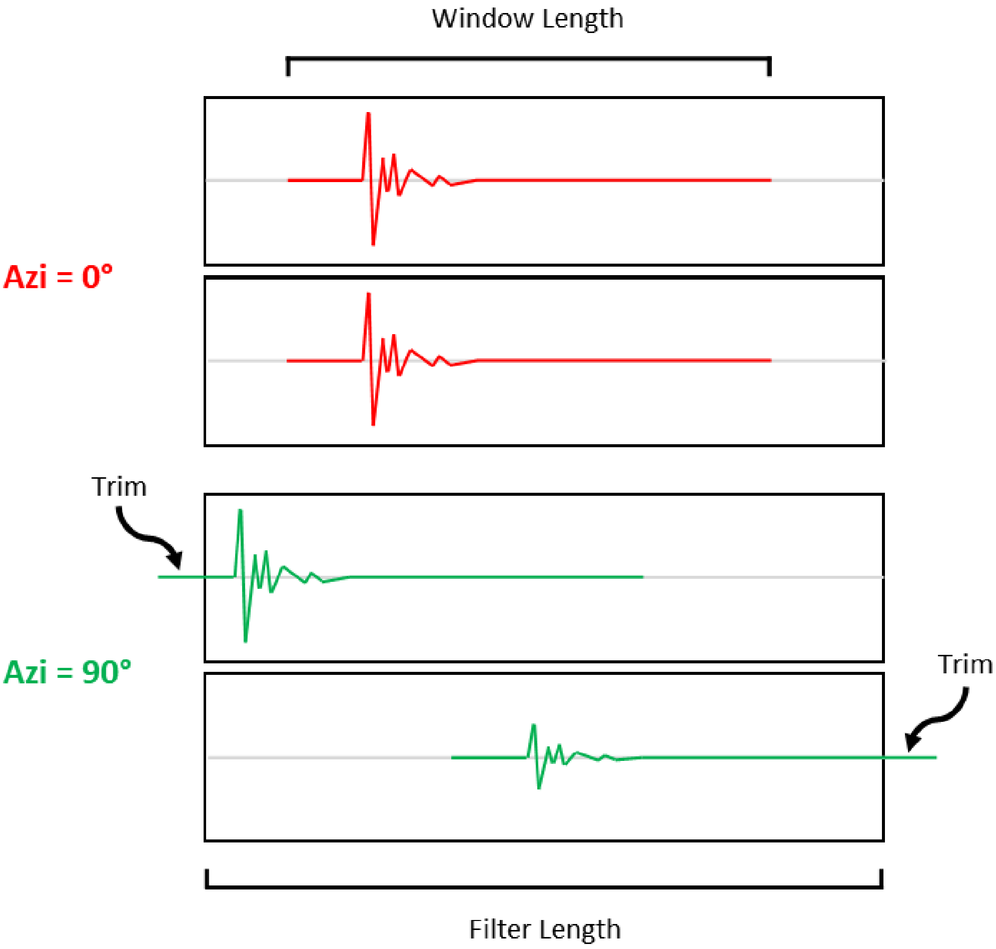
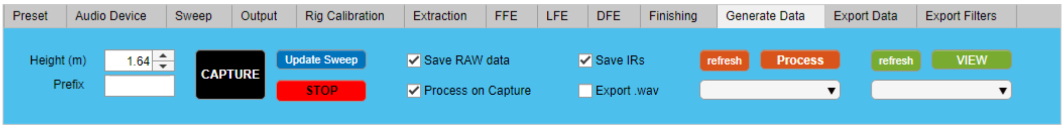


Figure 6.28: Trimming of HRTFs that exceed the filter length due to ITDs despite being windowed to fewer samples.

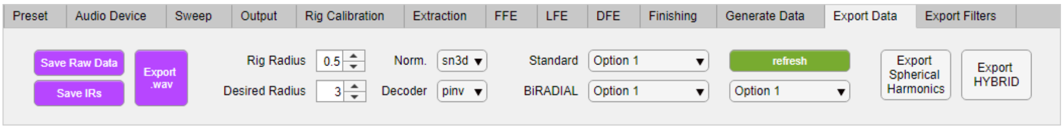
Finishing: With reference to Fig. 6.27, once the filters have been processed they must be windowed and trimmed to their final length. Again, the *Window* shape and size may be defined by two opening and closing half-Hanning windows and a pre- and post-peak pad. The length to which the filters are trimmed is defined by *Filter Length* and should be set in terms of samples.

Ideally, the total window length should be shorter that the filter length. However, due to ITDs it may be the case that as a set some of the filters exceed the desired filter length, for example, Fig. 6.28. In that case the start and end of the overlapping filters are trimmed equally and a short opening/closing Hanning window is applied to either edge. A warning is also displayed to the console.



Height (m) 1.64
Prefix
☑ Save RAW data
☑ Process on Capture
☑ Save IRs
□ Export .wav

Figure 6.29: MARC Workflow: Generate Data settings.



Rig Radius 0.5
Desired Radius 3
Norm. sn3d
Decoder pinv
Standard Option 1
BiRADIAL Option 1
Option 1

Figure 6.30: MARC Workflow: Export Data settings.

Measurement: With reference to Fig. 6.29, once the rig is calibrated and the post-processing parameters have been set the system is ready to *Capture* a set of HRTFs. The person being measured should insert the binaural microphones into their ears and step into the rig. The internal frame should then be positioned to the correct height and the person should align themselves to the center of the array using the laser alignment system. The *Height* of the internal frame must be input in order to correctly compensate for the variable radius of the bottom loudspeaker fixed to the floor. If changes have been made to the sweep setting, the program should be updated with *Update Sweep*.

Options are available to automatically save *RAW Data* (the completely unprocessed recordings taken directly from the microphones) and *RAW IRs* (after deconvolution of the exponential sweeps and separation of the HRTFs). This data allows for changes to be made to the extraction and post-processing settings and for the final HRTFs to be re-generated without requiring the person to undergo a second measurement. As such, the tab also includes options to load previously recorded data and re-process it using the drop down options on the right to select, *Process* and *View* the HRTFs.

Two further options are available to automatically *Process* and *Export* the HRTFs once the measurement is complete. As exporting multiple .wav files takes some time it is useful to turn this feature off if multiple measurements are to be taken in quick succession. There is also an emergency *STOP* button which immediately mutes all output and aborts the measurement.

Export: With reference to Fig. 6.30, the final tab presents the options to export the data in a number of formats. The most basic output is the measured HRTFs in either .wav of .sofa format. Further options include the implementation of novel methods to simulate far-field HRTFs from the measured signals using *Export Spherical Harmonics* and *Export Hybrid*. These methods are discussed in Section 6.5. Near-field compensation and format options are selectable for the spherical harmonic formats.

Displays of recorded data pre- and post-extraction, time domain plots, frequency domain plots and diffuse-field response plots throughout the post-processing chain allow the operator to quickly analyse and assess the basic quality of the measurements.

6.5 Simulation of Far-Field Measurements

6.5.1 Overview

Whilst near-field measurements are convenient in a practical sense, they do not provide optimal externalisation for a binaural renderer. Generally, there are no techniques to independently alter a single HRTF measurement to make it appear as though it was measured from a greater distance. However, by exploiting a high quality binaural Ambisonic renderer inclusive of NFC filters it is possible to synthesise accurate planar far-field HRTFs using the near-field measurements e.g. Pollow et al., 2012; Duraiswami, Zotkin and Gumerov, 2004.

The accuracy of a traditional (loudspeaker-based) Ambisonic workflow is limited by the spatial aliasing frequency. In the case of 5^{th} order Ambisonics this is approximately 3kHz (Daniel, Rault and Polack, 1998). However, the synthesis of HRTFs only requires the system to produce a binaural output, therefore, a binaural Ambisonic workflow may be implemented instead. It is therefore possible to implement Time-Alignment or BiRADIAL techniques to significantly improve the high-frequency accuracy of the rendered signals.

There are several ways in which this synthesis may be performed with variable output performance. Some of the techniques allow exportation of synthesised HRTF filters whilst others skip this step and simply aim to directly emulate far-field sources within the renderer itself.

6.5.2 Near-Field Compensation of HRTFs

Implementing the NFC filters required to simulate far-field HRTFs from measurements taken by MARC is a more complex procedure than in the standard case.

Typically, NFC filters are applied to an Ambisonic input signal within the spherical harmonic domain to either simulate a source of finite radii or counter the physical effects of a non-infinite loudspeaker array. It is generally assumed that the reproduction array is of a fixed and constant radii. This is not the case for MARC.

MARC captures HRTF measurements at a range of different radii due to the space required for a person to walk in and out of the rig. This means that the measurements are not directly suitable for virtual Ambisonic rendering as the wavefront curvatures of the different loudspeakers are not the same at the center of the array. The 3D signals therefore do not superimpose with each other as they are expected to during their derivation based on the predicted mode excitation within the array. Further, taking the approach of filtering an incoming Ambisonic signal leads to increases in real-time complexity which are best avoided.

Individualised compensation filters are therefore applied directly to the HRTFs. To do this, the HRTFs are first transformed into the spherical harmonic domain, as would be the case in a typical spherical harmonic decoder. It then does not matter whether the compensation filters are applied to the incoming Ambisonic signal or to the spherical harmonic representation of the HRTFs. Each HRTF is encoded and filtered individually before summing the individual spherical harmonic representations of each HRTF.

$$\sum_{l=1}^{L} H_m \cdot Y_k(\vec{v}_l) \cdot h_l \tag{6.5}$$

for all k such that $H_m \cdot Y_k(\vec{v}_l) \cdot h_l$ forms a column vector of length, k, of spherical harmonic encoded HRTFs and for each stereo channel where:

- L is the number of loudspeakers,
- H_m are the degree dependant distance compensation filters,
- K is the number of Ambisonic Channels,
- $Y_k(\vec{v}_l)$ is the Decoding matrix coefficient representative of the Ambisonic channel, k, for the loudspeaker l, for the configuration of loudspeakers, L,
- h_l is the binaural filter measured from the position of that loudspeaker,

This allows each measurement to be compensated individually with respect to its specific radii compared to the other loudspeakers.

Fig. 6.31 describes this workflow. Each HRTF is encoded into spherical harmonics using the weights held in the relevant column of the transposed decoding matrix.

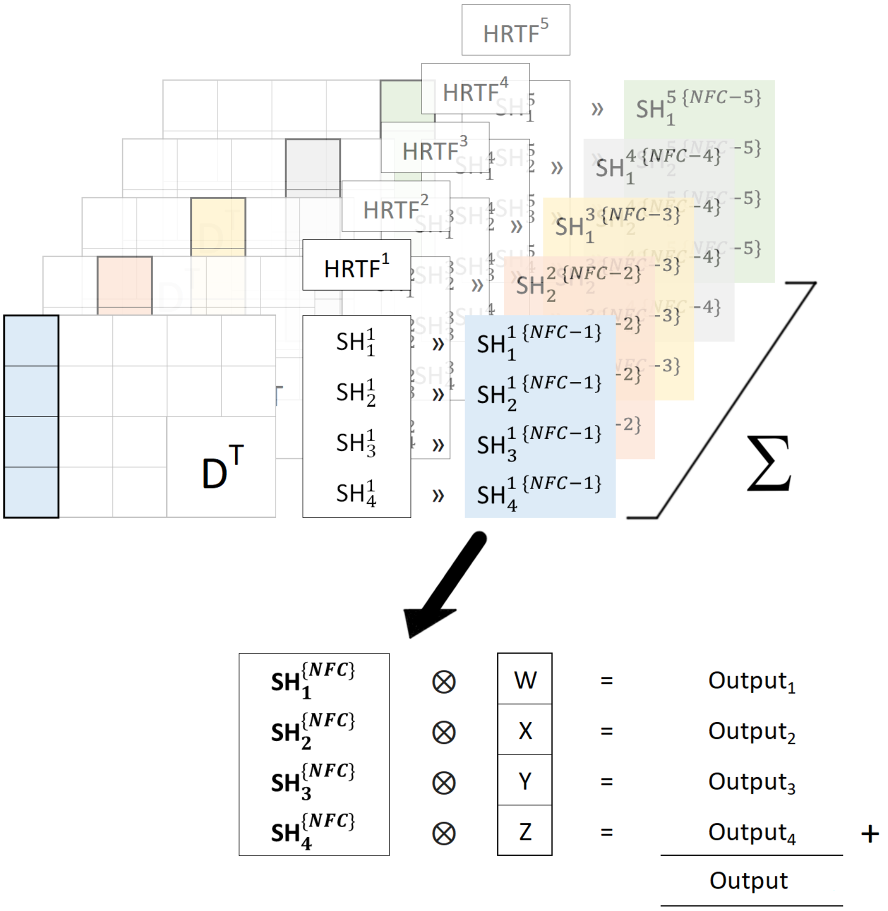


Figure 6.31: Workflow of individual NFC filtering of HRTFs. Each HRTF is encoded into spherical harmonics and individually filtered by NFC filters based on the radial value of that particular measurement. The results are then summed over each spherical harmonic channel.

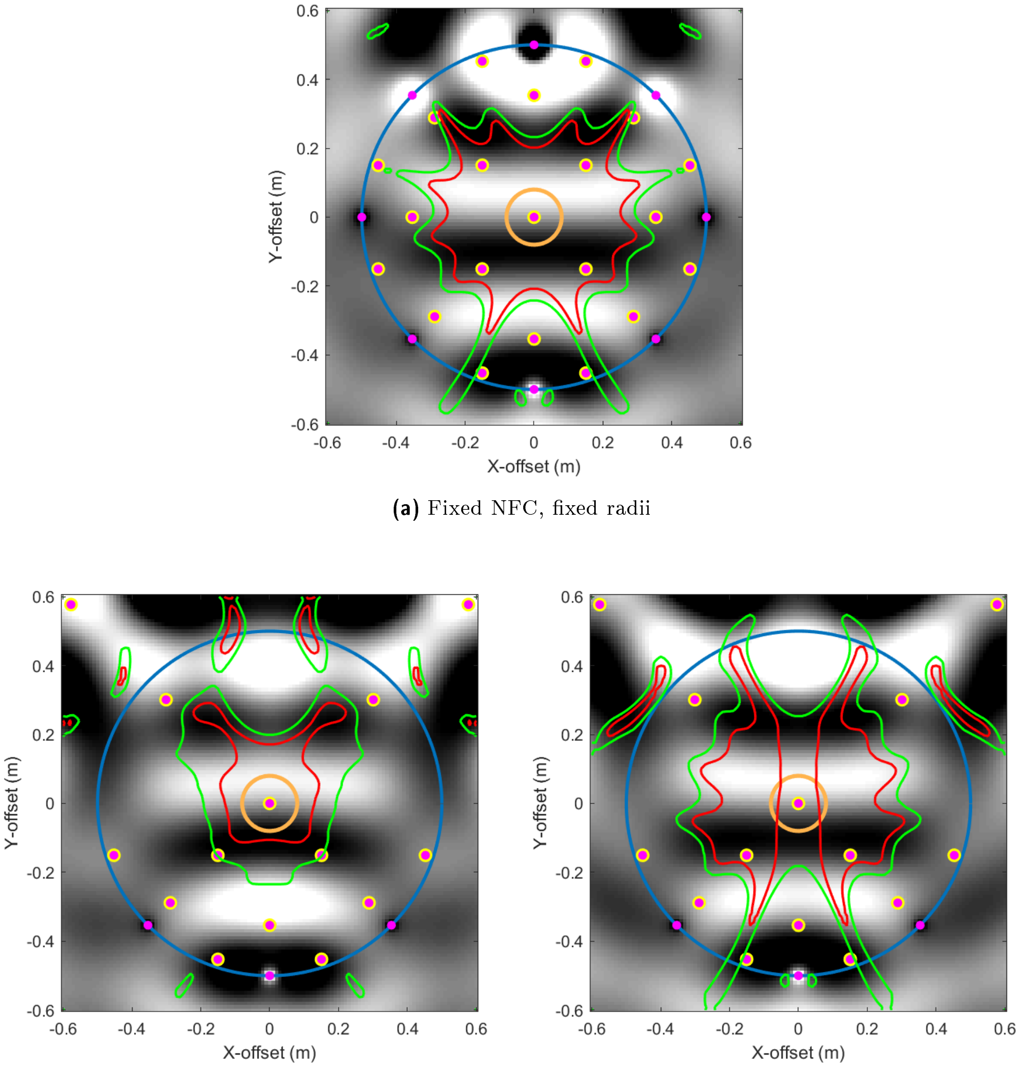
The encoded HRTFs are then each individually filtered by NFC filters that are designed based on the radial value of that particular measurement. As the aim is to synthesise planar HRTFs using this technique it is recommended that NFC filters are used that adjust the wave front curvature to match a significantly far-field source, for example 100m (i.e. significant bass cuts in addition to phase adjustments).

The approach is not exact; NFC filters are designed to manipulate the spherical harmonic components in such a way that the frequency and phase responses of each and every loudspeaker are adjusted to reshape the wavefront. When different filters are applied to the spherical harmonic representations of each HRTF their combined affects are no longer coherent. That being said, it may be demonstrated through the following simulations that this approach better represents planar HRTFs than if the variation in measurement radii is simply ignored.

Fig. 6.32 shows the reproduced soundfield of a distance compensated Ambisonic source reproduced over a standard fixed radii array and over a multi-radii array. In Fig. 6.32a the spherical harmonic components are correctly compensated to reproduce a planar wave front. In Fig. 6.32b the radii of half of the loudspeakers is doubled but the same compensation filters are applied. By increasing the loudspeaker radii they better approximate infinite sources. An infinite source would require no distance compensation. Therefore, by not reducing the compensation of the NFC filters the wave front curvature is effectively being overcompensated. As a result the wave front begins to curve in the opposite direction (note the location of the encoded source is straight ahead). In Fig. 6.32c the compensation applied to the more distant loudspeakers is reduced and a planer wave front curvature is restored.

Examination of Fig. 6.32c reveals a region of imperfect reproduction ($10\% <$ error $< 20\%$) through the center of the array. This error is due to an error in reproduction amplitude caused by the inconsistent NFC filters. The error is better shown in Fig. 6.33 which plots the sound pressure of the central point in the array over time for one full wave cycle.

NFC filters apply a frequency dependant amplitude and phase shift to the spherical harmonic components that result in similar manipulations to the resultant loudspeaker signals. The combination of these manipulations have the effect of reshaping the wave front reproduced by the speakers. By opting to implement multiple



(a) Fixed NFC, fixed radii

(b) Fixed NFC, variable radii

(c) Variable NFC, variable radii

Figure 6.32: Ambisonic soundfield reconstructions of a frontal source distance compensated to a radius of 100m. (a) has a loudspeaker radius of 0.5m. (b) and (c) have variable loudspeaker radii: front hemisphere = 1m, rear hemisphere = 0.5m. Fixed NFC filters designed for a loudspeaker radius of 0.5 m have been applied in (a) and (b). Variable NFC filters have been applied in (c). Loudspeaker positions are shown as pink dots. The source is a sine wave of 1000Hz. Amplitude is plotted on a scale of black to white. An orange ring of radius 8cm is plotted in the center of each array to indicate the approximate size of a human head. Areas of accurate reproduction are highlighted by contours: $< 20\%$ error (green); $< 10\%$ error (red).

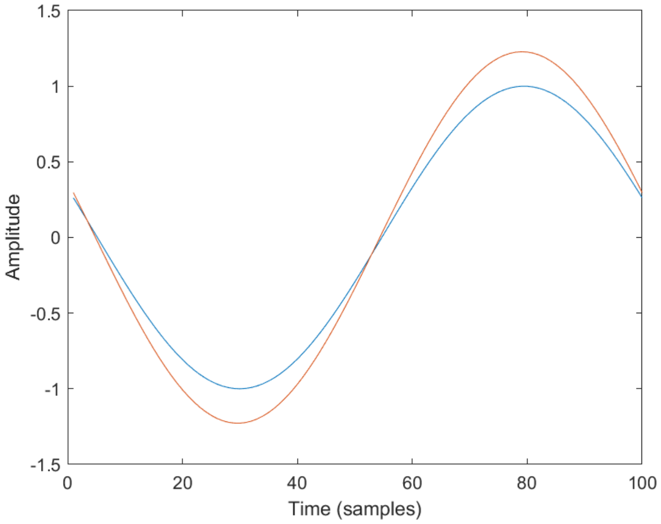


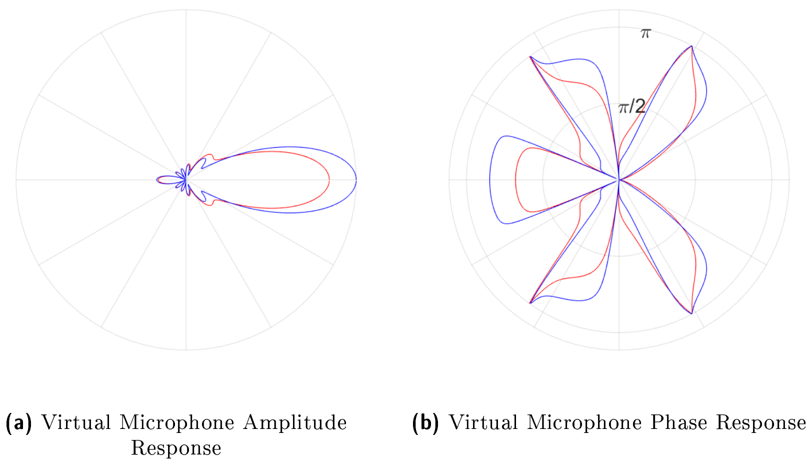
Figure 6.33: Amplitude error of reproduced wave due to inconsistent NFC filters. Blue: original wave. Orange: Reproduced wave.

and independently calculated NFC algorithms it is no longer fair to assume that the manipulations remain correctly balanced.

Fig. 6.34 displays the absolute amplitude and phase responses of the virtual microphone pick up patterns for a 5^{th} order projection decoded 1000Hz Ambisonic source simulated at a distance of 100m to be rendered over a 0.5m and 1m radii loudspeaker array. It is shown that within the frontal hemisphere the phase differences between the two NFC algorithms are similar but that the amplitude response of the $100 \rightarrow 1m$ algorithm is inflated. This increase in amplitude is usually compensated by the response of the rear hemisphere in which the amplitude differences are minimal, but the general phase response of the $100 \rightarrow 1m$ algorithm is closer to 180°. The increase in phase difference leads to greater destructive interference of the signals entering the array and therefore the overall amplitude of the reproduced wave front is maintained in either case.

Referring back to the example in Fig. 6.32, a source is rendered over a loudspeaker array in which the front hemisphere of loudspeakers have been filtered using a $100 \rightarrow 1m$ NFC filter and the rear hemisphere of loudspeakers have been filtered using a $100 \rightarrow 0.5m$ filter. As a result, the inflated amplitude response in the frontal hemisphere is not counter balanced by an increased phase difference in the rear hemisphere. This results in the increase in amplitude shown in Fig. 6.33.

The impact of multiple NFC algorithms is frequency dependant as the amplitude and phase corrections of the filters are applied mainly to the lower end of the spectrum.



(a) Virtual Microphone Amplitude Response

(b) Virtual Microphone Phase Response

Figure 6.34: Absolute amplitude and phase responses of the virtual microphone pick up patterns for a source (1000Hz) panned to 0° and simulated at a distance of 100m to be rendered over a 0.5m and 1m radii loudspeaker array using NFC filters. Blue: the response of the 100 → $1m$ NFC filter. Red: the response of the 100 → $0.5m$ NFC filter.

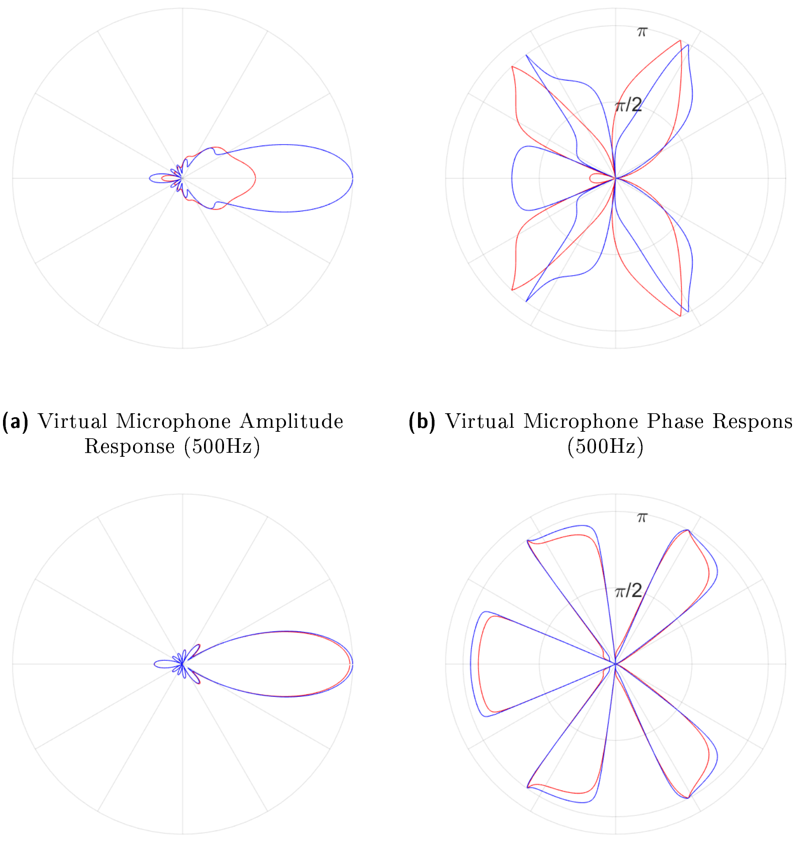
This may be demonstrated by plotting the virtual microphone responses for the same 5$^{\text{th}}$ order Ambisonic source at a frequency of 500Hz and 3000HZ, as in Fig. 6.35. Note how the responses are settling towards a more typical pickup pattern, as seen in Fig. 4.10b, by 3000Hz.

It is found that preserving the wave front curvature over the exact amplitude gives a more consistent and reliable timbre in the reproduced signal, providing the variation in loudspeaker radii is not excessive. This method is therefore implemented within the post-processing stages of MARC.

6.5.3 Methods of synthesis

Further steps may be taken in addition to NFC to simulate far-field measurements from MARC measurements using a binaural Ambisonic rendering approach. 4 approaches are evaluated, as are presented in Fig. 6.36. The methods utilize either Time-Alignment or BiRADIAL rendering methods to either reproduce a far-field reproduction array directly (1 and 2), or via an intermediate far-field representation of HRTFs (3 and 4).

The encoding of HRTFs into spherical harmonics and application of the multi-radial NFC method outlined in Section 6.5.2 is depicted by an orange block. In addition, a dual-band (basic/max-$\boldsymbol{r}_E$) decoder is implemented by manipulating the weights on the spherical harmonic components in the very high frequencies ($>$ 15000Hz). A similar conversion is depicted by a purple box, but without the application of any NFC filters.



(a) Virtual Microphone Amplitude Response (500Hz)

(b) Virtual Microphone Phase Response (500Hz)

(c) Virtual Microphone Amplitude Response (3000Hz)

(d) Virtual Microphone Phase Response (3000Hz)

Figure 6.35: Absolute amplitude and phase responses of the virtual microphone pick up patterns for a a&b) 500Hz and c&d) 3000Hz) panned to 0° and simulated at a distance of 100m to be rendered over a 0.5m and 1m radii loudspeaker array using NFC filters. Blue: the response of the 100 → $1m$ NFC filter. Red: the response of the 100 → $0.5m$ NFC filter.

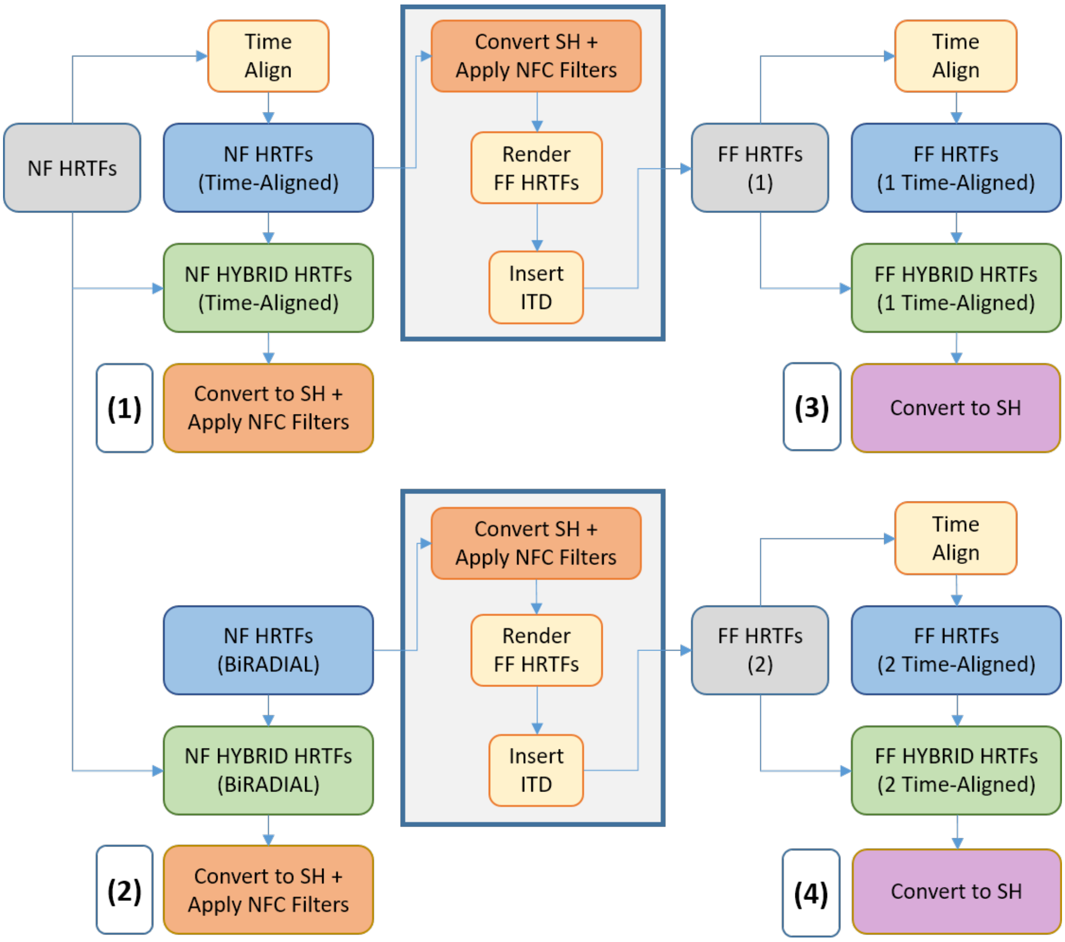

Time Align
NF HRTFs
NF HRTFs (Time-Aligned)
NF HYBRID HRTFs (Time-Aligned)
(1)
Convert to SH + Apply NFC Filters
Convert SH + Apply NFC Filters
Render FF HRTFs
Insert ITD
FF HRTFs (1)
Time Align
FF HRTFs (1 Time-Aligned)
FF HYBRID HRTFs (1 Time-Aligned)
(3)
Convert to SH
NF HRTFs (BiRADIAL)
NF HYBRID HRTFs (BiRADIAL)
(2)
Convert to SH + Apply NFC Filters
Convert SH + Apply NFC Filters
Render FF HRTFs
Insert ITD
FF HRTFs (2)
Time Align
FF HRTFs (2 Time-Aligned)
FF HYBRID HRTFs (2 Time-Aligned)
(4)
Convert to SH

Figure 6.36: An outline of the 4 methods (1-4) suggested to produce far-field Ambisonic decoders (weighted HRTFs encoded into spherical harmonic format) from a set of near-field measurements. The 3-step process highlighted by the light blue background is the synthesis of far-field HRTFs.

Direct, Time Alignment (1 - see Fig. 6.36) The first method is the most simple and requires the fewest physical measurements. Hybrid HRTFs are generated by time-aligning the high frequency portion of the near-field HRTFs. They are then encoded and distance compensated in the spherical harmonic domain ready to be implemented within a spherical harmonic format binaural Ambisonic decoder.

Direct, BiRADIAL (2) The second method is almost identical to the first but requires BiRADIAL HRTFs to replace the time-aligned standard measurements in the high frequencies.

Through Intermediate Far-Field Representation, Time Alignment (3) The third method implements a two stage approach and allows for the actual synthesis of natural far-field HRTFs. The technique first encodes and distance compensates full-band time aligned HRTFs in the spherical harmonic domain.

Single delta functions are then encoded with standard coefficients as Ambisonic sources into spherical harmonics and convolved with the HRIRs to synthesis time aligned far-field HRIRs. This may be done for as many directions as required. ITDs are then calculated using a spherical head model based on the desired directions and radius of the measurements and are re-introduced to the HRIRs by means of a discrete sample delay to give realistic approximations of the far-field filters.

A typical Time-Alignment decoder is then constructed using these synthesised measurements in place of the near-field ones. A frequency dependant time delay is introduced and hence far-field equivalent Hybrid HRTFs are derived. The HRTFs are represented in spherical harmonic format ready to be implemented within a spherical harmonic format binaural Ambisonic decoder.

Through Intermediate Far-Field Representation, BiRADIAL (4) The final method is almost identical to the third but requires BiRADIAL HRTFs to replace the time-aligned standard measurements in the high frequencies during the synthesis of the far-field HRTFs.

A Time-Alignment approach is still used within the final renderer (implementing the BiRADIAL synthesised far-field HRTFs) as the technique is equivalent to BiRADIAL decoding given the planar nature of the synthesised measurements. Further, it does not require the synthesis of a second set of far-field BiRADIAL HRTFs.

6.5.4 Evaluation

The 4 methods of far-field HRTF synthesis were evaluated though the author's critical listening and analysis of the PSDM, ITDs and ILDs. 5^{th} order binaural Ambisonic renders using KU100 HRTFs measured in MARC were compared to far-field reference HRTFs. As very far-field reference measurements were not available, analysis was performed by modelling a radial distance of just 1.2m and comparing the output to the SADIIE Database. Differences between modelling a distance of 1.2m and 100m were then compared directly. The analysis here is carried out over the horizontal axis as this remains the primary focus for many current applications, however, the results presented in Table 7.2 demonstrate how the accuracy of these results are in fact representative of results over the entire sphere.

In total 6 renderers are considered. The first 2 are the direct NFC filtered spherical harmonic representations of the HRTFs (Time Alignment and BiRADIAL). A further 2 are derived from the intermediate far-field representations of the near-field HRTFs - the result of synthesising and re-encoding 50 far-field HRTFs arranged in a Lebedev Grid (Time Alignment and BiRADIAL). The final 2 are the same but for 2702 points. Within each pair both a Time-Alignment and BiRADIAL approach are considered.

Comparing generally the direct representations vs the intermediate far-field representations the intermediate far-field representations better reproduce ITDs at the extreme lateral angles on the horizontal plane, as shown in Fig. 6.37. ILDs were similarly rendered, see Fig. 6.38,

The spectral response is in general also quite similar, however, there exists a small increase in the overall amplitude in the >15kHz band in each case. The increase in amplitude is greater in the intermediate far-field representation method, see Fig. 6.39.

It is thought that this amplitude discrepancy is a result of the energy preservation of the max-$\boldsymbol{r}_E$ weighting scheme. Note the lack of any significant red colour in Fig. 6.39a. The effect is made worse when the energy preserving weighting is applied twice (once in each stage) in the intermediate far-field representation method. The max-$\boldsymbol{r}_E$ scheme increases the width of the frontal lobe and therefore the spread of a source over multiple reproduction loudspeakers. In order to preserve the reproduction energy the overall reproduction amplitude must therefore be increased.

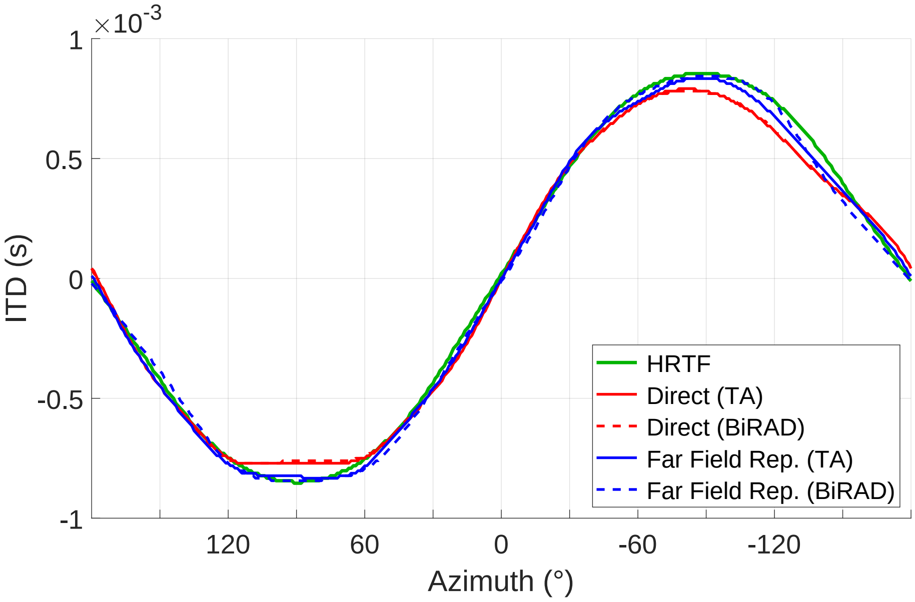


Figure 6.37: Comparing ITD estimations for direct NFC filtered spherical harmonic rendering against intermediate generation and re-synthesis of far-field HRTFs

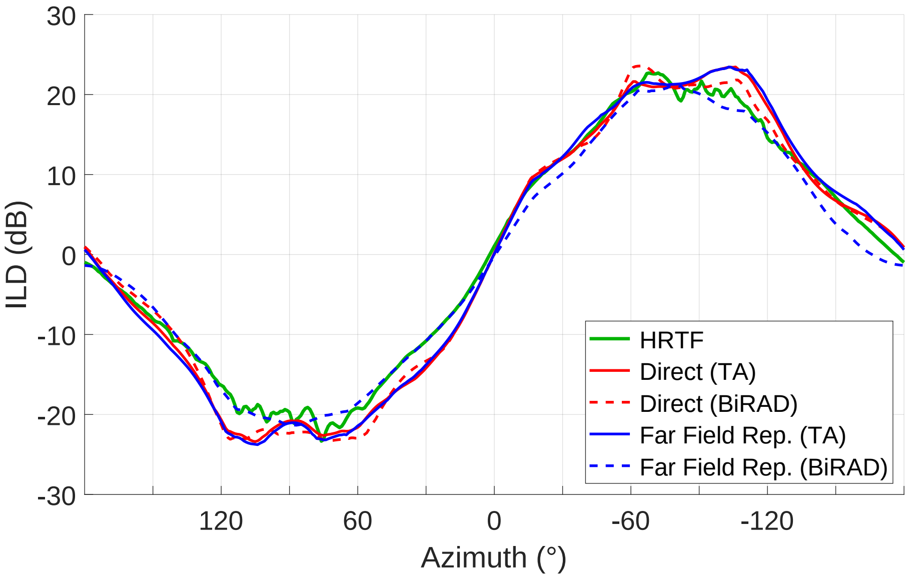
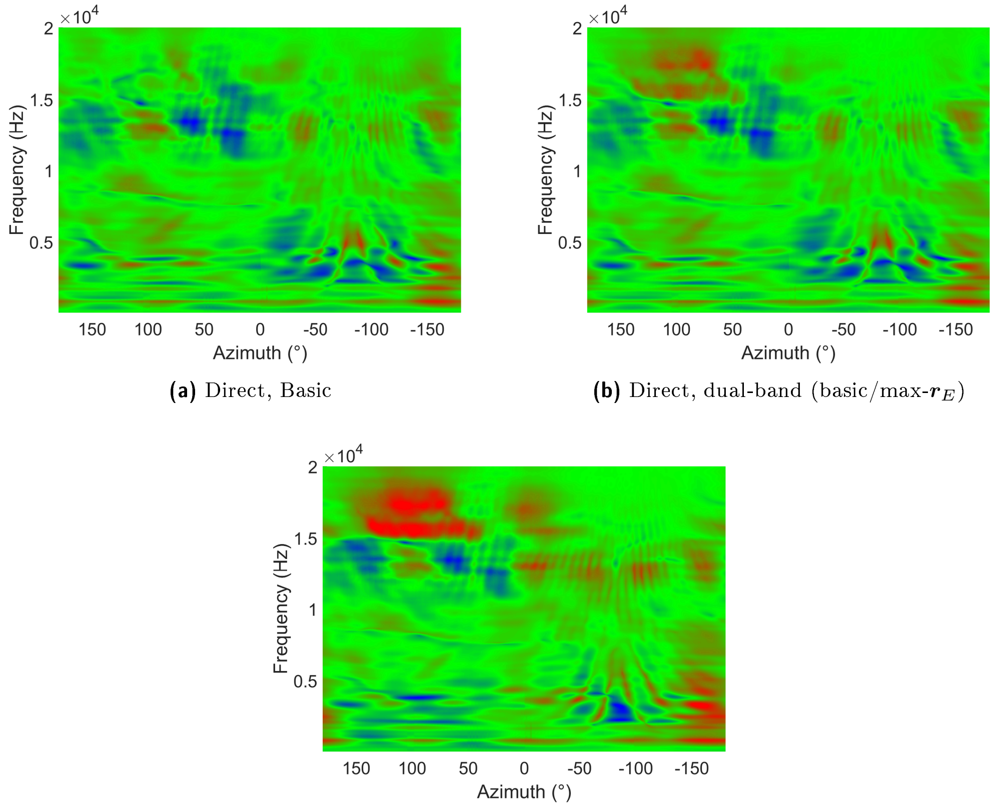


Figure 6.38: Comparing ILD estimations for direct NFC filtered spherical harmonic rendering against intermediate generation and re-synthesis of far-field HRTFs

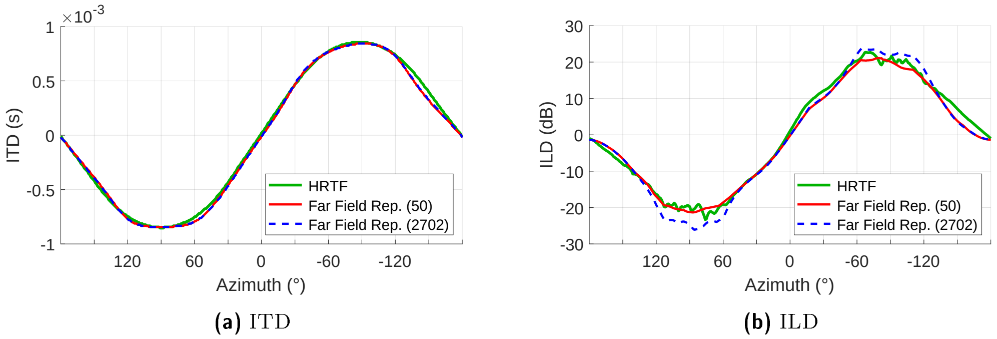


(a) Direct, Basic

(b) Direct, dual-band (basic/max-$\boldsymbol{r}_E$)

(c) Far-Field Rep., dual-band (basic/max-$\boldsymbol{r}_E$)

Figure 6.39: Spectral response comparisons of BiRADIAL MARC renderers: basic and dual-band (basic/max-$\boldsymbol{r}_E$) direct NFC filtered spherical harmonic rendering and intermediate generation and re-synthesis of far-field HRTFs (50 points). Comparisons are made with respect to SADIIE HRTFs. Red indicates an amplitude difference >8dB whilst blue indicates a difference of <-8dB. Green indicates an accurate reproduction. The increased amplitude in >15kHz band is a result of the energy preserving normalisation scheme of the max-$\boldsymbol{r}_E$ decoding weights.

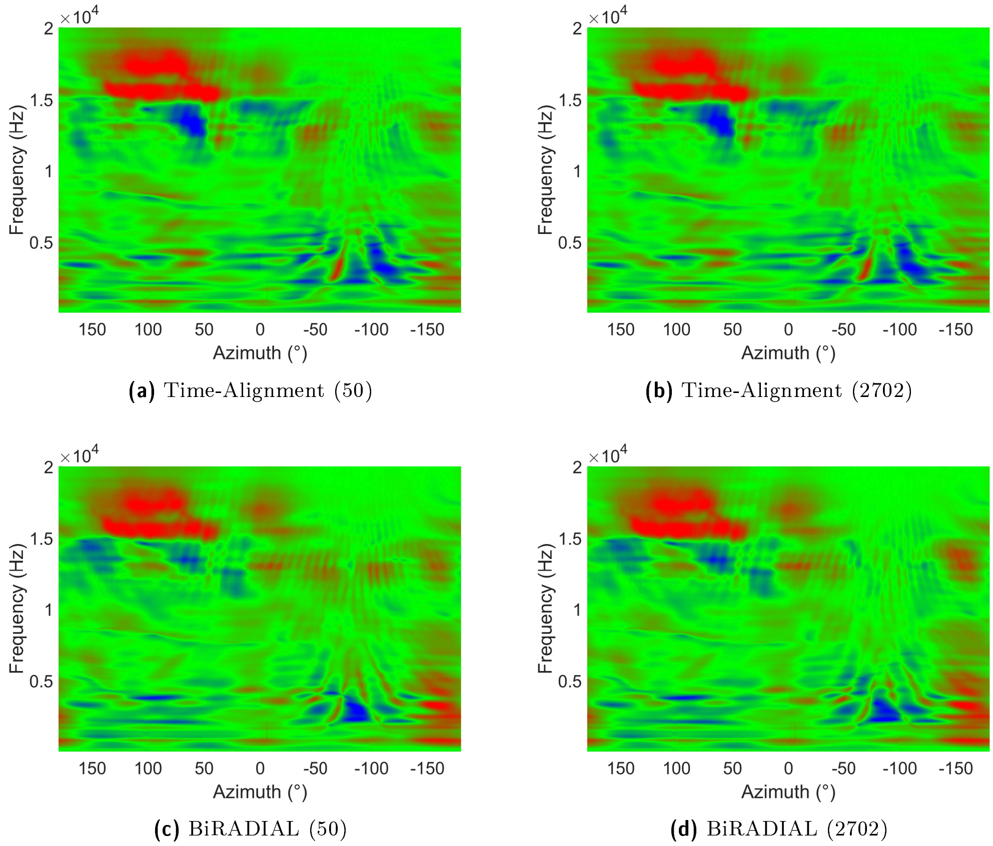


(a) ITD **(b)** ILD

Figure 6.40: Comparing ITD and ILD estimations for BiRADIAL intermediate generation and re-synthesis of far-field HRTFs using 50 and 2702 points

When comparing the intermediate far-field representation of 50 points vs 2702 points there is no difference in the response of the Time-Alignment method. This is believed to be because the synthesised HRTFs are derived from the same original set of 50 filters by the same method (Time-Alignment). However, by implementing a BiRADIAL approach and synthesising 2702 HRTFs instead of 50 a slight increase in the reproduced ILDs is seen. ITD remains the same, see Fig. 6.40. Although not necessarily any more or less accurate, this change does help to focus and externalise lateral sources to either side of the head. The spectral response is also very similar, both cases are shown in Figure 6.41. To demonstrate directly the similarities between the results of SADIIE and MARC systems intensity plots of the HRTFs generated from each are presented in figure 6.42. The average PSDM error in this case is 1.39 Sones which may be compared to values presented in Table 5.8 for a sense of scale. However, the reader is reminded that in this case comparisons are being made between HRTFs measured within two completely separate environments and so although the absolute value may at first appear large, the fact it is even of similar order to those presented in Table 5.8 is an impressive result.

Directly comparing modelling a loudspeaker radii of 1.2m and 100m, it is noted that the 100m reproduction has a slightly brighter timbre to the 1.2m reproduction (regardless of the method used - Time-Alignment, BiRADIAL, 50/2702 points). This is particularly true for sources that approach a $-90°$ elevation. At 1.2m sources tend to sound a little boomy as they pass underneath the listener. Whilst in a sense this may invoke a clear sense of direction, the dramatic change in timbre is less convincing and the 100m model is far more consistent. There are minor changes in the reproduction of ITD and ILD but these are not significant. This is shown for a BiRADIAL intermediate far-field representation renderer with 2702 points in Fig. 6.43. Spectral comparisons are similarly shown in Figure 6.44 revealing very



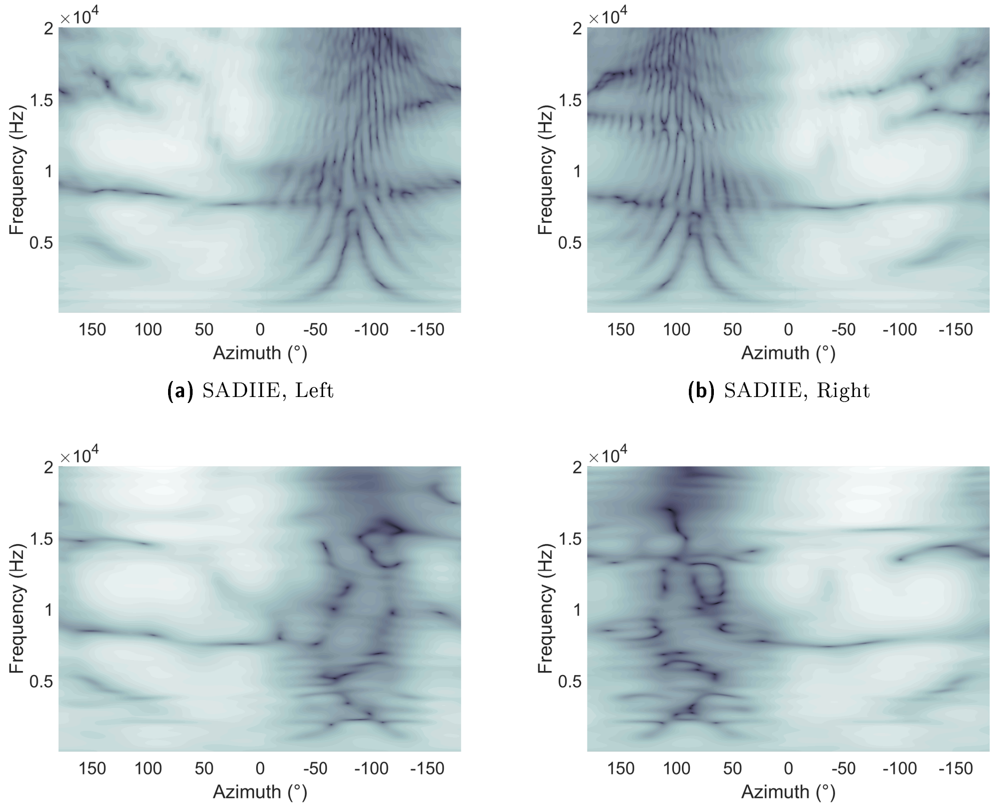
(a) Time-Alignment (50)

(b) Time-Alignment (2702)

(c) BiRADIAL (50)

(d) BiRADIAL (2702)

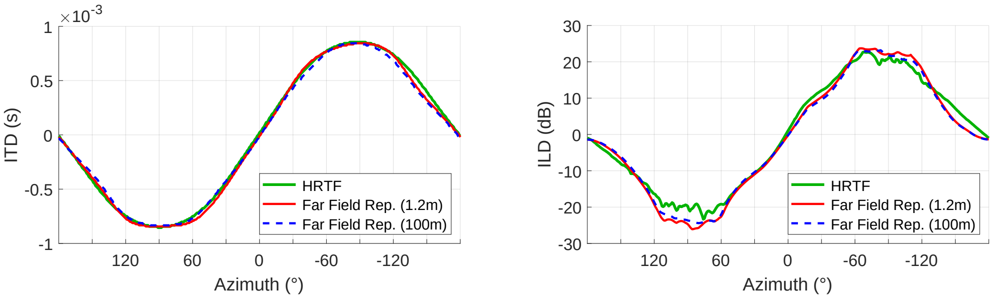
Figure 6.41: Comparison of the PSDM (left channel only) of the intermediate far-field representation of 50 points vs 2702 points for Time-Alignment and BiRADIAL rendering of HRTFs within the MARC system. Comparisons are made with respect to SADIIE HRTFs. Note how the plots of the Time-Alignment renderer are practically identical irrespective of the number of points rendered whilst the responses of the BiRADIAL are very similar (subtle differences can be seen on the contralateral/right hand side of the plot).



(a) SADIIE, Left **(b)** SADIIE, Right

(c) MARC, Left **(d)** MARC, right

Figure 6.42: Directly comparing HRTF intensity plots from the SADIIE database and BiRADIAL intermediate generation and re-synthesis of far-field HRTFs using 2702 points from MARC. Amplitude is shown by colour. Black: -60dB, White: +20dB. Note the reconstruction of similar peaks and notches in each case.



(a) ITD **(b)** ILD

Figure 6.43: Comparing ITD and ILD estimations for BiRADIAL intermediate generation and re-synthesis of far-field HRTFs 2702 points at 1.2 and 100m

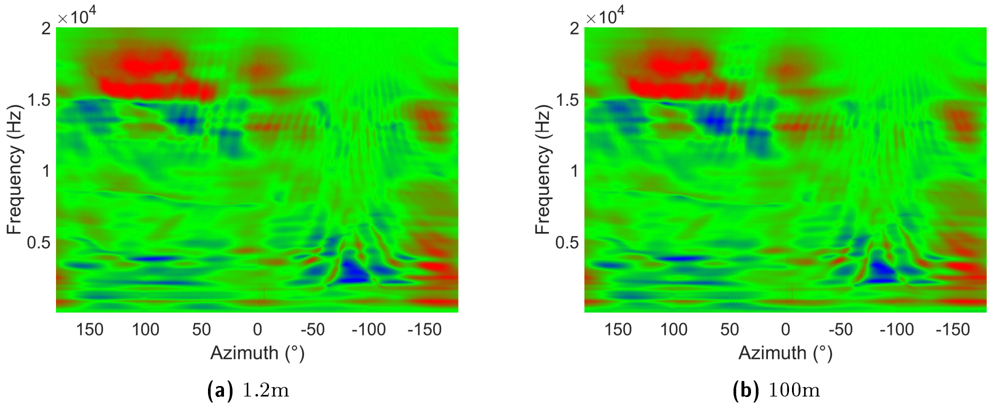


Figure 6.44: Comparison of the PSDM (left channel only) of BiRADIAL intermediate generation and re-synthesis of far-field HRTFs 2702 points at 1.2 and 100m. Comparisons are made with respect to SADIIE HRTFs.

few, if any, significant discrepancies.

Considering these findings it is recommended to use BiRADIAL methods to synthesise a high number of intermediate far-field HRTFs and then use these synthesised HRTFs within a Time-Alignment (with delay crossover ≈ 2000Hz) dual-band (basic, max-$\boldsymbol{r}_E$, 15kHz crossover) spherical harmonic decoder. If the BiRADIAL method is unavailable then a Time-Alignment method should be used to synthesise the far-field HRTFs. A high number of far-field HRTFs may still be synthesised if required for a particular application, but they are no longer beneficial to an Ambisonic decoder.

6.6 Summary

A compact HRTF measurement chamber, known as MARC, has been presented and discussed with regards to future commercial needs and applications. Methods, workflows and considerations have been suggested for the fast capture of near-field HRTFs and the subsequent accurate synthesis of infinite far-field measurements using the Time-Alignment and BiRADIAL binaural Ambisonic rendering approaches outlined in Chapter 5. Objective analysis has identified the overall benefit, alongside drawbacks, of applying multiple inconsistent NFC filters to individual HRTFs (prior to summation in the spherical harmonic domain) in order to compensate for the variable radii of the measurement rig.

MARC provides a solution to the need for convenient and rapid capture of individual HRTFs for a wide audience. The application of such a system ranges from research and development through to the amateur and professional computer games markets. Objective analysis in Section 6.5 has revealed that the synthesised HRTFs are a good match to those measured as part of the SADIIE database. In Chapter 7 subjective evaluations will go on to confirm these results and justify the use of MARC in such scenarios.

7

Assessing the Subjective Performance of MARC

Chapter Overview

This chapter consolidates the work of this thesis with a perceptual listening test in which Time-Alignment and BiRADIAL rendering methods are compared utilizing HRTFs from both MARC and SADIIE. Individual and non-individual measurements are also compared with respect to both localisation and timbral qualities. Results show that individual HRTFs provide the best localisational performance whilst BiRADIAL rendering with KU100 HRTFs measured in MARC gives the best timbral performance.

7.1 Introduction

Objective comparisons in Chapter 6 confirmed that HRTF measurements from MARC were of similar composition to those from the SADIIE database. However, in order to fully evaluate the HRTF measurement and optimization processes developed within this thesis a final listening test was proposed to consider the equivalent subjective results and validate the findings of Chapters 3, 5 and 6.

The test serves two purposes. Firstly, to evaluate the performance of MARC and to ascertain whether or not this new near-field fast-capture technique is perceptually comparable to the more traditional HRTF measurement procedures, for example, the approach taken for the SADIIE database. Secondly, it would provide the opportunity to directly compare the Time-Alignment (Zaunschirm, Schörkhuber and Höldrich, 2018) and BiRADIAL optimization strategies for near-field HRTFs.

A collection of binaural renderers were constructed using SADIIE and MARC HRTFs and using the techniques discussed in Sections 4.9.3 and 6.5. Individual HRTFs were again compared to generic sets as in the perceptual listening test of the SADIIE database, presented in Section 3.4. This time, both the localisation and timbral performances of the renderers were considered.

Results from the SADIIE test, presented in Section 3.4.4, had already shown that individual measurements were not necessarily the most preferred generally. However, it was important to identify whether improvements in localisation (and indeed externalisation) commonly attributed to individual measurements (Begault et al., 2000) would be preserved given the *close-to-the-head* nature of the near-field capture technique. Further, these tests would help to provide a better understanding of the performance differences between the BiRADIAL and Time-Alignment rendering techniques.

7.2 Listening Test

7.2.1 Overview

A listening test was conducted to investigate the performance of generic and individualised binaural Ambisonic renderers implementing Time-Alignment and BiRADIAL optimization strategies. 10 participants, all experienced expert listeners, were recruited for the test. All subjects gave their informed consent for inclusion before

Table 7.1: Summary of the HRTF renderers being compared. TA = Time Alignment.

	1	2	3	4	5
Source	SADIIE	MARC	MARC	MARC	MARC
Subject	KU100	KU100	KU100	Individual	Individual
Rendering Technique	TA	TA	BiRADIAL	TA	BiRADIAL
Number of HRTFs	50	2702	2702	2702	2702

they participated in the study. The protocol was approved by the University of York Physical Sciences Ethics Committee.

Five binaural Ambisonic renderers were compared to assess their performance both in terms of localization accuracy and timbral qualities. The binaural renderers were devised from a selection of HRTFs encoded into 5th order spherical harmonic format with pseudo-inverse decoding weights, as described in Section 4.9.3. Generic HRTFs from the SADIIE database (KU100 dummy head) were compared against both generic and individual HRTFs measured in MARC. A summary of the renderers is given in Table 7.1.

In the case of the SADIIE renderer a 50 point Lebedev grid distribution of HRTFs were encoded into spherical harmonic format. This was the highest resolution quasi-regular distribution of HRTFs measured as part of this database and represents the state-of-the-art or virtual loudspeaker rendering. In the case of the MARC renderers a 2702 point Lebedev grid distribution of HRTFs were encoded into spherical harmonic format having been synthesised from 50 near-field measurements. This is the best objectively performing format of binaural renderer currently available as part of the MARC measurement system, as shown in Section 6.5.4, and represents the state-of-the-art of the near-field measurements.

Each renderer was used to binauralise a selection of 5th order Ambisonic material over of a pair of sennheiser HD650 headphones. 3D head-tracking was implemented using Motive software[27] (Version 1.8.0 Final 64-bit). 6 Optitrack Flex-3 Infra-Red motion capture cameras tracked a 5-point array of reflective markers secured to the top of the headphones, shown in Fig. 7.1

Headphone equalisation filters were not implemented for this test. This is justified for a number of reasons. Firstly, the addition of linear phase filters has the effect of smearing the original signal in the time domain. Whilst it is generally agreed that

[27] optitrack.com/products/motive/tracker/

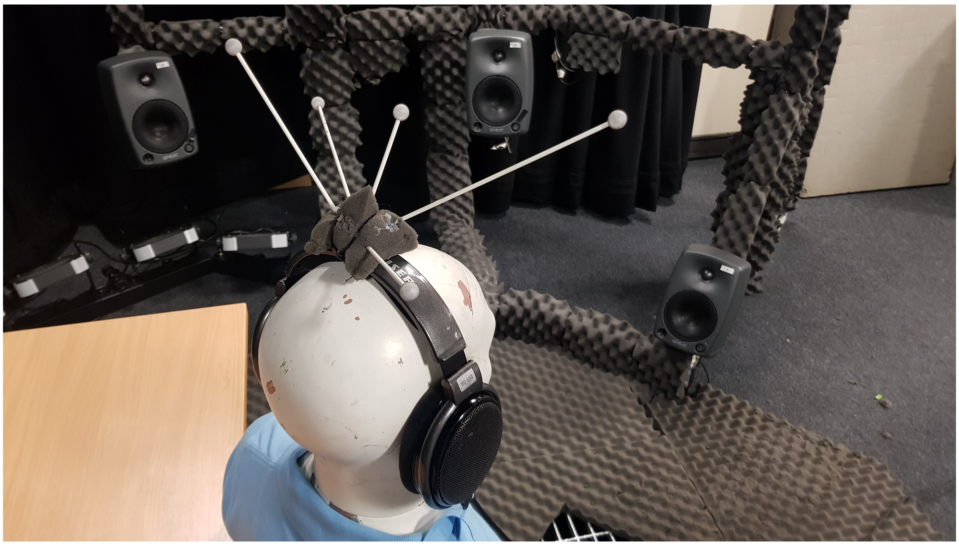
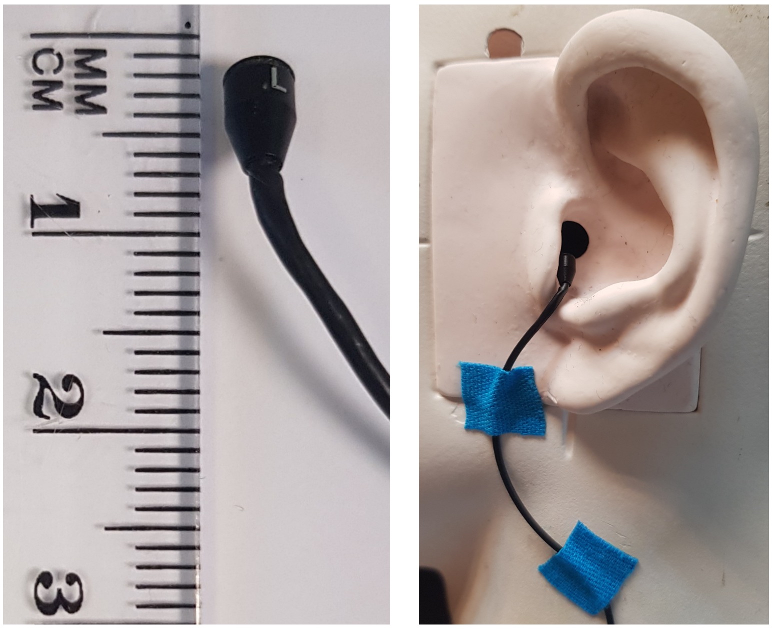


Figure 7.1: 5-point array of reflective markers secured to sennheiser HD650 headphones for the purpose of head-tracking

this is a small price to pay for a flat EQ (Schärer and Lindau, 2009) it is unclear exactly what other effects this may have on a binaural renderer. For example, the filters would need to be re-windowed after filtering in order to preserve their original tap length. Previous work in Section 3.3.3 has indicated how even minor differences in windowing technique can effect the resulting free-field and diffuse-field response of the HRTFs.

Work by Adams and Boland, 2010 also illustrates the variance in headphone equalisation with repeated seatings/measurements. Variation of several dB is evident throughout the magnitude curve of the HD650 headphones. Falsely accounting for such a shift in the high frequency amplitude could easily have an effect on the resultant ILDs of the signals and therefore the perceived lateralisation. As the following tests will assess potentially very minor differences in the binaural outputs of these renderers it is important not to introduce any other possible sources of HRTF distortion. Further, as only the differences between the renderers are being assessed, it is acceptable that the natural headphone EQ is applied to each stimuli in the same way.

In a practical sense, it is also of interest to assess the feasibility and performance of the renderers in a real-world scenario. Whilst it is one thing to imagine a future in which people will each have a copy of their own HRTFs, it is quite another to imagine a future in which they will also have a headphone equalisation filter for each and every pair of headphones they may ever buy. Although some software packages



(a) Loudspeaker mounter perpendicular to frame

(b) Loudspeaker mounted at angle to frame

Figure 7.2: Voice Technologies VT202 stereo lavalier microphone scale and placement behind tragus for HRTF measurement in MARC.

may include example filters, they will not be individualised to the user. High quality open back reference headphones are therefore utilized for the test and the HRTFs under test are left unedited.

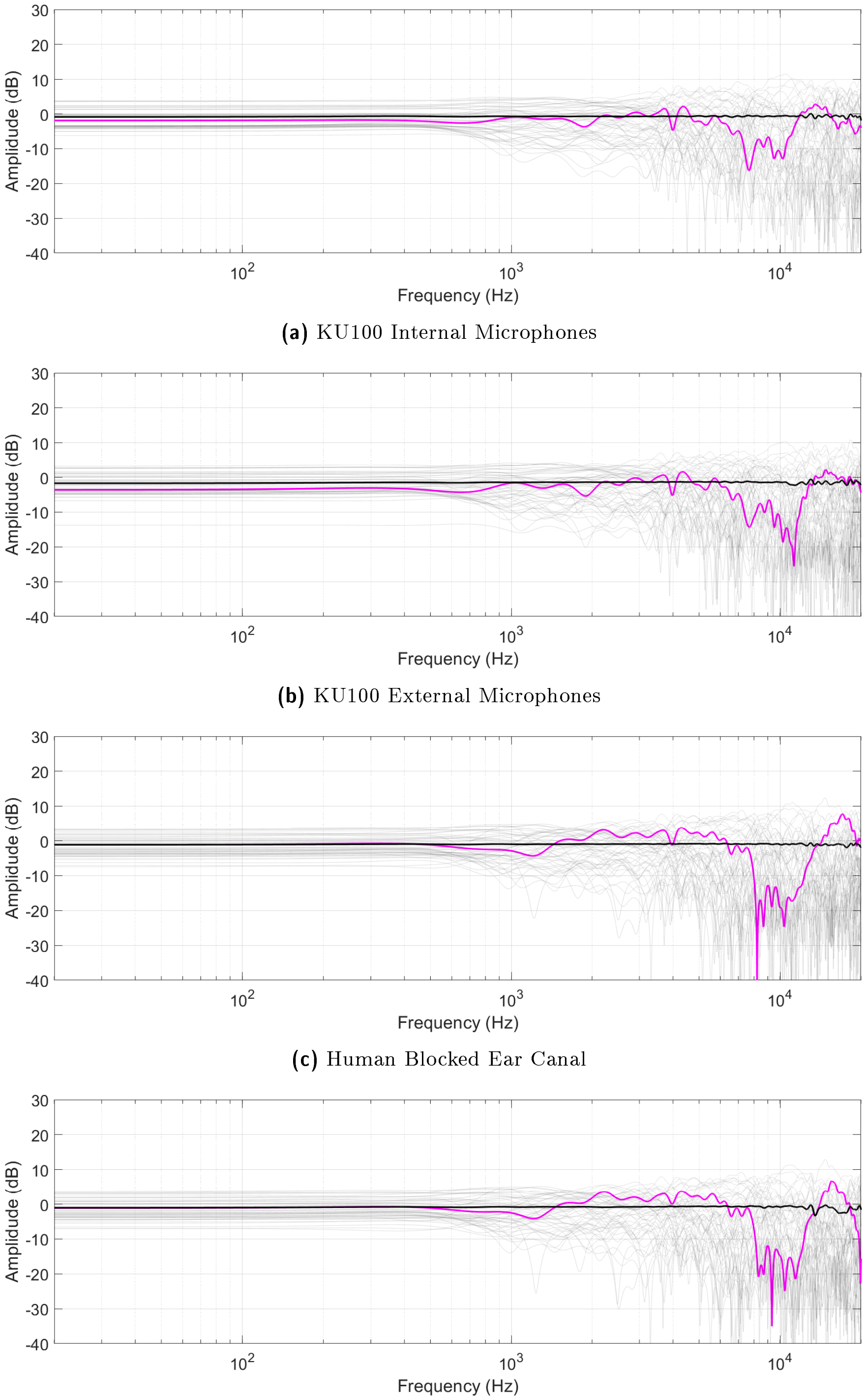
7.2.2 HRTF Acquisition

Each participant had their individual HRTFs measured within MARC. This was done using an open ear canal approach with Voice Technologies VT202 stereo lavalier microphones[28] placed just behind the tragus and secured in place with tape, see Fig. 7.2. Measurements of the KU100 dummy head were were also taken within MARC, but utilised the devices own internal microphones.

Informal listening tests by at least 2 expert listeners and visual comparisons of frequency spectra revealed no perceptually relevant differences between the resultant HRTFs of:

1. The KU100 with its internal microphones vs the KU100 with the lavalier microphones

2. A human subject with the lavalier microphones and an open ear canal vs the same subject with a blocked ear canal (Zoggs Silicone Swim Ear Plugs)



(a) KU100 Internal Microphones

(b) KU100 External Microphones

(c) Human Blocked Ear Canal

(d) Human Open Ear Canal

Figure 7.3: Spectral plots of the KU100 and human HRTFs measured in MARC with internal/external microphones and a blocked/open ear canal respectively. Diffuse-field equalised HRTFs are shown in grey. The HRTF of angle ($\varphi = 0°, \vartheta = 0°$) is highlighted in purple. The diffuse-field response of the diffuse-field equalised HRTFs is shown in black.

Table 7.2: Average (ERB weighted) PSDM values for the comparison of binaural renderings using the HRTFs of the KU100 measured with internal/external microphones and a human with a blocked/open ear canal. Values are given for comparisons made over the sphere and horizontal axis.

		Sphere	Horizontal Axis
KU100 (int. vs ext.)	Left	0.60	0.57
	Right	0.69	0.65
Human (Blocked vs Open)	Left	0.37	0.38
	Right	0.40	0.40

Example spectral plots of the two subjects are given in Fig. 7.3 Similarities between the equalised diffuse-field responses, and example spectra (highlighted for the angle $\varphi = 0°, \vartheta = 0°$) can be seen in each case.

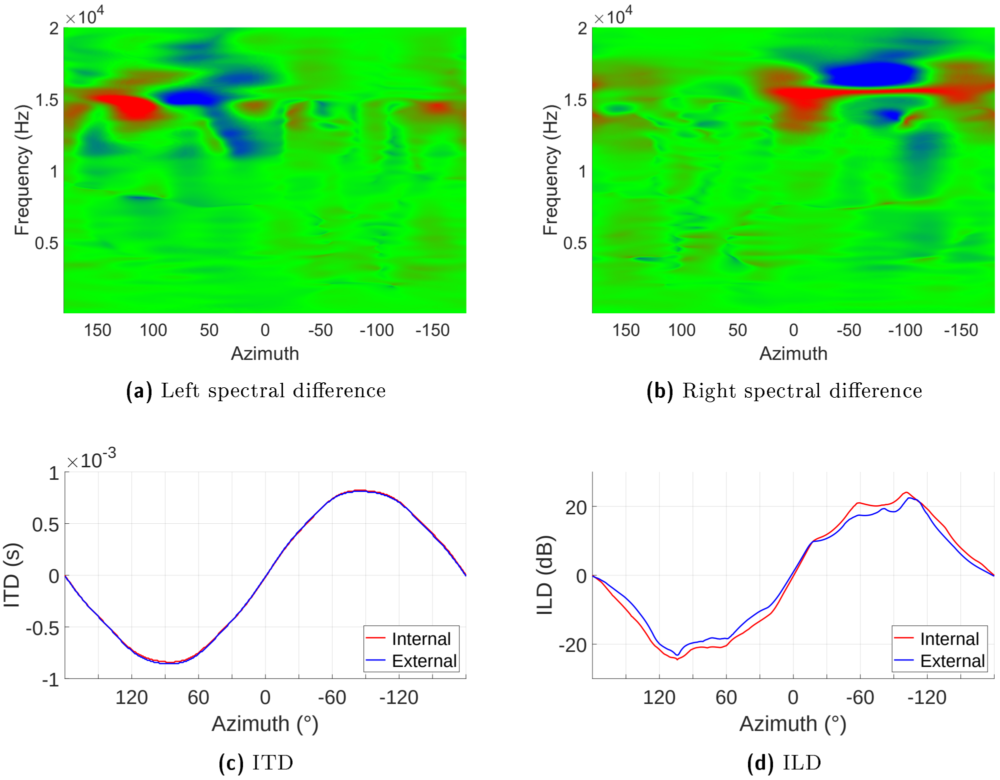
These results were assured using the PSDM. The HRTFs in question were used to synthesise 2702 far-field measurements which were encoded into 5^{th} order tri-band (time aligned at 2100Hz and max-$\boldsymbol{r}_E$ at 15kHz) Time-Alignment Ambisonic decoders. Binaural renderings were then compared over the sphere (5° azimuth/elevation resolution) and about the horizontal axis (1° resolution). Average (ERB weighted) PSDM values are presented in Table 7.2. Perceptual spectral difference plots, ITD and ILD plots are also given for the KU100 in Fig. 7.4 and for the human in Fig. 7.5.

The PSDM values are low. The reader is referred back to Fig. 5.24 for a sense of scale and and is reminded that at a listening level of 75dBSPL a difference of 1dB (1JND) equates to 0.8 Sones. Further the results show no differences in ITD and only minor differences in ILD as a result of spectral changes (primarily) above 10kHz only. Within this region of the frequency spectrum it is common to see discrepancies between multiple HRTF measurements regardless due only to the fact that the subject has been reseated in between recordings (Andreopoulou, Begault and Katz, 2015).

This analysis therefore justifies both the use of the lavalier microphones as well as the open ear canal approach. It is thought that the multiple stages of free-field and diffuse-field equalisation prior to and post-synthesis of the intermediate far-field HRTFs are the reason for the consistent output across technique.

Following the workflow defined in Section 6.4.2, playback of sine sweeps and binaural microphone recordings were made via a 96 channel in/out MOTU PCI 424 soundcard

[28] vt-switzerland.com/en/ (product no longer available.)



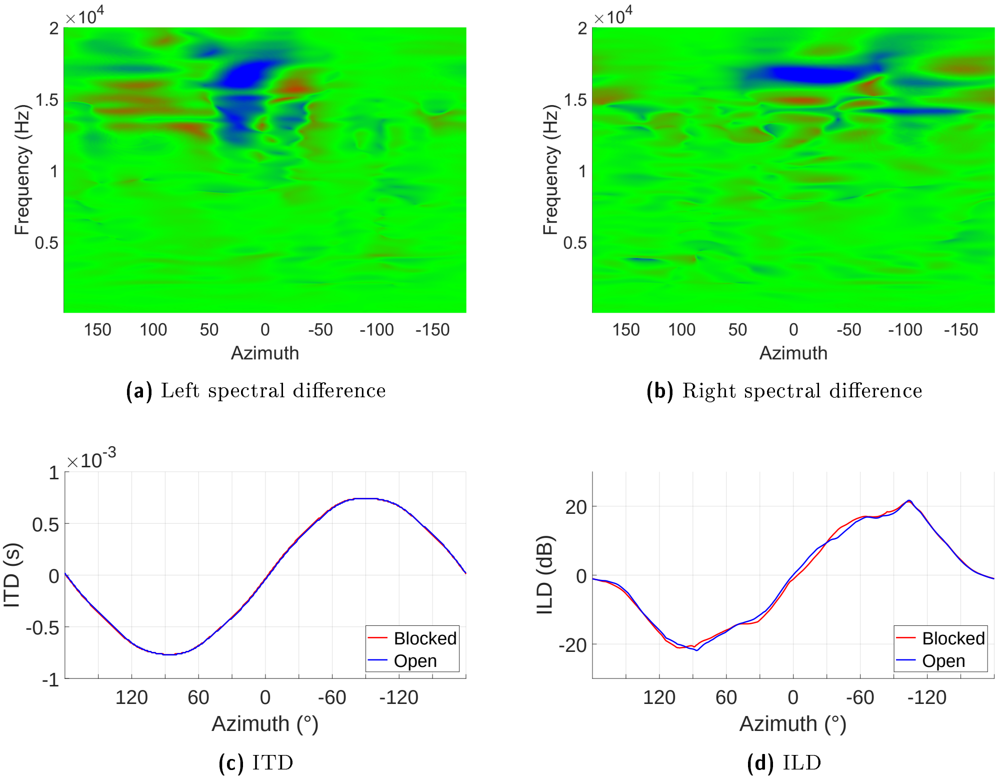
(a) Left spectral difference

(b) Right spectral difference

(c) ITD

(d) ILD

Figure 7.4: PSDM results for the KU100 comparing binaural renderings made with HRTFs measured with the dummy head's internal microphones vs external lavalier microphones.



(a) Left spectral difference

(b) Right spectral difference

(c) ITD

(d) ILD

Figure 7.5: PSDM results for a human comparing binaural renderings made with HRTFs measured with a blocked vs open ear canal and lavalier microphones.

Table 7.3: The extraction window as used in the MARC HRTF measurement system.

Opening (half-Hanning)	Pre-Peak Pad	Post-Peak Pad	Closing
50	100	75	25

Table 7.4: The final windowing parameters as used in the MARC HRTF measurement system.

Opening (half-hanning)	Pre-Peak Pad	Post-Peak Pad	Closing
40	30	65	20

via MOTU 2408 interfaces at 48kHz sampling rate and 2048 sample frame size. The binaural microphones were input to the interface via an IMG Stageline MPA-202 2-channel stepped preamplifier. 50 2.5s exponential sine sweeps were output from the 50 loudspeakers over the range 250-22000Hz with peak level -3dBFS. Free-field equalisation filters were employed to flatten the response of the loudspeakers ($\pm \approx$ 15dB) down to approximately 350Hz. The sweeps were output from individual loudspeakers with a time delay of 0.08s between each sweep interleaving the IRs to be extracted with the overlapped exponential swept sine wave technique (Majdak, Balazs and Laback, 2007). This was a similar workflow to that implemented in the measurements of the SADIIE database.

Individual loudspeaker levels were normalised with respect to the average RMS amplitude of their free-field measurements. HRTFs were extracted from the raw, deconvolved file with a 4-part window as described in Table 7.3. Further free-field equalisation ($\pm \approx$ 3dB in the mid-high frequency range with a $+ \approx$ 15dB boost in the low frequencies 350Hz-200Hz) was applied to each HRTF to account for the remaining frequency responses of the loudspeakers and also the binaural microphones. A low frequency model was applied to each measurement with a crossover frequency of 475Hz.

Diffuse-field equalisation of the measurements was undertaken with 1/4 octave band smoothing. Maximum corrections of ± 20dB were imposed within the frequency range 20Hz-20kHz and ± 10dB outside of that range. The equalisation filter was truncated to 4096 samples with a Hanning window. The final (equalised) HRTFs filters were windowed as described in Table 7.4 and truncated as a set to 128 samples (48kHz). Note that this is half the length of the SADIIE filters. Is is beneficial to truncate the filters as short as possible (and to a power of 2) in order to reduce latency concerns in real-time systems. However, it was found that a reliable and

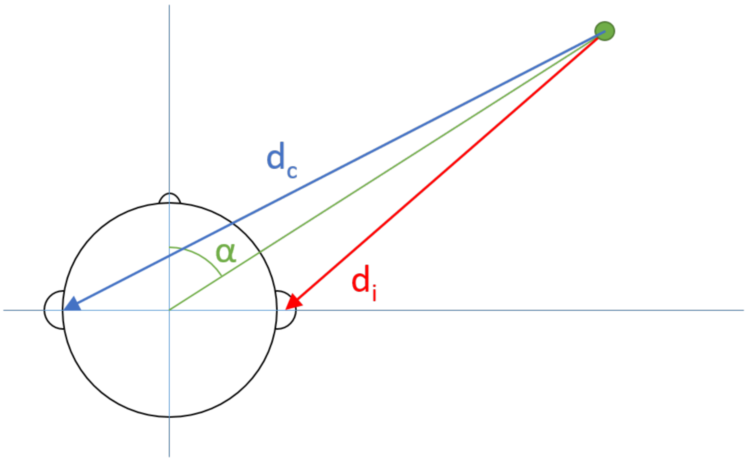


Figure 7.6: A simplified 2D distance-only model for approximating ITD.

accurate frequency response was unable to be maintained with a tap length of 64 samples or below.

Head centred and ear centred HRTFs were captured for each subject. The head centred HRTFs were used to synthesise far-field HRTFs using the Time-Alignment approach and vice versa for the ear centred HRTFs and a BiRADIAL approach. For the Time-Alignment approach, a spherical head model with head radius 8cm was utilised to approximate the ITDs of the HRTFs and align the measurements. Similarly, a spherical head model was used to re-introduce the same time delays into the far-field HRTFs. In the BiRADIAL approach a distance-only based model, demonstrated in Fig. 7.6, was used to re-introduce the ITD into the far-field HRTFs. This is because there already exists a partial ITD within the BiRADIAL measurements due to the physical placement of the head interrupting the direct line of sight between the ear being measured and the contralateral loudspeakers.

Once the far-field HRTFs were synthesised, Hybrid HRTFs were generated using a Time-Alignment approach. This is because for large radii the perceptual differences between a Time-Alignment and BiRADIAL renderer are negligible. This is shown through the similarity of laterally spaced HRTFs beyond a distance of 10m in Section 6.2. A crossover frequency of 2100Hz was chosen to implement the Time-Alignment. Gradual Group delay filters were implemented from approximately 1800Hz to smooth the transition between the varying delays. A max-$\boldsymbol{r}_E$ weighting scheme was applied at 15kHz.

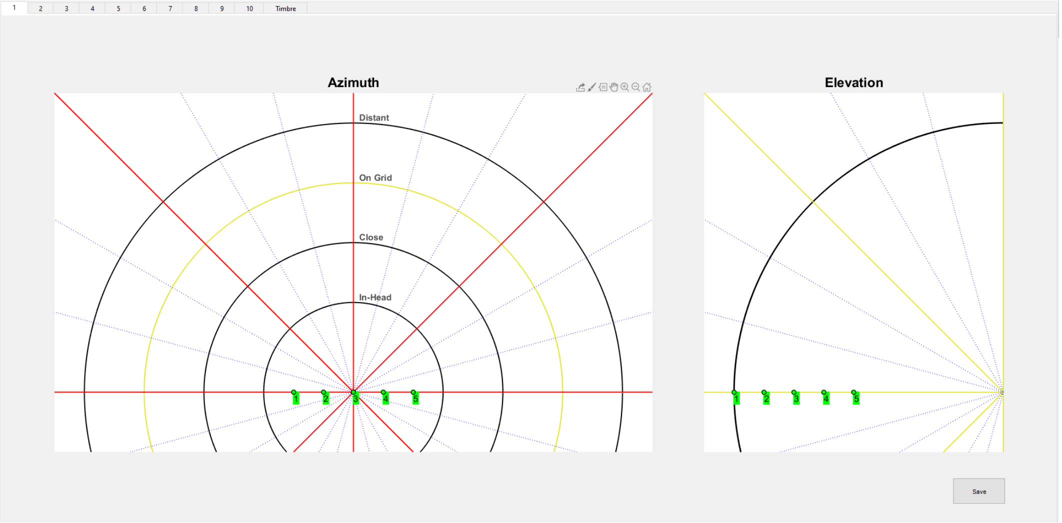


Figure 7.7: Matlab GUI for inputting perceived location of sources. Subject is asked to input azimuthal and radial position on the left, and elevation on the right. Radial positions are labelled: 'In-Head', 'Close', 'On Grid' and 'Distant'.

7.2.3 Localisation

The localisation performance of each renderer was tested by synthesising a series of 5 white noise point sources at locations about the sphere and asking the participant to mark the perceived location (azimuth, elevation and radial distance) of each source on a GUI generated in Matlab, see Fig. 7.7. It is a similar approach to that by Gilkey et al., 1995 and Begault et al., 2000 with the inclusion of a physical quarter-sphere reference grid with markings (red, yellow, blue) corresponding to the GUI, see Fig. 7.8. The grid consisted 15° intervals between −105° and 105° azimuth and −15° and 90° elevation. 45° azimuthal intervals were highlighted in red and 45° elevation intervals were highlighted in yellow. The purpose of the grid was to help ensure that the angle being recorded was truly the correct angle being perceived and that a participant's results were not being skewed by any internal bias. To that end, the source locations being tested were restricted to those belonging within the confines of the grid.

Alternative approaches may have included an egocentric pointing method, e.g. Bahu et al., 2016, where a participant is asked to physically point towards a location with their hand/a wand. However, this is a somewhat more complex procedure to measure and relies on good hand-eye coordination from the subject. Another common method is the method of adjustment (Cardozo, 1965; Thresh, Armstrong and Kearney, 2017; Rudzki et al., 2019). However, such a technique would not be appropriate in this case. Firstly, the method requires a reference signal that

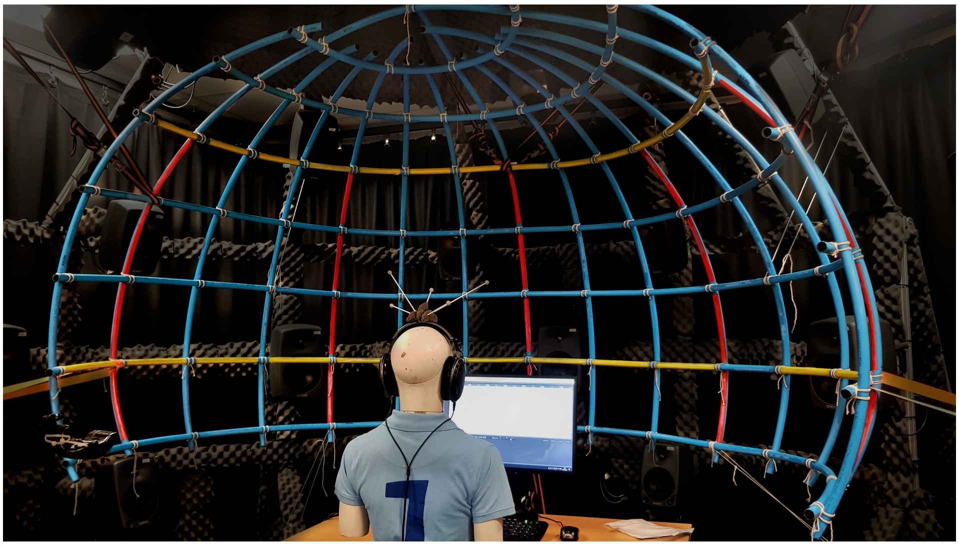


Figure 7.8: The colour coded reference grid that surrounded the participant. Red and yellow highlights define 45° increments about the azimuth and elevation.

Table 7.5: The source locations tested for each HRTF renderer.

	1	2	3	4	5
Azimuth (°)	0	45	90	0	-45
Elevation (°)	0	0	0	45	45

would have been impossible to provide over headphones and be sure of its perceived location. Secondly, the results can be skewed by a participant attempting to match the timbre of the test signal to a reference signal rather than its location. Finally, the method tests a participant's ability to match the location of a test signal to a reference signal rather than provide any indication as to the true perceived location of either source.

In order to properly assess the localisation performance of the renderers for differently panned sources participants were instructed to remain facing in a forward direction whilst listening to the stimuli. Head-tracking was employed in an attempt to help resolve general front/back confusions with minor head movements, however, participants were told not to turn to face a source. They were, however, permitted to turn their head in between playback in order to locate a source upon the reference grid.

The true locations of the 5 sources put under test are shown in Table 7.5. For each renderer the 5 sources were presented sequentially as a series of pulse trains. Each source was presented twice (back to back) and each presentation consisted of 3 groups of 7 pulses, as shown in Fig. 7.9. The order of the sources could not be

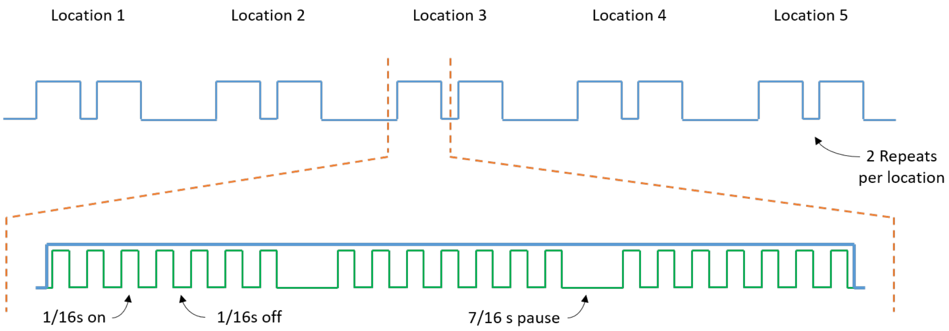


Figure 7.9: Pulse train sequence for testing the localisation performance of each renderer

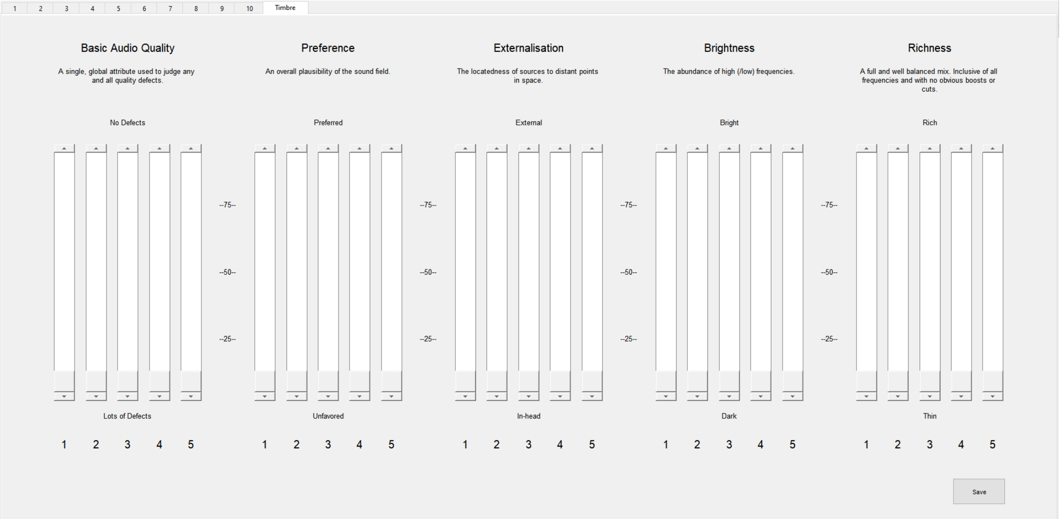
randomised as they were predefined in a single audio track within a Digital Audio Workstation (DAW) workflow, however, the order of the renderers was randomised for each participant. This was deemed acceptable as it was not the general location of each source that was under test, rather the perceptual misalignments of the different renderers. Each renderer was tested twice for a total of 10 trials.

7.2.4 Timbre

The timbral performance of each renderer was tested in a reference-free MUSHRA style test. It was impossible to present a reliable reference stimuli as this would have corresponded to a real world source. In order to assess this reference participants would have had to either repeatedly take on/off their headphones (requiring a head tracker re-alignment each time) or complex headphone compensation filters would have been required to negate the shadowing effects of the hardware. Variations in the high frequency details of such a filter could have then led to discrepencies in the results that would not have otherwise been present. Further, any such filter would have required consideration of its variation for multiple source locations and would have required adaptation with head-tracking.

Instead, participants were simply asked to rate the renderers directly against each other and against their own internal sense of preference. This was done via a second Matlab GUI, shown in Fig. 7.10. It is the same approach as was taken in the SADIIE listening test presented in Section 3.4. The only difference being, in order to familiarise the participant with the basic virtual rendering set up, they were played an example of a pair of real world loudspeakers (located in the same positions as the virtual ones would be) for basic reference before beginning this part of the test.

Participants were asked to rate each renderer on 5 scales: Preference, Externalisation, Richness, Brightness and Basic Audio Quality. The first 4 of these scales were



Basic Audio Quality

A single, global attribute used to judge any and all quality defects.

Preference

An overall plausibility of the sound field.

Externalisation

The locatedness of sources to distant points in space.

Brightness

The abundance of high (/low) frequencies.

Richness

A full and well balanced mix. Inclusive of all frequencies and with no obvious boosts or cuts.

Figure 7.10: Matlab GUI for inputting timbral preferences. Each renderer (1-5) is rated on each of the 5 attributes.

previously justified in Section 3.4.2. The latter, Basic Audio Quality, was included after it was standardised for timbral comparisons of stimuli in ETSI TS 126 259 V15.0.0 (2018-10) (ETSI, 2018). It is defined as 'A single, global attribute used to judge any and all quality defects' (ITU, 2015) and the scale was labelled from *No Defects* to *Lots of Defects*. Whilst potentially similar to overall preference, it is interesting that participants did not always rate the two attributes the same. Although, the overall trends are similar (shown in Section 7.2.5).

Two popular stereo music tracks (All I Really Want[29] and IRIS[30]) were binaurally rendered over a pair of virtual loudspeakers located in front of the subject ±45° on the horizontal plane (left channel to left speaker, right channel to right speaker). During the test they were also given the opportunity to play a plain stereo representation of the two stimuli over their headphones. Although they were told not to compare the binaural renderers to the stereo representation directly, being able to hear the original tracks gave them the opportunity to familiarise themselves with the general timbre of the music.

The participants were asked to provide a single rating for each renderer after considering both tracks. From this point they were allowed to move their head and play, pause, select or loop any section of music they liked. They were encouraged to explore the full selection of music but doing so was left up to them. This section of the test had no repeats, however there was also no time limit to complete the ratings. They were free to listen to the tracks through any of the binaural renderers at

[29] Alanis Morrissette, Glen Ballard. In *Jagged Little Pill* by Alanis Morrissette, Maverick. 1996.
[30] John Rzeznik. In *Dizzy Up the Girl* by Goo Goo Dolls, Warner Bros. 1998.

any time and flick between them as they pleased. The order in which the renderers were presented was randomised for each participant. It was noted that participants tended to take their time on this part of the test, taking on average upwards of 10 minutes to complete their ratings.

7.2.5 Results

Throughout this section the different binaural renderers are identifiable by their colour (and the order in which they are presented) as detailed below:

Blue: SADIE KU100 Time-Alignment
Red: MARC KU100 Time-Alignment
Green: MARC KU100 BiRADIAL
Pink: MARC Personal Time-Alignment
Yellow: MARC Personal BiRADIAL

The complete set of localisation results are given in Fig. 7.11. The graphs plot the azimuth and elevation angles and externalisation ratings given by each subject for each of the 5 locations tested and for each of the 5 different renderers. Anomalies were identified using Tukey's Fences (Tukey, 1977) where a result, x, existed outside of the range

$$Q_1 - 1.5(Q_3 - Q_1) < x < Q_3 + 1.5(Q_3 - Q_1) \tag{7.1}$$

where Q_1 and Q_3 are the lower and upper quartiles respectively.

The same results (excluding anomalies) are replotted onto linear graphs throughout Figs. 7.12, 7.13 and 7.14. In each case the interquartile range is also plotted in the renderer's respective colour.

The average solid angle error of each renderer based on the participants' responses to the stimuli is plotted in Fig. 7.15. The error is calculated as the solid angle between the true source location, and the median perceived azimuth and elevation. Externalisation is not taken into account in this calculation.

Results of the timbral test are shown in Fig. 7.16. Individual responses from each participant to each attribute were normalized by mean and standard deviation as recommended by ITU-R BS.1284-1 (ITU, 2003a).

Due to the limited sample size, it would have been unfair and unreliable to test for normality. Non-parametric statistical analysis of the results was therefore undertaken in the form of a Kruskal-Wallis test. p-values are presented in Table 7.6 for

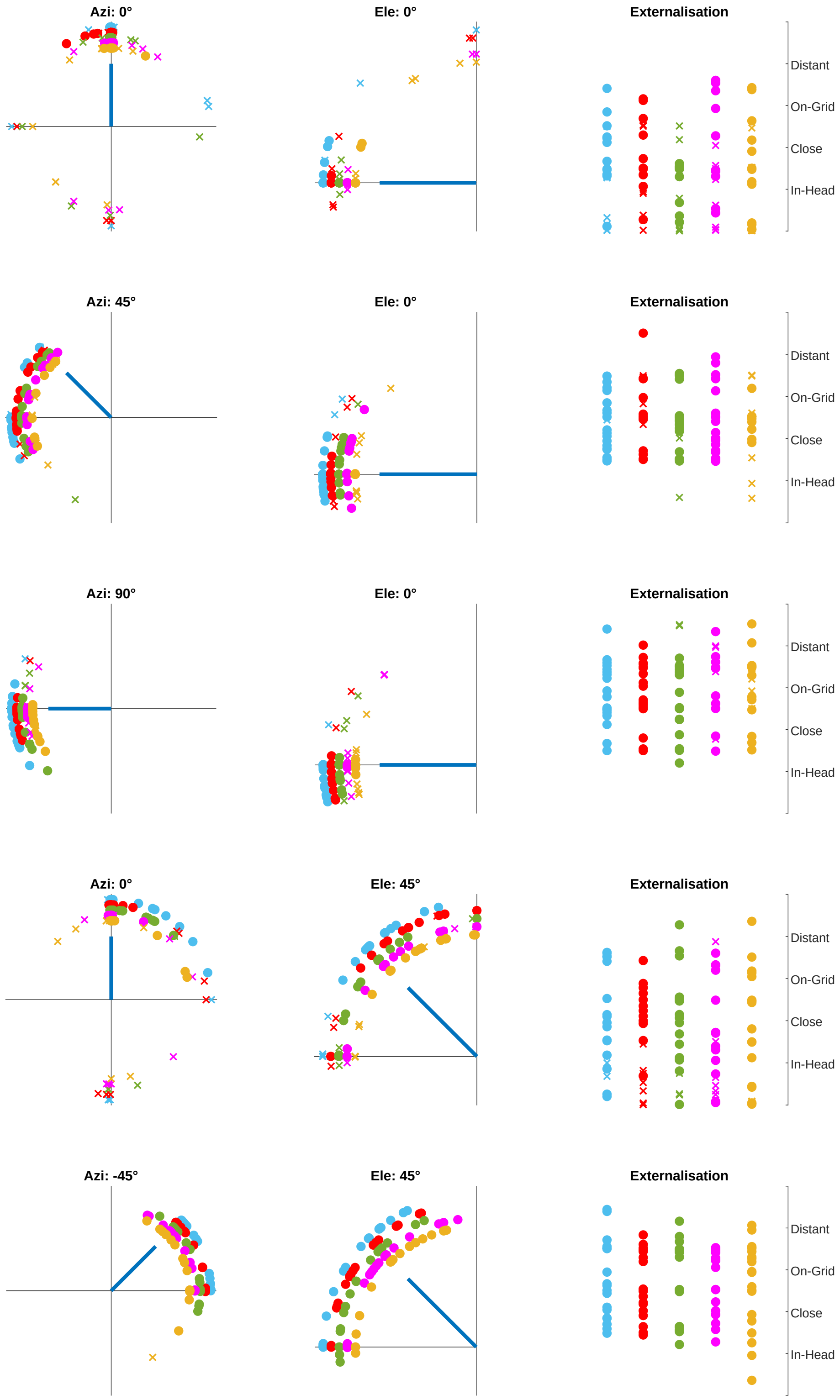


Figure 7.11: Complete localisation results. Each row represents a single source localisation. Correct azimuth and elevation are detailed in the titles of the plots and by a solid blue line from the origin. Externalisation results are shown on the right. Each of the 5 renderers are identified by colour (and radial/lateral position). Anomalies are shown as crosses.

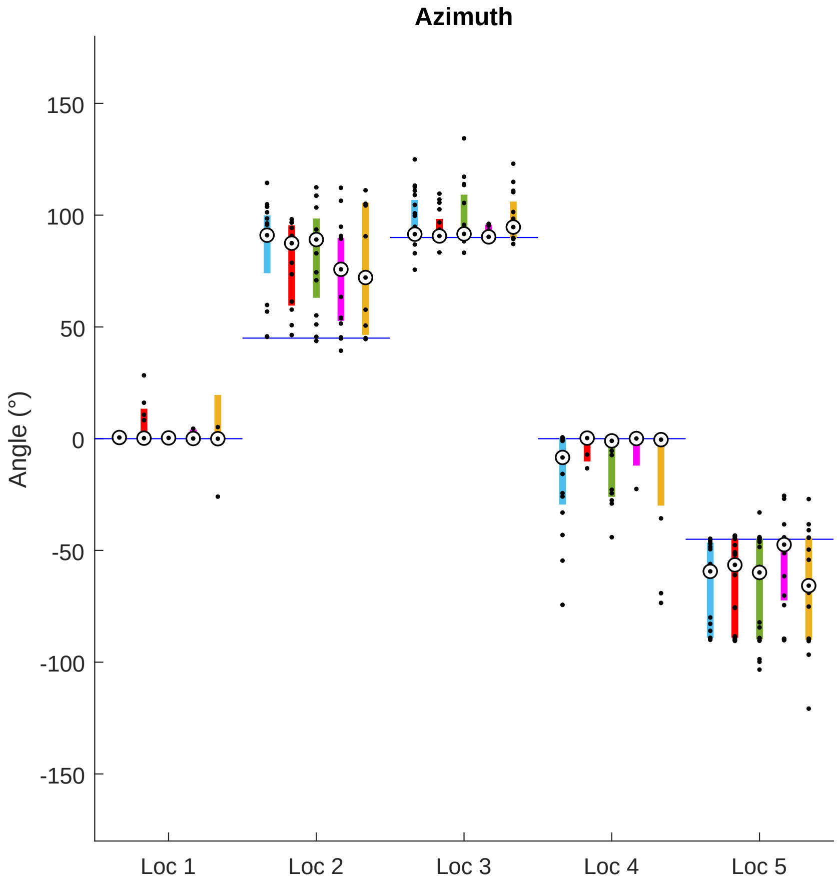


Figure 7.12: Azimuth localisation results. Source locations are grouped upon the y-axis. The correct location is depicted by a horizontal blue line. Median values are presented as a black dot within a solid white circle. Each of the 5 renderers are identified by colour (and lateral position).

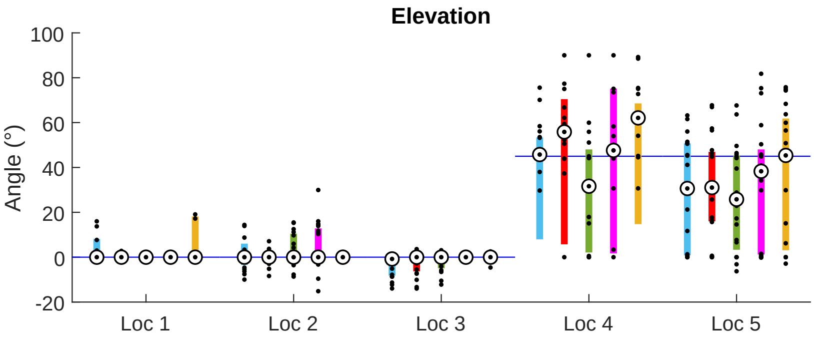


Figure 7.13: Elevation localisation results. Source locations are grouped upon the y-axis. The correct location is depicted by a horizontal blue line. Median values are presented as a black dot within a solid white circle. Each of the 5 renderers are identified by colour (and lateral position).

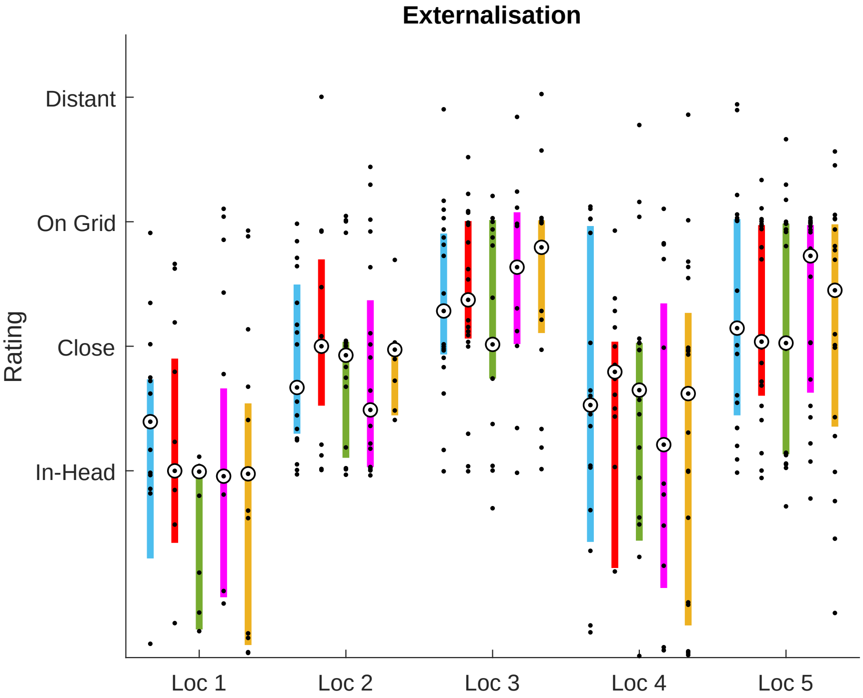


Figure 7.14: Externalisation localisation results. Source locations are grouped upon the y-axis. Median values are presented as a black dot within a solid white circle. Each of the 5 renderers are identified by colour (and lateral position).

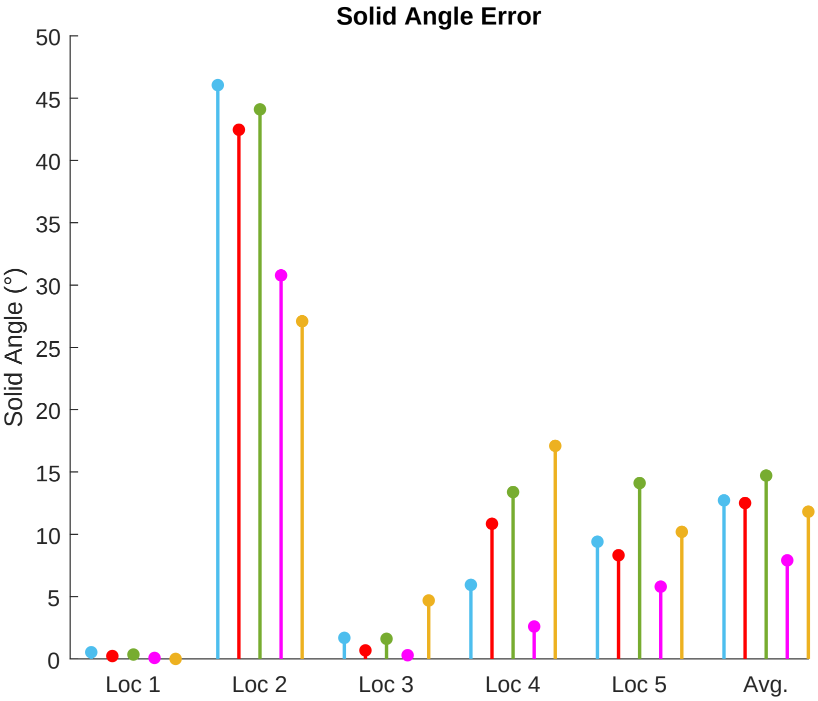


Figure 7.15: Average solid angle error of renderers. Errors for individual source locations are grouped upon the y-axis. An additional group to the right represents the mean of the previous 5 values. Each of the 5 renderers are identified by colour (and lateral position).

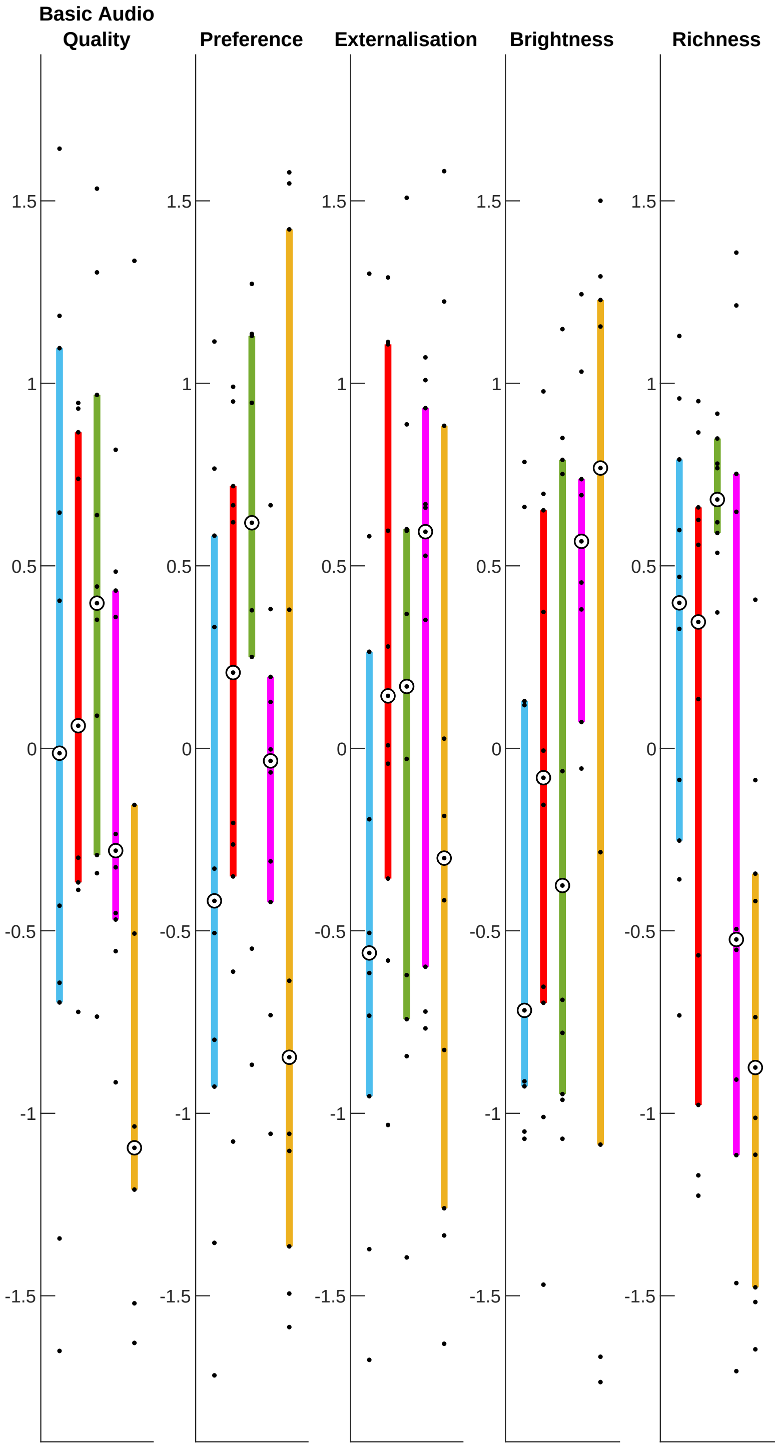


Figure 7.16: Timbral performance results. The results for each attribute are split by graph. Median values are presented as a black dot within a solid white circle. Each of the 5 renderers are identified by colour (and lateral position).

Table 7.6: p-values (%) output from a Kruskal-Wallis test regarding the data of perceived azimuth, elevation and externalisation across each of the 5 renderers for each of the 5 locations tested. A p-value below 5% is deemed significant (*).

	Loc 1	Loc 2	Loc 3	Loc 4	Loc 5
Azimuth (°)	31.42	19.25	86.05	3.66*	64.37
Elevation (°)	13.51	27.31	29.36	0.95*	48.00
Externalisation	48.86	80.48	67.28	77.42	96.37

Table 7.7: p-values (%) output from a Kruskal-Wallis test regarding the data of perceived timbral quality across each of the 5 renderers for each of the 5 timbral attributes tested. A p-value below 5% is deemed significant (*).

	Basic Audio Quality	Preference	Externalisation	Brightness	Richness
Timbral Quality	5.23	26.63	38.49	23.42	0.88*

the data regarding perceived azimuth, elevation and externalisation across each of the 5 renderers for each of the 5 locations tested. p-values are presented in Table 7.7 for the data regarding perceived timbral quality across each of the 5 renderers for each timbral attribute test. A p-value below 5% was deemed significant.

7.2.6 Discussion

The results presented in Section 7.2.5 are indicative of several independent conclusions. Many of these relate to the specific renderers under test, but others relate more generally to binaural reproduction as a whole.

Statistical Significance

Using a Kruskal-Wallis test only 3 comparisons indicated statistically significant differences in results, the reader is referred back to Tables 7.6 and 7.7. These cases were for perceived azimuth and elevation, Location 4, and perceived richness. Post-hoc analysis was undertaken in each case on the average group ranks based on Tukey's honest significant difference criterion at the 5% confidence interval.

During post-hoc analysis, no significant differences were found between the mean rank values in the case of perceived azimuth at Location 4. 1 significant difference was found in the case of perceived elevation at Location 4. This was found between the *KU100 HRTFs - BiRADIAL* and *Individual HRTFs - BiRADIAL* renderers. The individual renderer was found to have higher rank (indicative of greater elevation). 1 significant difference was also found between the same renderers in the case of the comparison of *Richness*. The KU100 renderer was found to have the higher rank in this instance.

General Findings

Firstly, consider the clear and obvious localisation errors identified as anomalous in Fig. 7.11. The majority of these cases can be attributed to a front back confusion of some sort but other examples, such as an increased perception of elevation for horizontal sources, are not so easily explained. It is unfortunate that these results are not completely avoided even with the implementation of head-tracking. However, such results are also present within the literature. For example, Begault found that although head-tracking will help to resolve some locational confusion it will not help in general with localisation accuracy (Begault et al., 2000). The reader should also recall that although head-tracking was implemented, participants were asked to remain facing in a forward direction and so it is more than likely that the full benefit was not realised.

A common trend is the over-lateralisation of sources, most clearly evident at locations 2 and 5 with an azimuthal component of ± 45°. These sources are almost exclusively perceived as closer to the interaural axis than they are in fact positioned. More interesting though, is the perceived externalisation of the sources in this region. Comparing the azimuthal and externalisation results of locations 2 and 3 it is noted that although the perceived location of the sources is actually somewhat similar (Fig. 7.12), the source positioned at 90° azimuth is perceived as more external across every renderer (Fig. 7.14). This is an important finding as it perhaps indicates a potential overlap between the perception of location and distance.

A potential bias in the experiment is also highlighted with regards location 4. It is unusual that the spread of results exists only to the right of the true source location. Despite accurate median values, shown in Fig. 7.12, that imply the majority of results fall close to the 0° angle, almost the entire interquartile range of every renderer protrudes counter-clockwise of these positions. As the order in which the sources were presented to the participant was not randomised, it is suggested that this result may then be a reaction to having previously heard two sources from the left hemisphere. Being presented with a frontal source after focusing for too long on the left hemisphere could therefore have led a participant to overcompensate and perceive the source as being further to the right than they would have done otherwise. This is another key finding as such a result implies a non-linearity in the ability to accurately localise a source and could have significant implications in, for example, source positioning in game design.

Individual HRTFs

In general, it is found that individual HRTFs improve the localisation accuracy of the renderers over generic HRTFs. This is shown in Fig. 7.15. The difference is most obvious for Location 2 where the generic HRTFs almost entirely collapsed the source image to the left hand side whereas the individual measurements maintained some level of frontal spatial positioning (see Fig. 7.12). There are also some cases of increased externalisation (locations 3 and 5) but these are not consistent.

However, there remains a dramatic reduction in timbral quality considering the individual measurements. Not only do they appear to be less preferred, but they have a lower Basic Audio Quality and are perceived as bright and thin. This was not predicted especially as the tests shown in Section 7.2.2 indicated no perceptual difference in results between the KU100's internal microphones and the lavalier microphones used for human measurements. At this time, the conclusion must therefore be that there is something specific about the HRTF measurements of human subjects that has a significant negative impact of perceived timbral quality. Much further research is needed on this topic but explanations could range from minor head movements during measurement to acoustic skin absorption coefficients.

Time-Alignment and BiRADIAL

This comparison has possibly the most interesting and relevant findings as the results differ depending on whether the HRTFs utilised were generic or individual. On average, BiRADIAL rendering appears to increase the source localisation error by approximately 5°, see Fig. 7.15. This is unexpected given the increased accuracy of ILD and wavefront reconstruction shown in Section 6.5.4, but at the same time is a very small difference in comparison to the overall range of results.

Regarding the timbral attributes, BiRADIAL rendering appears to offer clear advantages over Time-Alignment for generic HRTFs whist suffering severe decreases in performance compared to Time-Alignment for individual measurements. Whilst initially this is a somewhat concerning result there are a few possible explanations.

For generic HRTFs, BiRADIAL rendering is rated on average higher than Time-Alignment with regards to overall preference, Basic Audio Quality and richness - all deemed to be positive attributes. In fact, it receives the highest ratings of all the stimuli tested. However, renderers that implemented individual HRTFs are rated very differently. They are ranked the worst performing with regards to preference, basic audio quality and richness and are notably brighter than all other renderers. It is possible though, that the reason for this is not due to the BiRADIAL technique, but the way in which the HRTFs are measured.

It has been suggested in this thesis (Section 3.4.5) that individual HRTF performance may well suffer due to the fact that a participant is unable to stay perfectly still during measurement. The fallout of this is that there are mid-high frequency errors in the HRTFs. This problem is made worse when 2 separate measurements are required to construct a BiRADIAL HRTF as the errors are no longer even consistent between the two ears. Such an explanation would not only help to explain the decreased performance of individual HRTFs in general, but further the worse performance of individual BiRADIAL rendering.

The premise is also supported by the especially large range (and interquartile range) of results for individual BiRADIAL rendering. For example, if it so happens that a person is able to remain still enough during their measurements then they will experience superior performance. Again, further tests examining the accuracy of individual measurements are needed to support this argument. It can be said with some certainty though, that there are significant perceptual differences between the BiRADIAL and Time-Alignment approach.

SADIIE and MARC

The final and most meaningful conclusion is that of the validity of the MARC HRTFs. The main concern of the near-field measurements was that they would not perform as well with respect to timbre given the lower quality of loudspeaker and would suffer with respect to externalisation given the distance of the loudspeakers from the head. However, despite savings in time, cost and complexity the results show that not only are the MARC HRTFs comparable to the SADIIE HRTFs but the renderers are in fact capable of out-performing the SADIIE measurements. This is shown throughout Figs. 7.14, 7.15 and 7.16 by comparing the results of the generic KU100 HRTFs (first 3 renderers: blue [SADIIE], red [MARC] and green [MARC]). In particular the first 2 renderers may be compared (blue and red) as each uses the Time-Alignment optimization.

One reason for this may be the natural progression and incremental improvements in the windowing and filtering parameters constantly being adjusted and tweaked to improve results. However, a big advantage in optimizing these parameters is the ability to capture an entire set of HRTFs in such a short space of time. At this point it is interesting to recall that the SADIIE HRTFs being tested are 256 tap (@48,000Hz sampling rate) whilst the MARC HRTFs have been reduced to just 128 tap.

Another difference is that the MARC renderers employed 2702 virtual loudspeakers whilst the SADIIE renderer was restricted to 50. Whilst one could argue that this is simply a benefit of the capture system (that it is capable of outputting any number of HRTFs) one could also argue that it is an unfair comparison. However, as discussed in Section 6.5.4, there is no benefit to synthesising large number of HRTFs when using the Time-Alignment approach. The results are the same as when synthesising only 50. It is therefore fair to compare the SADIIE and MARC Time-Alignment renderers. What is clear in this case, is that there are no general, obvious, clear or statistical differences in either the localisation or timbral performances of these two capture systems. Further, considering the median values and interquartile ranges it is the MARC system that performs superior.

7.3 Summary

A perceptual listening test was conducted to investigate the performance of generic HRTFs from the SADIIE databse and generic and individual HRTFs measured in MARC. Time-Alignment and BiRADIAL optimization techniques were used to represent state of the art binaural Ambisonic rendering. There were 3 key findings: the first is that dummy head HRTFs measured in MARC are comparable in terms of both localisation and timbral quality to those measured as part of the SADIIE database. This is despite significant reductions in time, cost and efforts acquiring the measurements and a 50% reduction in the tap length of the filters.

The second and third finding relate to the performance of BiRADIAL with respect to Time-Alignment. Preliminary results show that although BiRADIAL may offer a substantial timbral improvement over Time-Alignment for generic HRTFs, the same cannot be said for individual measurements. In the case of individual measurements there was in fact a large reduction in timbral quality when using a BiRADIAL renderer. It is possible that this may be attributed to the random movement of a human subject during the measurement phase, but this is not confirmed. BiRADIAL rendering also shows on average a slightly larger error with respect to localisation ($< 5°$), but the significance of this error has not been identified.

Future work would look to increase the number of participants within the study and investigate the statistical significance of these findings. An improved test workflow would also allow for the randomisation of the order of point sources in the localisation study in order to eliminate potential bias in the results. However, results to date combined with a theoretically more accurate wavefront reconstruction show great promise for the new HRTF measurement technique and BiRADIAL optimisation.

This Page Has
Intentionally Been
Left Blank

8 Conclusion

Chapter Overview

This chapter concludes the thesis by revisiting the Hypothesis and assessing its output against the original objectives. A summary of the work presented, followed by suggestions of future work and final remarks.

8.1 Introduction

This thesis has focused on the development and perceptual analysis of binaural Ambisonic renderers. To summarise the work the hypothesis, originally stated in Chapter 1, will be revisited and conclusions will be drawn as to whether the accompanying objectives were met. This will be followed by a more in depth summary of the preceding chapters followed by a note on future work and final remarks.

8.2 Review of Thesis

8.2.1 Consideration of Hypothesis

In Section 1.2.1 the hypothesis was stated:

> *Improvements can be made to binaural Ambisonic rendering*
> *workflows through the spatio-temporal manipulation of HRTFs*
> *within a feasible measurement procedure.*

which was to be answered though the following objectives:

1. To compile a comprehensive review of the human auditory system and Ambisonic reproduction technologies

2. To compare the performance of spatially and temporally manipulated HRTFs in the context of binaural Ambisonic rendering following a traditional HRTF capture workflow

3. To deliver a fast and convenient HRTF capture system able to exploit the findings of Objective 2 to deliver optimized individual HRTFs for the end user.

The first objective was met in Chapters 2 and 4 with a comprehensive review of binaural audio, spatial hearing, Ambisonic reproduction and binaural rendering. The second was met in Chapter 5 which includes comparisons of the BiRADIAL and Time-Alignment renderers with HRTF measurements that were taken as part of the SADIIE database, presented in Chapter 3.

Soundfield reconstruction simulations and objective analysis supported the case that spatio-temporal manipulations of HRTFs can improve the performance of binaural Ambisonic renderers. Fig. 5.3 shows how the bounding edge of an Ambisonic sweet spot may be shifted to center around a person's individual ears within a typical decoder. Improvements in binaural output are then evident with respect to both ILDs, shown in Fig. 5.27, and spectral response, shown in Fig. 5.28.

Following these results, the third objective was met in Chapter 6 with the development of MARC. The hypothesis is then answered by the results of the perceptual listening test in Chapter 7. It is found that improvements can be made to the binaural rendering of Ambisonics though the temporal and spatial alignment of HRTFs. It is also found that by implementing these manipulations within a quick and convenient HRTF capture system infinite HRTFs may be generated that are perceptually similar quality to traditionally measured HRTFs.

The overall accuracy of the binaural renderers remains undefined as the analysis of this would require real world reference sources to be compared against within a listening test. However, it can be said with some certainty that the methods presented in this thesis for capturing and optimizing HRTFs specifically for binaural Ambisonic rendering are both more convenient than ever before, and result in a superior quality of binaural signal compared to previous virtual loudspeaker rendering techniques.

8.2.2 Summary of Work

Chapter 2 discusses binaural audio, the human auditory system and the methods with which people localise sounds in space. These are important metrics to consider before methods can be developed to improve the way in which sound in presented to a listener. As such, these features were used throughout the thesis as points of comparison between prospective new technologies.

The PSDM is a perceptually driven model for comparing the frequency spectra of similar HRTFs in order to determine their perceived similarity. It is based on equal loudness contours, perceptual loudness scales and ERB weightings to account for the frequency and amplitude sensitivities and resolutions of the human hear. Its output is shown in Chapter 2 to vary more in line with human perception compared to a typical spectral difference calculation.

The SADIIE database is a collection of over 60,000 binaural IRs which is now being used internationally for research and development by academic institutions and within industry. It provides over 2100 HRTFs for 18 human subjects and over 8800 HRTFs for the KU100 and KEMAR mannequins. The HRTFs measured are compatible with at least 20 Ambisonic loudspeaker configurations ranging from 1^{st} to 5^{th} order 2D and 3D Ambisonic layouts. BRIRs for the 50 point (nesting the 26 and 8 point) Lebedev grid and headphone equalisation filters are also available together with anthropomorphic data.

A perceptual listening test in Chapter 3 has shown that individual HRTFs are not necessarily the optimal rendering solution with regards to timbral performance. Although this test did not consider localisation, the results indicated with statistical significance that particular HRTF sets are more generally preferred to others. In particular, HRTFs of the KU100 had the highest rated median and mean value. Second to the KU100 was the KEMAR.

Chapter 4 presented a complete review of Ambisonics and its role within a binaural Ambisonic renderer. The principles of Ambisonic rendering are crucial in understanding the formation of the sweet spot and hence how this can be manipulated and exploited within BiRADIAL and Time-Alignment renderers. The chapter summarises with another key consideration of binaural audio, head-tracking, and the ways and methods in which Ambisonics in well suited to handle this.

Chapter 5 summarises a new understanding of the Time-Alignment approach from the view of displacing the sweet spot within either a virtual or real loudspeaker array. Previously, the approach had only been considered as a reduction in the energy of the HRTFs within the higher spherical harmonic orders. However, this has no benefit to a real world context. The approach of shifting the sweet spot would allow the Ambisonic reproduction to be tracked to a real-listener no matter where they are situated within an array.

Chapter 5 also presents BiRADIAL as a technique for rendering binaural Ambisonics. Wavefront reconstruction simulations have shown how this method better approximates planar sources compared to a Time-Alignment approach, given the use of near-field compensation filters. Objective analysis of the rendered signals show it is comparable to Time-Alignment for semi-far-field (1.2m radii) virtual loudspeaker arrays.

MARC is presented in Chapter 6. It is a two-part HRTF measurement system that encompasses both the measurement of near-field HRTFs and the synthesis of far-field HRTF approximations. This is achieved via the optimised binaural rendering of Dirac pulses encoded into Ambisonics with near-field compensation filters and reproduced with either a BiRADIAL or Time-Alignment approach. Objective comparisons have shown that the far-field HRTFs are of similar composition (with respect to ITDs, ILDs and spectral cues) to the HRTFs from the SADIIE database.

This is confirmed by a perceptual listening test in Chapter 7. The test compares binaural Ambisonic reproductions with a variety of HRTFs and rendering techniques. HRTFs measured within MARC were, in general, rated similarly if not higher with regards to overall preference compared to those from SADIIE. For generic HRTFs, BiRADIAL also outperformed Time-Alignment with regards to preference, although the opposite was true for individual HRTFs. This tests represents the first timbral comparison test of musical stimuli rendered with binaural Ambisonics using individual and non-individual HRTFs. Once again, it supports the conclusion that individual HRTFs are not necessarily the best option to use in every case. However, an accompanying localisation test did confirm an increase in localisation accuracy with the individual measurements.

8.3 Future Work

It is quite possible that the methods of optimisation presented in this thesis based on the temporal and spatial alignment of the virtual loudspeaker arrays could see incremental improvements with further research and modifications. However, it is also of great importance at this stage to further define the metrics with which humans value spatial audio reproduction.

For example, is there a tangible benefit to developing a rendering method that provides superior localisation performance at the cost of timbre? Currently this question represents the difference between individual and non-individual HRTFs. It is important to define a set of test criteria that sufficiently describe the performance of binaural renderers in every relevant way. This would mean first assessing the use cases of binaural audio and then analysing the most valuable attributes on a case by case basis. By doing this renderers may be tailored such that they are best suited to their application for optimal results.

More specific to the work presented here, further tests are needed to confirm the performance of MARC in order to obtain statistical significance within the tests. One obvious direction in which this could lead is the compilation of another HRTF database, SADIIIE! Having confirmed the quality of the synthesised far-field HRTFs this database could focus on gathering mass data on human subjects. It is feasible to imagine a database of several hundred subjects each with as many HRTFs as were desired. This quantity of data could be useful in analysing patterns/differences between subjects or even providing personalised measurements without requiring every future subject to have their own HRTFs actually measured.

8.4 Final remarks

This thesis has explored the measurement of HRTFs and their spatio-temporal optimisation for binaural Ambisonic rendering. Improvements in binaural output have been identified both objectively and subjectively. A new approach to individual HRTF measurement has been suggested that utilizes these optimization techniques to infinitely synthesise accurate far-field HRTFs from fast and convenient near-field measurements. The application of such technology stands not only to improve current rendering systems, through nothing more than the substitution of HRTFs, but to facilitate large scale HRTF data capture and individualisation of new rendering systems throughout the marketplace.

Fin.

This Page Has
Intentionally Been
Left Blank

Glossary

2D	2-dimensional
3D	3-dimensional
ACN	Ambisonic Channel Number
AR	Augmented Reality
ASD	Absolute Spectral Difference
BEM	Boundry Element Method
BiRADIAL	Binaural Rendering of Audio through Duplex Independant Auralised Listening
BRIR	Binaural Room Impulse Response
DAW	Digital Audio Workstation
ERB	Equivalent Rectangular Bandwidth
FOA	First Order Ambisonics
FFT	Fast Fourier Transform
HATS	Head And Torso Simulator
HOA	Higher Order Ambisonics
HRIR	Head Related Impulse Response
HRTF	Head Related Transfer Function
IACC	InterAural Cross-Correlation
ILD	Interaural Level Difference
IR	Impulse Responses
ITD	Interaural Time Difference
JND	Just Noticeable Difference
LAN	Local Area Network
MARC	Miniaturised Acoustic Response Chamber
MaxIACC	Maximum InterAural Cross-Correlation
MLS	Maximum Length Sequence
NFC	Near-Field Control
OME	Outer-Middle-Ear
PSDM	Perceptual Spectral Difference Model
SNR	Signal to Noise Ratio

SVD	Singular Value Decomposition
VBAP	Vector Based Amplitude Panning
VR	Virtual Reality
WFS	Wave Field Synthesis

Bibliography

Aaronson, Neil L. and William M. Hartmann (2014). "Testing , correcting , and extending the Woodworth model for interaural time difference". In: *The Journal of the Acoustical Society of America* 135.2. DOI: 10.1121/1.4861243.

Abhayapala, Thushara D. and Darren B. Ward (2002). "Theory and design of high order sound field microphones using spherical microphone array". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1949–1952. ISSN: 15206149. DOI: 10.1109/ICASSP.2002.5745011.

Adams, Stephen and Frank Boland (2010). "On the Distortion of Binaural Localization Cues using Headphones". In: *IET Irish Signals and Systems Conference*. Cork.

Algazi, V. Ralph, Carlos Avendano and Richard O. Duda (2001). "Elevation localization and head-related transfer function analysis at low frequencies". In: *The Journal of the Acoustical Society of America* 109.3, pp. 1110–1122. ISSN: 0001-4966. DOI: 10.1121/1.1349185.

Algazi, V. Ralph et al. (2002a). "Approximating the head-related transfer function using simple geometric models of the head and torso". In: *The Journal of the Acoustical Society of America* 112.5, pp. 2053–2064. ISSN: 0001-4966. DOI: 10.112 1/1.1508780.

– (2002b). "Approximating the head-related transfer function using simple geometric models of the head and torso". In: *The Journal of the Acoustical Society of America* 112.5, pp. 2053–2064. ISSN: 00014966. DOI: 10.1121/1.1508780.

Algazi, V.R. et al. (2001). "The CIPIC HRTF database". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz. ISBN: 0-7803-7126-7. DOI: 10.1109/ASPAA.2001.969552.

Andreopoulou, Areti, Durand R. Begault and Brian F. G. Katz (2015). "Inter-Laboratory Round Robin HRTF Measurement Comparison". In: *Inter-Laboratory Round Robin HRTF Measurement Comparison* 9.5, pp. 895–906. DOI: 10.1109/J STSP.2015.2400417.

Andreopoulou, Areti and Brian F. G. Katz (2015). "On the Use of Subjective Hrtf Evaluations for Creating Global Perceptual Similarity Metrics of Assessors and Assessees". In: *The 21st International Conference on Auditory Display (ICAD 2015)*, pp. 13–20.

– (2016a). "Investigation on Subjective HRTF Rating Repeatability". In: *AES 140th Convention*.

– (2016b). "Subjective HRTF evaluations for obtaining global similarity metrics of assessors and assessees". In: *Journal on Multimodal User Interfaces* 10.3, pp. 259–271. ISSN: 17838738. DOI: 10.1007/s12193-016-0214-y.

Aoshima, Nobuharu (1981). "Computer-generated pulse signal applied for sound measurement". In: *The Journal of the Acoustical Society of America* 69.5, pp. 1484–1488. ISSN: NA. DOI: 10.1121/1.385782.

Armstrong, Cal and Gavin Kearney (2020). "Acoustic measurements". Pat. GB1918010.8. UK Patent Office.

Armstrong, Cal, Damian Murphy and Gavin Kearney (2018). "A Bi-RADIAL Approach to Ambisonics". In: *AES International Conference on Audio for Virtual and Augmented Reality*.

Armstrong, Cal et al. (2017). "Simultaneous HRTF Measurement of Multiple Source Configurations Utilizing Semi-Permanent Structural Mounts". In: *AES 143rd Convention*.

Armstrong, Cal et al. (2018a). "A Perceptual Evaluation of Individual and Non-Individual HRTFs : A Case Study of the SADIE II Database". In: *Applied Sciences* 8.11. DOI: 10.3390/app8112029.

Armstrong, Cal et al. (2018b). "A Perceptual Spectral Difference Model for Binaural Signals". In: *AES 145th Convention*.

Avni, Amir et al. (2013). "Spatial perception of sound fields recorded by spherical microphone arrays with varying spatial resolution". In: *The Journal of the Acoustical Society of America* 133.2711. DOI: 10.1121/1.4795780.

Bahu, Hélène et al. (2016). "Comparison of Different Egocentric Pointing Methods for 3D Sound Localization Experiments". In: *Acta Acustica united with Acustica* 102. DOI: 10.3813/AAA.918928.

Bajnok, Bela (1991). "Construction of Designs on the 2-Sphere". In: *European Journal of Combinatorics* 12.5, pp. 377–382. ISSN: 01956698. DOI: 10.1016/S0195-6698(13)80013-3.

Balmages, Ilya and Boaz Rafaely (2007). “Open-Sphere Designs for Spherical Microphone Arrays”. In: *IEEE Transactions on Audio, Speech and Language Processing* 15.2. ISSN: 1098-6596. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

Bauer, B. B. and E. L. Torick (1966). “Researches in Loudness Measurement”. In: *IEEE Transactions on Audio and Electroacoustics* 14.3, pp. 141–151. ISSN: 00189278. DOI: 10.1109/TAU.1966.1161864.

Beerends, John G. and Jan A. Stemerdink (1994). “A Perceptual Speech-Quality Measure Based on a Psychoacoustic Sound Representation”. In: *Journal of the Audio Engineering Society* 42.3, pp. 115–123.

Begault, Durand R. (1992). “Perceptual Effects of Synthetic Reverberationon on Three-Dimensional Audio Systems”. In: *Journal of the Audio Engineering Sociaty* 40.11.

Begault, Durand R. et al. (2000). “Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source”. In: *AES 108th Convention*. Paris. ISBN: 0780356128. DOI: 10.1109/ASPAA.1999.810884.

Berg, Jan and Francis Rumsey (1999). “Identification of Perceived Spatial Attributes of Recordings by Repertory Grid Technique and other methods”. In: *AES 106th Convention*. Munich.

Berge, Svein and Natasha Barrett (2010). “High Angular Resolution Planewave Expansion”. In: *Proceedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*.

Berkhout, A J (1988). “A Holographic Approach to Acoustic Control”. In: *Journal of the Audio Engineering Society* 36.12, pp. 977–995. ISSN: 00047554.

Bernschütz, Benjamin (2013). “A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100”. In: *the 40th Italian (AIA) Annual Conference on Acoustics and the 39th German Annual Conference on Acoustics (DAGA)*. Merano.

Bertet, Stéphanie et al. (2013). “Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources”. In: *Acta Acustica united with Acustica* 99.4, pp. 642–657. ISSN: 16101928. DOI: 10.3813/AAA.918643.

Best, Catherine T., Gerald W McRoberts and Nomathemba M. Sithole (1988). “Examination of Perceptual Reorganization for Nonnative Speech”. In: *Journal of Experimental Psychology: Human Perception and Performance* 14.3, pp. 345–360.

Best, Virginia et al. (2013). “Spatial release from masking as a function of the spectral overlap of competing talkers”. In: *The Journal of the Acoustical Society of America* 133.6, pp. 3677–3680. ISSN: 0001-4966. DOI: 10.1121/1.4803517.

Blauert, Jens (1970). “Ein Versuch zum Richtungshören bei gleichzeitiger optischer Stimulation”. In: *Acta Acustica united with Acustica* 23, pp. 118–119.

– (1997). *Spatial hearing: the phychophysics of human sound localization.* MIT press.

Bomhardt, Ramona, Matias De La Fuente Klein and Janina Fels (2016). “A high-resolution head-related transfer function and three-dimensional ear model database”. In: *Proceedings of Meetings on Acoustics.* Vol. 29. DOI: 10.1121/2.0000467.

Brainard, Michael S, Eric I Knudsen and Steven D Esterly (1992). “Neural derivation of sound source location: Resolution of spatial ambiguities in binaural cues”. In: *The Journal of the Acoustical Society of America* 1015.91. DOI: 10.1121/1.402627.

Brinkmann, Fabian et al. (2019). *A cross-evaluated database of measured and simulated HRTFs including 3d head meshes, anthropometric features, and headphone impulse responses.* Tech. rep. 9, pp. 705–718. DOI: 10.17743/jaes.2019.0024.

Brungart, Douglas S (1999). “Auditory parallax effects in the hrtf for nearby sources”. In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*

Brungart, Douglas S and William M Rabinowitz (1999). “Auditory localization of nearby sources . Head-related transfer functions Auditory localization of nearby sources . Head-related”. In: *The Journal of the Acoustical Society of America* 106.3. DOI: 10.1121/1.427180.

Brungart, Douglas S. and Brian D. Simpson (2002). “The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal”. In: *The Journal of the Acoustical Society of America* 112.2, pp. 664–676. ISSN: 0001-4966. DOI: 10.1121/1.1490592.

Brungart, Douglas S., Brian D. Simpson and Alexander J. Kordik (2005). “The detectability of headtracker latency in virtual audio displays”. In: *Eleventh Meeting of the International Conference on Auditory Display (ICAD 05).* Limerick.

Brungart, Douglas S. et al. (2004). “The interaction between head-tracker latency, source duration, and response time in the localization of virtual sound sources”. In: *Tenth Meeting of the International Conference on Auditory Display (ICAD 04).* Sydney.

Bücklein, Roland (1981). “The audibility of frequency response irregularities”. In: *Journal of the Audio Engineering Society* 29.3.

Burkhard, M. D. and R. M. Sachs (1975). “Anthropometric manikin for acoustic research”. In: *The Journal of the Acoustical Society of America* 58.214. DOI: 10.1121/1.380648.

Cardozo, B L (1965). “Adjusting the Method of Adjustment : SD vs DL”. In: *The Journal of the Acoustical Society of America* 37.786. DOI: 10.1121/1.1909439.

Carpentier, Thibaut et al. (2014). “Measurement of a head-related transfer function database with high spatial resolution”. In: *Proceedings of Forum Acusticum*. ISBN: 9788361402282.

Ceperley, Peter (2016). *Resonances, waves and fields: Spherical harmonics*. URL: http://resonanceswavesandfields.blogspot.co.uk/2016/05/spherical-harmonics.html.

Chardon, Gilles, Wolfgang Kreuzer and Markus Noisternig (2014). “Design of a robust open spherical microphone array”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. ISBN: 9781479928927. DOI: 10.1109/ICASSP.2014.6854919.

– (2015). “Design of Spatial Microphone Arrays for Sound Field Interpolation”. In: *IEEE Journal on Selected Topics in Signal Processing*. ISSN: 19324553. DOI: 10.1109/JSTSP.2015.2412097.

Chen, Xiaojun (2009). “Numerical Verification Methods for Spherical t -Designs”. In: *Japan Journal of Industrial and Applied Mathematics* 26.2-3, pp. 317–325. DOI: 10.1007/bf03186537.

Courrieu, Pierre (2008). “Fast Computation of Moore-Penrose Inverse Matrices”. In: *Neural Information Processing* 8.2, pp. 25–29. arXiv: 0804.4809.

Damaske, P and B Wagener (1969). “Richtungshorversuche fiber einen nachgebildeten Kopf”. In: *Acta Acustica united with Acustica* 21, pp. 30–35.

Daniel, Jérôme (2001). “Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia”. PhD thesis. l’Université Paris, p. 319.

– (2003). “Spatial Sound Encoding Including Near Field Effect: Introducing Distance Coding Filters and a Viable, New Ambisonic Format”. In: *AES 23rd International Conference*. Copenhagen.

Daniel, Jérôme, Rozenn Nicol and Sébastien Moreau (2003). “Further investigations of high order ambisonics and wavefield synthesis for holophonic sound imaging”. In: *AES 114th Convention*. Amsterdam. DOI: 10.1.1.459.117.

Daniel, Jérôme, Jean-Bernard Rault and Jean-Dominique Polack (1998). "Ambisonics Encoding of Other Audio Formats for Multiple Listening Conditions". In: *AES 105th Convention*. California.

Dolby (2016). *Dolby Atmos Home Theater Installation Guidelines*. Tech. rep. URL: https://www.dolby.com/us/en/technologies/dolby-atmos/dolby-atmos-home-theater-installation-guidelines.pdf.

Dooley, Wes L. and Ron D. Streicher (1982). "M-S Stereo: A powerful technique for working in stereo". In: *Journal of the Audio Engineering Society* 30.10, pp. 707–718. ISSN: 00047554.

Dunn, Chris and Malcolm Omar Hawksford (1993). "Distortion Immunity of MLS-Derived Impulse Response Measurements". In: *Journal of the Audio Engineering Sociaty* 41.5.

Duraiswami, Ramani, Dmitry N. Zotkin and Nail A. Gumerov (2004). "Interpolation and range extrapolation of HRTFs". In: *IEEE International Conference on Acoustics, Speech and Signal Processing* 4.January 2004. ISSN: 15206149.

Durlach, N. I. et al. (1992). "On the Externalization of Auditory Images". In: *Presence Teleoperators & Virtual Environments* 1.2. DOI: 10.1162/pres.1992.1.2.251.

EBU (1997). *Assessment methods for the subjective evaluation of the quality of sound programme material – Music*. Tech. rep. 3286–E. European Broadcasting Union. URL: https://tech.ebu.ch/docs/tech/tech3286.pdf.

Epain, Nicolas, Craig T. Jin and Franz Zotter (2014). "Ambisonic decoding with constant angular spread". In: *Acta Acustica united with Acustica* 100.5, pp. 928–936. ISSN: 16101928. DOI: 10.3813/AAA.918772.

ETSI (2018). *TS 126 259 - V15.0.0 - 5G; Subjective test methodologies for the evaluation of immersive audio systems (3GPP TS 26.259 version 15.0.0 Release 15)*. Tech. rep. European Telecommunications Standards Institute.

Evans, Michael J., James A. S. Angus and Anthony I. Tew (1998). "Analyzing head-related transfer function measurements using surface spherical harmonics". In: *The Journal of the Acoustical Society of America* 104.4, pp. 2400–2411. DOI: 10.1121/1.423749.

Farina, Angelo (2000). "Simultaneous measurement of impulse response and distortion with a swept-sine technique". In: *AES 108th Convention*. Paris. ISBN: 0780356128. DOI: 10.1109/ASPAA.1999.810884.

– (2006). *A-format to B-format conversion.* URL: http://pcfarina.eng.unipr.it/Public/B-format/A2B-conversion/A2B.htm (visited on 11/22/2019).

Favrot, S. and J. M. Buchholz (2012). “Reproduction of nearby sound sources using higher-order ambisonics with practical loudspeaker arrays”. In: *Acta Acustica united with Acustica* 98, pp. 48–60. ISSN: 16101928. DOI: 10.3813/AAA.918491.

Fellgett, Peter (1975). “Ambisonics. Part one: Genereal System Description”. In: *Studio Sound* 17.8.

Fletcher, Harvey and W. A. Munson (1933). “Loudness, Its Definition, Measurement and Calculation”. In: *The Journal of the Acoustical Society of America* 5.82. DOI: 10.1121/1.1915637.

– (1937). “Relation between loudness and masking”. In: *The Journal of the Acoustical Society of America* 9.1. ISSN: 00014966. DOI: 10.1121/1.1915904.

Freyman, Richard L., Uma Balakrishnan and Karen S. Helfer (2001). “Spatial release from informational masking in speech recognition”. In: *The Journal of the Acoustical Society of America* 109.5, pp. 2112–2122. ISSN: 0001-4966. DOI: 10.1121/1.1354984.

Gardner, William G. and Keith D. Martin (1995). “HRTF measurements of a KEMAR”. In: *The Journal of the Acoustical Society of America* 97.6, pp. 3907–3908. ISSN: 0001-4966. DOI: 10.1121/1.412407.

Genelec (2014). *8010A Operating Manual.* Tech. rep. URL: www.genelec.com.

– (2017). *Genelec and IDA Audio to redefine immersive 3D Audio for professional headphone users.* URL: https://www.genelec.com/genelec-and-ida-audio-redefine-immersive-3d-audio-professional-headphone-users (visited on 11/22/2019).

Geronazzo, Michele, Simone Spagnol and Federico Avanzini (2018). “Do we need individual head-related transfer functions for vertical localization? the case study of a spectral notch distance metric”. In: *IEEE/ACM Transactions on Audio Speech and Language Processing* 26.7, pp. 1243–1256. ISSN: 23299290. DOI: 10.1109/TASLP.2018.2821846.

Gerzon, Michael A. (1973). “Periphony: With-Height Sound Reproduction”. In: *Journal of the Audio Engineering Sociaty* 21.1.

– (1975a). “Ambisonics. Part two: Studio techniques”. In: *Studio Sound* 17.8.

– (1975b). “The Design Of Precisely Coincident Microphone Arrays For Stereo And Surround Sound”. In: *AES 50th Convention.*

Gerzon, Michael A. (1977). “Design of Ambisonic Decoders for Multispeaker Surround Sound”. In: *AES 58th Convention.* New York.

– (1980). “Practical Periphony: The Reproduction of Full-Sphere Sound”. In: *AES 65th Convention.* London.

– (1992a). “General Metatheory of Auditory Localization”. In: *AES 92nd Convention.* Vienna. DOI: 10.1111/j.1365-2141.1992.tb04620.x.

– (1992b). “Psychoacoustic Decoders for Multispeaker Stereo and Surround Sound”. In: *AES 93rd Convention.* San Francisco.

Gerzon, Michael A. and Geoffrey J. Barton (1992). “Ambisonic Decoders for HDTV”. In: *AES 92nd Convention.* Vienna.

Gilkey, Robert H. et al. (1995). “A pointing technique for rapidly collecting localization responses in auditory research”. In: *Behavior Research Methods, Instruments, & Computers* 27.1.

Glyde, Helen et al. (2013). “The importance of interaural time differences and level differences in spatial release from masking”. In: *The Journal of the Acoustical Society of America* 134.2, EL147–EL152. ISSN: 0001-4966. DOI: 10.1121/1.481244 1.

Goldberg, J. N. et al. (1967). “Spin-s Spherical Harmonics and ð”. In: *Journal of Mathematical Physics* 8.11, p. 2155. ISSN: 00222488. DOI: 10.1063/1.1705135.

Golub, G . and W . Kahan (1965). “Calculating the Singular Values and Pseudo-Inverse of a Matrix”. In: *Journal of the Society for Industrial and Applied Mathematics* 2.2, pp. 205–224. DOI: 10.1137/0702016.

Gorzel, Marcin, Gavin Kearney and Frank Boland (2014). “Investigation of Ambisonic Rendering of Elevated Sound Sources”. In: *AES 55th International Conference*, pp. 1–10. ISBN: 9780937803981.

Gorzel, Marcin et al. (2019). “Efficient encoding and decoding of binaural sound with resonance audio”. In: *AES Conference on Immersive and Interactive Audio.*

Gover, Bradford N., James G. Ryan and Michael R. Stinson (2004). “Measurements of directional properties of reverberant sound fields in rooms using a spherical microphone array”. In: *The Journal of the Acoustical Society of America* 116.4, pp. 2138–2148. ISSN: 0001-4966. DOI: 10.1121/1.1787525.

Graham, Craven Peter and Michael A. Gerzon (1977). *Coincident microphone simulation covering three dimensional space and yielding various directional outputs.*

Griesinger, David (1989). “Equalization and Spatial Equalization of Dummy-head Recordings for Loudspeaker Reproduction”. In: *Journal of the Audio Engineering Society* 37.1/2, pp. 20–29. ISSN: 00047554.

Guldenschuh, Markus, Alois Sontacchi and Franz Zotter (2008). “HRTF modelling in due consideration variable torso reflections”. In: *Acoustics’08*. Paris.

Gumerov, Nail A. et al. (2010). “Computation of the head-related transfer function via the fast multipole accelerated boundary element method and its spherical harmonic representation”. In: *The Journal of the Acoustical Society of America* 127.1, pp. 370–386. ISSN: 0001-4966. DOI: 10.1121/1.3257598.

Gupta, Aastha and Thushara D. Abhayapala (2010). “Double sided cone array for spherical harmonic analysis of wavefields”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 77–80. ISSN: 15206149. DOI: 10.1109 /ICASSP.2010.5496193.

Gupta, Navarun et al. (2010). “HRTF database at FIU DSP Lab”. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 169–172. ISSN: 15206149. DOI: 10.1109/ICASSP.2010.5496084.

Haber, Howard (2012). *The Spherical Harmonics*. Santa Cruz. URL: http://scipp.ucsc.edu/~haber/ph116C/SphericalHarmonics_12.pdf.

Hardin, R. H. and N. J A Sloane (1996). “McLaren’s improved snub cube and other new spherical designs in three dimensions”. In: *Discrete & Computational Geometry* 15.4, pp. 429–441. ISSN: 0179-5376. DOI: 10.1007/BF02711518. arXiv: 0207211 [math.CO].

Heller, Aaron J, Eric M. Benjamin and Richard Lee (2010). “Design of Ambisonic Decoders for Irregular Arrays of Loudspeakers by Non-Linear Optimization”. In: *AES 129th Convention*. San Francisco. ISBN: 9781617821943.

– (2012). “A Toolkit for the Design of Ambisonic Decoders”. In: *Linux Audio Conference*. Stanford.

Heller, Aaron J, Richard Lee and Eric M. Benjamin (2008). “Is My Decoder Ambisonic?” In: *AES 125th Convention*. San Francisco.

Herre, Jürgen et al. (2014). “MPEG-H Audio - The New Standard for Universal Spatial/3D Audio Coding”. In: *Journal of the Audio Engineering Society* 62.12, pp. 821 –830. DOI: 10.17743/jaes.2014.0049.

Hofman, Paul M., Jos G.A. Van Riswick and A. John Van Opstal (1998). "Relearning sound localization with new ears". In: *Nature Neuroscience* 1.5, pp. 417–421. ISSN: 10976256. DOI: 10.1038/1633.

Hollerweger, Florian (2006). "Periphonic Sound Spatialization in Multi-User Virtual Environments". PhD thesis. University of California, p. 125.

– (2008). *An Introduction to Higher Order Ambisonic*. Tech. rep. URL: https://pdfs.semanticscholar.org/40b6/8e33d74953b9d9fe1b7cf50368db492c898c.pdf.

Huopaniemi, Jyri, Nick Zacharov and Matti Karjalainen (1999). "Objective and Subjective Evaluationof Head,Related Transfer Function Filter Design". In: *Journal of the Audio Engineering Sociaty* 47.4, pp. 218–239.

Hur, Y et al. (2008). "Efficient individualization of HRTF using critical-band based spectral cues control". In: *AES 124th Convention* 180, pp. 167–180.

ISO (2003). *Acoustics – Normal equal-loudness-level contours*. Tech. rep. Geneva: International Standardization Organization.

– (2009). *Acoustics Measurement of Room Acoustic Parameters Part 1: Performance Spaces*. Tech. rep. International Organization for Standardization.

ITU (2001). *BS.1387-1 Method for objective measurements of perceived audio quality*. Tech. rep. International Telecommunication Union.

– (2003a). *BS.1284-1 General methods for the subjective assessment of sound quality*. Tech. rep. International Telecommunication Union.

– (2003b). *BS.1534-1 Method for the subjective assessment of intermediate quality level of coding systems Annex 1*. Tech. rep. International Telecommunication Union.

– (2015). *BS.1534-3 Method for the subjective assessment of intermediate quality level of audio systems*. Tech. rep. International Telecommunication Union.

Iwaya, Yukio, Yôiti Suzuki and Daisuke Kimura (2003). "Effects of head movement on front-back error in sound localization". In: *Acoustical Science and Technology* 24.5, pp. 322–324. ISSN: 13463969. DOI: 10.1250/ast.24.322.

Jin, Craig T., Nicolas Epain and Abhaya Parthy (2014). "Design, optimization and evaluation of a dual-radius spherical microphone array". In: *IEEE Transactions on Audio, Speech and Language Processing* 22.1, pp. 193–204. ISSN: 15587916. DOI: 10.1109/TASLP.2013.2286920.

Jin, Craig T. et al. (2014). "Creating the Sydney York Morphological and Acoustic Recordings of Ears Database". In: *IEEE Transactions on Multimedia* 16.1, pp. 37–46. DOI: 10.1109/TMM.2013.2282134.

Jot, Jean Marc (1999). "Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces". In: *Multimedia Systems* 7, pp. 55–69. ISSN: 09424962. DOI: 10.1007/s005300050111.

Kato, Masaharu et al. (2003). "The effect of head motion on the accuracy of sound localization". In: *Acoustical Science and Technology* 24.5, pp. 315–317. ISSN: 13463969. DOI: 10.1250/ast.24.315.

Katz, Brian F. G. (2001a). "Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation". In: *The Journal of the Acoustical Society of America* 110.5, pp. 2440–2448. ISSN: 0001-4966. DOI: 10.1121/1.1412440.

– (2001b). "Boundary element method calculation of individual head-related transfer function. II. Impedance effects and comparisons to real measurements". In: *The Journal of the Acoustical Society of America* 110.5, pp. 2449–2455. ISSN: 0001-4966. DOI: 10.1121/1.1412441.

Katz, Brian F. G. and Markus Noisternig (2014). "A comparative study of interaural time delay estimation methods". In: *The Journal of the Acoustical Society of America* 135.6, pp. 3530–3540. ISSN: 0001-4966. DOI: 10.1121/1.4875714.

Katz, Brian F. G. and Gaëtan Parseihian (2012). "Perceptually based head-related transfer function database optimization". In: *The Journal of the Acoustical Society of America* 131.2, EL99–EL105. ISSN: 0001-4966. DOI: 10.1121/1.3672641.

Kearney, Gavin (2010). "Auditory Scene Synthesis using Virtual Acoustic Recording and Reproduction". PhD thesis. University of Dublin, p. 368.

Kearney, Gavin and Tony Doyle (2015a). "A HRTF Database for Virtual Loudspeaker Rendering". In: *AES 139th Convention*. New York.

– (2015b). "Height Perception in Ambisonic Based Binaural Decoding". In: *AES 139th Convention*. New York.

Kearney, Gavin et al. (2012). "Distance Perception in Interactive Virtual Acoustic Environments using First and Higher Order Ambisonic Sound Fields". In: *Acta Acustica united with Acustica* 98.1, pp. 61–71. DOI: 10.3813/AAA.918492.

Kirkeby, Ole and Philip A. Nelson (1999). “Digital Filter Design for Inversion Problems in Sound Reproduction”. In: *Journal of the Audio Engineering Society* 47.7/8, pp. 583–595. ISSN: 00047554.

Kistler, Doris J. and Frederic L. Wightman (1992). “A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction”. In: *The Journal of the Acoustical Society of America* 1637.91. DOI: 10.1121/1.402444.

Koivuniemi, Kalle and Nick Zacharov (2001). “Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training”. In: *AES 111th Convention*. NY.

Kreuzer, Wolfgang, Piotr Majdak and Zhengsheng Chen (2009). “Fast multipole boundary element method to calculate head-related transfer functions for a wide frequency range”. In: *The Journal of the Acoustical Society of America* 126.3, pp. 1280–1290. ISSN: 0001-4966. DOI: 10.1121/1.3177264.

Kronlachner, Matthias and Franz Zotter (2014). “Spatial transformations for the enhancement of Ambisonic recordings”. PhD thesis. Graz: University of Music and Performing Arts.

Kuhn, George F (1977). “Model for the interaural time differences in the azimuthal plan”. In: *The Journal of the Acoustical Society of America* 157.62. DOI: 10.1121/1.381498.

Langendijk, E. H. and Adelbert W. Bronkhorst (2000). “Fidelity of three-dimensional-sound reproduction using a virtual auditory display.” In: *The Journal of the Acoustical Society of America* 107.1, pp. 528–537. ISSN: 00014966. DOI: 10.1121/1.428321.

Lecomte, Pierre et al. (2015). “On the Use of a Lebedev Grid for Ambisonics”. In: *AES 139th Convention*. New York.

Lecomte, Pierre et al. (2016). “A fifty-node lebedev grid and its applications to ambisonics”. In: *Journal of the Audio Engineering Society* 64.11, pp. 868–881. ISSN: 15494950. DOI: 10.17743/jaes.2016.0036.

Lee, Ki Seung and Seok Pil Lee (2011). “A Relevant Distance Criterion for Interpolation of Head-Related Transfer Functions”. In: *IEEE Transactions on Audio, Speech and Language Processing* 19.6, pp. 1780–1790. ISSN: 15587924. DOI: 10.1109/TASL.2010.2101590.

Lindau, Alexander et al. (2014). “A spatial audio quality inventory (SAQI)”. In: *Acta Acustica united with Acustica* 100.5, pp. 984–994. ISSN: 16101928. DOI: 10.3813/AAA.918778.

Lokki, Tapio et al. (2012). “Disentangling preference ratings of concert hall acoustics using subjective sensory profiles”. In: *The Journal of the Acoustical Society of America* 132.5, pp. 3148–3161. ISSN: 0001-4966. DOI: 10.1121/1.4756826.

Lorho, Gaëtan et al. (2000). “Efficient HRTF synthesis using an interaural transfer function model”. In: *European Signal Processing Conference* 2015-March.March, pp. 80–83. ISSN: 22195491.

Macpherson, Ewan A and John C Middlebrooks (2002). “Listener weighting of cues for lateral angle: The duplex theory of sound localization revisited”. In: *The Journal of the Acoustical Society of America* 111.2219. DOI: 10.1121/1.1471898.

Majdak, Piotr (2013). *ARI HRTF Database.* URL: http://www.kfs.oeaw.ac.at/content/view/608/606/ (visited on 11/22/2019).

Majdak, Piotr, Peter Balazs and Bernhard Laback (2007). “Multiple exponential sweep method for fast measurement of head-related transfer functions”. In: *Journal of the Audio Engineering Society* 55.7-8, pp. 623–636. ISSN: 15494950.

Majdak, Piotr, Matthew J. Goupell and Bernhard Laback (2010). “3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training”. In: *Attention, Perception, & Psychophysics* 72.2. DOI: 10.3758/APP.72.2.454.

Malham, David G. (1992). “Experience with Large Area 3-D Ambisonic Sound Systems”. In: *Institute of Acoustics Autumn Conference on Reproduced Sound 8*. Institute of Acoustics.

– (2003a). *Higher order Ambisonic systems.* Aabstracted from "Space in Music - Music In Space". University of York. URL: http://www.york.ac.uk/inst/mustech/3d_audio/higher_order_ambisonics.pdf.

– (2003b). “Space in Music - Music in Space”. PhD thesis. University of York.

Masiero, Bruno Sanches (2012). “Individualized Binaural Technology. Measurement, Equalization and Perceptual Evaluation”. PhD thesis. URL: http://publications.rwth-aachen.de/record/197486.

Mattila, Ville-Veikko (2001). “Descriptive Analysis of Speech Quality in Mobile Communications: Descriptive Language Development and External Preference Mapping”. In: *AES 111th Convention*, Convention Paper 5455.

McAnally, Ken I and Russell L Martin (2002). “Variability in the Headphone-to-Ear-Canal Transfer Function”. In: *Journal of the Audio Engineering Society* 50.4, pp. 263–266. ISSN: 00047554.

Mckeag, Adam and David McGrath (1996). “Sound Field Format to Binaural Decoder with Head Tracking”. In: *AES 6th Australian Regional Convention*. Melbourne.

Mckenzie, Thomas, Damian Murphy and Gavin Kearney (2017). “Diffuse-Field Equalisation of First-Order Ambisonics”. In: *Proceedings of the 20th International Conference on Digital Audio Effects*, pp. 389–396.

– (2019). “Interaural Level Difference Optimization of Binaural Ambisonic Rendering”. In: *Applied Sciences*. DOI: 10.3390/app9061226.

McKenzie, Thomas, Damian T. Murphy and Gavin Kearney (2018). “Diffuse-Field Equalisation of binaural ambisonic rendering”. In: *Applied Sciences (Switzerland)* 8.10. ISSN: 20763417. DOI: 10.3390/app8101956.

Meng, Q. et al. (2008). “Impulse Response Measurement With Sine Sweeps and Amplitude Modulation Schemes”. In: *IEEE 2nd International Conference on Signal Processing and Communication Systems*. Gold Coast. ISBN: 9781424442423. DOI: 10.1109/ICSPCS.2008.4813749.

Menzies, Dylan and Marwan Al-Akaidi (2007). “Nearfield binaural synthesis and ambisonics”. In: *The Journal of the Acoustical Society of America* 121.3, p. 1559. ISSN: 00014966. DOI: 10.1121/1.2434761.

Meyer, Jens (2001). “Beamforming for a circular microphone array mounted on spherically shaped objects”. In: *The Journal of the Acoustical Society of America* 109.1, pp. 185–193. ISSN: 0001-4966. DOI: 10.1121/1.1329616.

Meyer, Jens and Gary Elko (2002). “A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield”. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 1781–1784. ISSN: 15206149. DOI: 10.1109/ICASSP.2002.5744968.

Middlebrooks, John C, James C. Makous and David M. Green (1989). “Directional sensitivity of sound-pressure levels in the human ear canal”. In: *Journal of the Acoustical Society of America* 86.1, pp. 89–108. ISSN: NA. DOI: 10.1121/1.398224.

Mills, A. W. (1958). “On the Minimum Audible Angle”. In: *The Journal of the Acoustical Society of America* 30.237. DOI: 10.1121/1.1909553.

– (1960). “Lateralization of High-Frequency Tones”. In: *The Journal of the Acoustical Society of America* 32.132. DOI: 10.1121/1.1907864.

Møller, Henrik (1992). “Fundamentals of binaural technology”. In: *Applied Acoustics* 36.3-4, pp. 171–218. ISSN: 0003682X. DOI: 10.1016/0003-682X(92)90046-U.

Møller, Henrik and M. F. Sørensen (1996). “Binaural technique: Do we need individual recordings?” In: *Journal of the Audio Engineering Society* 44.6, pp. 451–469. ISSN: 0004-7554.

Møller, Henrik et al. (1995). “Head-Related Transfer-Functions of Human-Subjects”. In: *Journal of the Audio Engineering Society* 43.5, pp. 300–321.

Monro, Gordon (2000). “In-phase corrections for Ambisonics”. In: *International Computer Music Conference*. Berlin.

Moore, Brian C. J. and Brian R. Glasberg (1995). “A Revision of Zwicker’s Loudness Model”. In: *Acta Acustica united with Acustica* 82.2, pp. 335–345.

Moore, Brian C. J., Brian R. Glasberg and Thomas Baer (1997). “A Model for the Prediction of Thresholds, Loudness, and Partial Loudness”. In: *Journal of the Audio Engineering Society* 45.4.

Moreau, Sébastien, Jérôme Daniel and Stéphanie Bertet (2006). “3D Sound field recording with higher order Ambisonics – objective measurements and validation of a 4th order spherical microphone”. In: *AES 120th Convention*.

Morse, Philip McCord and Uno Ingard (1968). *Theoretical Acoustics*. Princeton University Press, p. 949.

NCAR (2018). *Description of Gaussian, fixed, fixed offset, regular, curvilinear grids*. DOI: http://dx.doi.org/10.5065/D6WD3XH5. URL: http://www.ncl.ucar.edu/Document/Functions/sphpk_grids.shtml{\#}GaussianGrids (visited on 11/22/2019).

Neuman. *KU100 Operating instructions*.

Nicol, Rozenn et al. (2014). “A Roadmap for Assessing the Quality of Experience of 3D Audio Binaural Rendering”. In: *EAA Joint Symposium on Auralization and Ambisonics*, pp. 100–106.

Noisternig, Markus, Franz Zotter and Brian F. G. Katz (2011). “Reconstructing sound source directivity in virtual acoustic environments”. In: *Principles and Applications of Spatial Hearing*. Ed. by H. Kato, D. S. Brungart and Y. Suzuki. World Scientific Publishing Co. Pte. Ltd., pp. 357–373.

Noisternig, Markus et al. (2003a). “A 3D ambisonics based binaural sound reproduction system”. In: *AES 24th International conference on Multichannel Audio.* Banff.

Noisternig, Markus et al. (2003b). “A 3D Real Time Rendering Engine for Binarual Sound Reproduction”. In: *International Conference on Auditory Display.* Boston.

Ortolani, Francesca (2015). *Introduction to Ambisonics (Rev. 2015).* Tech. rep. Ironbridge Electronics. URL: https://www.academia.edu/6931506/Introduction_to_Ambisonics_Rev._2015_.

Ospina, Felipe Rugeles, Marc Emerit and Brian F.G. Katz (2015). “The three-dimensional morphological database for spatial hearing research of the BiLi project”. In: *Proceedings of Meetings on Acoustics.* Vol. 23. DOI: 10.1121/2.0000050.

Otani, Makoto, Tatsuya Hirahara and Shiro Ise (2009). “Numerical study on source-distance dependency of head-related transfer functions Numerical study on source-distance dependency of head-related”. In: *The Journal of the Acoustical Society of America* 125.5. DOI: 10.1121/1.3111860.

Paul, Stephan (2009). “Binaural recording technology: A historical review and possible future developments”. In: *Acta Acustica united with Acustica* 95, pp. 767–788. ISSN: 16101928. DOI: 10.3813/AAA.918208.

Pearce, Andy, Tim Brookes and Russell Mason (2017). “Timbral attributes for sound effect library searching”. In: *AES Conference on Semantic Audio.* Erlangen. DOI: 10.5281/zen-odo.167392.2.2.

Penrose, R. and J. A. Todd (1955). “A generalized inverse for matrices”. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 51.3, p. 406. ISSN: 0305-0041. DOI: 10.1017/S0305004100030401.

Perrett, Stephen and William Noble (1997). “The contribution of head motion cues to localization of low-pass noise”. In: *Perception and Psychophysics* 59.7, pp. 1018–1026. ISSN: 00315117. DOI: 10.3758/BF03205517.

Pfluger, Martin (1997). “Modelle des peripheren Gehors am Beispiel der menschlichen Lautheitsempfindung”. PhD thesis. Technischen Universitat Graz.

Pike, Chris and Frank Melchior (2013). *An Assessment of Virtual Surround Sound Systems for Headphone Listening of 5.1 Multichannel Audio.* Tech. rep. BBC.

Pike, Chris, Frank Melchior and Anthony I. Tew (2016). “Descriptive analysis of binaural rendering with virtual loudspeakers using a rate-all-that-apply approach”. In: *AES Conference on Headphone Technology.* Aalborg.

Plaga, John A. et al. (2005). *Design and Development of Anthropometrically Correct Head Forms for Joint Strike Fighter Ejection Seat Testing.* Tech. rep. Air Force Research Laboratory.

Poletti, Mark (1996). "The Designof Encoding Functions for Stereophonic and Polyphonic Sound Systems". In: *Journal of the Audio Engineering Society* 44.11, pp. 948 –963.

– (2000). "A Unified Theory of Horizontal Holographic Sound Systems". In: *Journal of the Audio Engineering Society* 48.12, pp. 1155–1182. ISSN: 00047554.

Pollow, Martin et al. (2012). "Calculation of head-related transfer functions for arbitrary field points using spherical harmonics decomposition". In: *Acta Acustica united with Acustica* 98, pp. 72–82. ISSN: 16101928. DOI: 10.3813/AAA.918493.

Pulkki, Ville (1997). "Virtual Sound Source Positioning Using Vector Based Amplitude Panning". In: *Journal of the Audio Engineering Society* 45.6, pp. 456–466.

Pulkki, Ville, Matti Karjalainen and Jyri Huopaniemi (1999). "Analyzing virtual sound source attributes using a binaural auditory model". In: *Journal of the Audio Engineering Society* 47.4, pp. 203–217. ISSN: 00047554.

Qualcomm (2015). *Scene based audio: A novel paradigm for immersive and interactive audio user experience.* Tech. rep. Qualcomm Technologies Inc.

Rafaely, Boaz (2005). "Analysis and design of spherical microphone arrays". In: *IEEE Transactions on Speech and Audio Processing* 13.1, pp. 135–143. ISSN: 10636676. DOI: 10.1109/TSA.2004.839244.

– (2008). "The Spherical-Shell Microphone Array". In: *IEEE Transactions on Audio, Speech and Language Processing* 16.4.

Rahim, T. and D. E.N. Davies (1982). "Effect of Directional Elements on the Directional Response of Circular Antenna Arrays". In: *IEE Proceedings H: Microwaves Optics and Antennas* 129.1, pp. 18–22. ISSN: 01437097. DOI: 10.1049/ip-h-1.1982 .0004.

Richter, Jan-gerrit et al. (2014). "Spherical Harmonics Based HRTF Datasets : Implementation and Evaluation for Real-Time Auralization". In: *Acta Acustica united with Acustica* 100, pp. 667–675. DOI: 10.3813/AAA.918746.

Rix, A. W. et al. (2001). "Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing.* Salt Lake

City. ISBN: 0-7803-7041-4. DOI: 10.1109/ICASSP.2001.941023. arXiv: arXiv:1011.1669v3.

Rudzki, Tomasz et al. (2019). “Auditory Localization in Low-Bitrate Compressed Ambisonic Scenes”. In: *Applied Sciences* 2019.9. DOI: 10.3390/app9132618.

Rumsey, Francis (2014). “Binaural challenges: Spatial audio”. In: *Journal of the Audio Engineering Society* 62.11, pp. 798–802. ISSN: 15494950.

Schärer, Zora and Alexander Lindau (2009). “Evaluation of Equalization Methods for Binaural Signals”. In: *AES 126th Convention*, p. 17.

Schönstein, David and Brian F. G. Katz (2012). “Variability in Perceptual Evaluation of HRTFs”. In: *Journal of the Audio Engineering Society* 60.10, pp. 783–793. ISSN: 15494950.

Schörkhuber, Christian, Markus Zaunschirm and H Robert (2018). “Binaural Rendering of Ambisonic Signals via Magnitude Least Squares”. In: *Fortschritter der Akustik (DAGA)*.

Schroeder, M. R. (1979). “Integrated-impulse method measuring sound decay without using impulses”. In: *The Journal of the Acoustical Society of America* 66.2, pp. 497–500. ISSN: NA. DOI: 10.1121/1.383103.

Searle, C. L. et al. (1975). “Binaural pinna disparity: another auditory localization cue”. In: *Journal of the Acoustical Society of America* 57.448.

Seeber, Bernhard U, Hugo Fastl and Others (2003). “Subjective selection of non-individual head-related transfer functions”. In: *International Conference on Auditory Display*.

Simon, Laurent S. R., Nick Zacharov and Brian F. G. Katz (2016). “Perceptual attributes for the comparison of head-related transfer functions”. In: *The Journal of the Acoustical Society of America* 140.5, pp. 3623–3632. ISSN: 0001-4966. DOI: 10.1121/1.4966115.

Smith, Julius O and Jonathan S Abel (1999). “Bark and ERB Bilinear Transforms”. In: *IEEE Transactions on Speech and Audio Processing* 7.6, pp. 697–708. DOI: 10.1109/89.799695.

Smyth, Mike and Stephen Smyth (2016). “Headphone Virtualisation for Immersive Audio Monitoring”. In: *AES 140th Convention*. Paris.

Spagnol, Simone (2015). “On distance dependence of pinna spectral patterns in head-related transfer functions”. In: *The Journal of the Acoustical Society of America* 137.1. DOI: 10.1121/1.4903919.

Stern, Richard M., Guy J. Brown and DeLiang Wang (2005). “Binaural sound localization”. In: *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Ed. by DeLiang Wang and Guy J. Brown. John Wiley & Sons, Inc. Chap. 5. ISBN: 978-0-471-74109-1.

Stevens, S. S. (1936). “A scale for the measurement of a psychological magnitude: loudness”. In: *The psychological review* 43.5.

– (1955). “The Measurement of Loudness”. In: *The Journal of the Acoustical Society of America* 27.5, pp. 815–829. DOI: 10.1121/1.1908048.

Stitt, Peter (2015). “Ambisonics and Higher-Order Ambisonics for Off-Centre Listeners : Evaluation of Perceived and Predicted Image Direction”. PhD thesis. Queen’s University Belfast.

Stitt, Peter, Lorenzo Picinali and Brian F. G. Katz (2019). *Auditory Accommodation to Poorly Matched Non-Individual Spectral Localization Cues Through Active Learning*. Tech. rep. 9. DOI: 10.1038/s41598-018-37873-0.

Strutt (Lord Rayleigh), John William (1907). “On our perception of sound direction”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 13.74. DOI: 10.1080/14786440709463595.

Takemoto, Hironori et al. (2012). “Mechanism for generating peaks and notches of head-related transfer functions in the median plane”. In: *The Journal of the Acoustical Society of America* 132.6. DOI: 10.1121/1.4765083.

Theile, Gunther (1984). “The Dummy Head - theory and Practice”. In: *13th Tonmeistertagung*. IRT Munich.

Thiede, Thilo et al. (2000). “PEAQ– The ITU standard for objective measurement of perceived audio quality”. In: *Journal of the Audio Engineering Sociaty* 48.1/2, pp. 3–29. ISSN: 0004-7554.

Thresh, Lewis, Cal Armstrong and Gavin Kearney (2017). “A Direct Comparison of Localisation Performance When Using First, Third and Fifth Order Ambisonics For Real Loudspeaker And Virtual Loudspeaker Rendering”. In: *AES 143rd Convention*.

Treviño, Jorge et al. (2011). “Evaluation of a New Ambisonic Decoder for Irregular Loudspeaker Arrays Using Interaural Cues”. In: *Ambisonics Symposium*. Lexington.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading MA: Addison-Wesley.

Usher, John and William L Martens (2007). "Perceived Naturalness Of Speech Sounds Presented Using Personalized Versus Non-personalized HRTFs". In: *13th International Conference on Auditory Display*, pp. 10–16.

Vorländer, Michael (2008). *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality.* First edit. Springer. DOI: 10.1007/978-3-540-48830-9.

Wabnitz, Andrew, Nicolas Epain and A McEwan (2011). "Upscaling Ambisonic sound scenes using compressed sensing techniques". In: *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. New Paltz. ISBN: 9781457706936. DOI: 10.1109/ASPAA.2011.6082301.

Wang, Shihua, Andrew Sekey and Allen Gersho (1992). "An Objective Measure for Predicting Subjective Quality of Speech Coders". In: *IEEE Journal on Selected Areas in Communications* 10.5, pp. 819–829. ISSN: 07338716. DOI: 10.1109/49.138987.

Ward, Darren B. and Thushara D. Abhayapala (2001). "Reproduction of a plane-wave sound field using an array of loudspeakers". In: *IEEE Transactions on Speech and Audio Processing* 9.6, pp. 697–707. ISSN: 10636676. DOI: 10.1109/89.943347.

Watanabe, Kanji et al. (2014). *Dataset of head-related transfer functions measured with a circular loudspeaker array.* Tech. rep. 3. DOI: 10.1250/ast.35.159. URL: http://www.riec.tohoku.ac.jp/pub/hrtf/index.html.

Watanabe, Kanji et al. (2016). "Estimation of interaural level difference based on anthropometry and its effect on sound localization". In: *The Journal of the Acoustical Society of America* 122.5. DOI: 10.1121/1.2785039.

Wenzel, Elizabeth M. et al. (1993). "Localization using nonindividualized head-related transfer functions". In: *The Journal of the Acoustical Society of America* 94.1, pp. 111–123. ISSN: 0001-4966. DOI: 10.1121/1.407089.

Wiggins, Bruce (2004). "An Investigation into the Real-Time Manipulation and Control of Three-Dimentional Sound Fields". PhD thesis. University of Derby.

Wightman, Frederic L. and Doris J. Kistler (1989a). "Headphone simulation of free field listening I: stimulus synthesis". In: *The Journal of the Acoustical Society of America* 85.1989, pp. 858–867. DOI: 10.1121/1.397557.

– (1989b). "Headphone simulation of free-field listening. II: Psychophysical validation". In: *The Journal of the Acoustical Society of America* 85.2, pp. 868–878. ISSN: 0001-4966. DOI: 10.1121/1.397558.

– (1999). “Resolution of front–back ambiguity in spatial hearing by listener and source movement”. In: *The Journal of the Acoustical Society of America* 105.5, pp. 2841–2853. ISSN: 0001-4966. DOI: 10.1121/1.426899.

Woodworth, R. S. (1938). *Experimental Psychology.* New York: Holt.

Xie, Bosun (2013). *Head-related transfer function and virtual auditory display.* Second edi. J. Ross Publishing, pp. 117–118. ISBN: 978-1-60427-070-9.

Yao, Shu-Nung, Tim Collins and Peter Jancovic (2015). “Timbral and spatial fidelity improvement in ambisonics”. In: *Applied Acoustics* 93, pp. 1–8. ISSN: 1872910X. DOI: 10.1016/j.apacoust.2015.01.005.

Yost, William A. (2000). *Fundamentals of Hearing: An Introduction.* San Diego: Academic Press. Chap. 10, pp. 149–164.

Young, Kat, Gavin Kearney and Anthony I. Tew (2018a). “Acoustic Validation of a BEM-Suitable 3D Mesh Model of KEMAR”. In: *AES Conference on Spatial Reproduction.*

– (2018b). “Loudspeaker Positions with Sufficient Natural Channel Separation for Binaural Reproduction”. In: *AES Conference on Spatial Reproduction.* Vol. 2.

Young, Kat, Anthony I. Tew and Gavin Kearney (2016). “Boundary element method modelling of KEMAR for binaural rendering : Mesh production and validation”. In: *Interactive Audio Systems Symposium.*

Young, Kat et al. (2019). “A Numerical Study into Perceptually-Weighted Spectral Differences between Differently-Spaced HRTFs”. In: *AES Conference on Immersive and Interactive Audio.*

Yu, Guang-zheng, Bo-sun Xie and Dan Rao (2010). “Characteristics of Near-Field Head-Related Transfer Function for KEMAR”. In: *AES 40th International Conference*, pp. 1–8.

Zaunschirm, Markus, Christian Schörkhuber and Robert Höldrich (2018). “Binaural rendering of Ambisonic signals by head-related impulse response time alignment and a diffuseness constraint”. In: *The Journal of the Acoustical Society of America* 143.6, pp. 3616–3627. ISSN: 0001-4966. DOI: 10.1121/1.5040489.

Zmölnig, Johannes M. (2002). “Entwurf und Implementierung einer Mehrkanal-Beschallungsanlage”. PhD thesis. kunstuniversität graz.

Zölzer, Udo, ed. (2011). *DAFX: Digital Audio Effects.* John Wiley & Sons. Chap. 2, pp. 50–55.

Zotkin, Dmitry N. et al. (2006). “Fast head-related transfer function measurement via reciprocity”. In: *The Journal of the Acoustical Society of America* 120.4, pp. 2202–2215. ISSN: 0001-4966. DOI: 10.1121/1.2207578.

Zotter, Franz (2009). “Analysis and synthesis of sound-radiation with spherical arrays”. PhD thesis. University of Music and Performing Arts.

Zotter, Franz and Matthias Frank (2012). “All-round ambisonic panning and decoding”. In: *Journal of the Audio Engineering Society* 60.10, pp. 807–820. ISSN: 15494950.

Zotter, Franz, Matthias Frank and Alois Sontacchi (2010). “The virtual t-design ambisonics-rig using vbap”. In: *EAA EuroRegio*. Ljublijana.

Zotter, Franz, Hannes Pomberger and M Frank (2009). “An alternative ambisonics formulation: Modal source strength matching and the effect of spatial aliasing”. In: *AES 126th Convention*. Munich. ISBN: 9781615671663.

Zotter, Franz, Hannes Pomberger and Markus Noisternig (2010). “Ambisonic Decoding With and Without Mode-Matching - A Case Study Using the Hemisphere”. In: *2nd International Symposium on Ambisonics and Spherical Acoustics*. Paris.

– (2012). “Energy-preserving ambisonic decoding”. In: *Acta Acustica united with Acustica* 98.1, pp. 37–47. ISSN: 16101928. DOI: 10.3813/AAA.918490.

Zwicker, Eberhard (1961). “Subdivision of the Audible Frequency Range into Critical Bands”. In: *The Journal of the Acoustical Society of America* 33.2. DOI: 10.1121/1.1908630.

Zwicker, Eberhard and Bertram Scharf (1965). “A Model of Loudness Summation”. In: *The psychological review* 72.1.