Binaural Impulse Response Rendering for Immersive Audio

A dissertation submitted to the University of Dublin
for the degree of Doctor of Philosophy

Claire Masterson
Trinity College Dublin, October 2010

Department of Electronic and Electrical Engineering
Trinity College Dublin

To my family and friends.

Declaration

I hereby declare that this thesis has not been submitted as an exercise for a degree at this or any other University and that it is entirely my own work.

I agree that the Library may lend or copy this thesis upon request.

Signed,

Claire Masterson

October 28, 2010.

Abstract

This thesis is concerned with the efficient reproduction of immersive and realistic 3D virtual auditory environments. This necessitates the correct virtual reproduction of the soundfield incident at both ears using signal processing tools. In order to mimic real world stimuli an equivalent sense of depth and directionality must be accorded to the listener which should be particular to the nature of the virtual listening environment, source-listener positioning and the size and shape of the listener's outer ear, head and torso.

A significant contribution of this thesis is a technique for the approximate factorisation of Head Related Impulse Responses (HRIRs). HRIRs describe the filtering caused by the outer ear, head and torso on source audio coming from a particular position in space. Commonly large datasets of these responses are measured for each ear for a dense grid of source positions. The factorisation allows for a direction independent common component to be extracted, or deconvolved, from a large number of HRIRs i.e. each individual filter is split into two filters which, when convolved together, result in a close approximation to the original filter. Hence convolution of the source audio with the common factor could be completed offline and stored leaving a shorter HRIR that would change with a relative movement between source and receiver in real time. This technique is extended to include two regularisation options applicable to either minimum phase or non minimum phase datasets and which allow for a more robust, initial condition independent factorisation process.

The anechoic nature of HRIR measurement means that the reverberant properties of the room are not included. This reverberance is measured in the Room Impulse Response (RIR) and its inclusion in a virtual audio reproduction is key for listener envelopment and the impartment of depth to a virtual source. Large grids of such RIRs are required in a fully interactive 'walk through' situation with full freedom accorded to the listener. This is further increased if the source position is not static. A novel spatial interpolation technique is presented here to reduce the measurement burden. The response is divided into an early reflection component and a diffuse decay component. A dynamic time warping based algorithm is used for the spatial interpolation of the early reflection part. The diffuse tail is synthesised from a measured tail from the available measured set which is decomposed into critical bands. These bands are then randomly time shifted to achieve decorrelation from the existing measured tail.

The combination of the anechoic HRIR and the reverberant RIR is examined and a real time implementation is proposed based on Ambisonic theory. This implementation, using the Fmod API, incorporates headtracking technology and allows for the creation of virtual loudspeakers which remain static in space even with listener head movement. This is made possible by rotation of the Ambisonic soundfield as opposed to HRIR switching which can lead to undesirable noise artifacts. The possible benefits of spatial audio to e-learning are investigated. The ReciTell Player is introduced and the application of the virtual loudspeaker system to it is examined.

Acknowledgments

Firstly I would like to thank my supervisor Prof Frank Boland, whose encouragement led me to embark upon this PhD and without whose guidance, support and good humour I would almost certainly never have finished it.

I also owe a debt of gratitude to Dr Gavin Kearney with whom I collaborated for portions of this work. His seemingly unending knowledge of audio reproduction techniques has been a huge help and inspiration. Thanks also to Stephen and Marcin for their help.

This work could not have been completed without the financial support of the Irish Research Council or Science Engineering and Technology (IRCSET).

A big thank you to all the support staff in the Department of Electronic and Electrical Engineering . Also a big shout out must go the whole DSP crew in Aras an Phiarsaigh, especially the inhabitants of Rooms 2.07, 2.18 and 2.19, for the support, welcome distraction and comedic interludes they provided to my postgrad experience. You know who you are!

Thanks to all my family and friends for their love and encouragement throughout these three years and beyond.

Lastly, thanks to Craig, whose understanding and support has kept me on an even keel.

Contents

Contents **iv**

List of Acronyms **vii**

1 Introduction **1**
1.1 Introduction . . . 1
1.2 Thesis outline . . . 5
1.3 Contributions of this Thesis . . . 7
1.4 Publications . . . 7

2 Background **9**
2.1 Human Auditory system . . . 9
2.2 Sound localisation . . . 11
2.2.1 Interaural Time Difference . . . 11
2.2.2 Interaural Level Difference . . . 14
2.2.3 Spectral cues . . . 14
2.2.4 Dynamic Cues . . . 16
2.2.5 Other cues . . . 17
2.3 Neurological Processing of Localisation Cues . . . 17
2.4 Spatial Resolution of Localisation . . . 19
2.5 Environmental Cues . . . 20
2.5.1 Measurement and Synthesis of RIRs . . . 22
2.6 Conclusion . . . 24

3 Head Related Impulse Response **25**
3.1 Physiological Basis . . . 25
3.2 Measurement and Modelling . . . 27
3.2.1 Individual Nature of the HRTF . . . 28
3.2.2 Equalisation . . . 29
3.3 The Minimum Phase Assumption . . . 30
3.4 HRTF Interpolation . . . 32

3.5 HRTF Reduction Techniques . . . 35
3.6 Conclusion . . . 37

4 HRIR Order Reduction using Approximate Factorisation 38
4.1 Introduction . . . 38
4.2 The Algorithm . . . 39
4.2.1 Convergence . . . 40
4.2.2 Results of Tests on CIPIC Database . . . 40
4.3 Regularisation . . . 48
4.3.1 Results when Applied to KEMAR HRTF Data . . . 49
4.3.2 Investigation of Regularisation using Human Data from CIPIC Database 57
4.4 Conclusion . . . 63

5 Room Response Modelling 64
5.1 Existing RIR Interpolation Techniques . . . 65
5.2 Transition Point Determination . . . 65
5.3 Dynamic Time Warping . . . 66
5.4 Application of DTW to RIR Interpolation . . . 68
5.5 Tail Synthesis . . . 73
5.6 Application of Interpolation to Wave Field Synthesis Reproduction . . . 74
5.6.1 Introduction to Wave Field Synthesis . . . 74
5.6.2 RIR Capture . . . 76
5.6.3 Implementation of Wave Field Synthesis . . . 79
5.6.4 Objective Analysis . . . 80
5.6.5 Perceptual Analysis . . . 81
5.7 Examination of Possible DTW Improvements . . . 85
5.8 Conclusion . . . 87

6 Virtual Auditory Environments 88
6.1 Binaural Room Impulse Response . . . 88
6.2 Ambisonics . . . 90
6.2.1 B-format Recording and Encoding . . . 90
6.2.2 B-Format Decoding . . . 92
6.2.3 Decoder Optimisation . . . 94
6.2.4 Near Field Effect . . . 94
6.2.5 Rotating the Soundfield . . . 95
6.2.6 Higher Order Ambisonics . . . 96
6.3 Virtual Loudspeaker Approach . . . 97
6.3.1 Application of HRIR Factorisation to VLA . . . 98
6.3.2 Application of RIR Interpolation to VLA . . . 99

6.4 Real-time Implementation 100
6.4.1 Fmod 108
6.5 Conclusion 110

7 Application to E-learning **111**
7.1 Application to E-learning 111
7.2 Applicability of Spatial Audio to E-learning 111
7.3 ReciTell 112
7.4 Implementation 114
7.5 Conclusion 115

8 Conclusion **117**
8.1 Future Work 119
8.2 Final Remarks 121

A Appendix **122**

Bibliography **126**

List of Acronyms

ASW Apparent Source Width

BMT Balanced Model Truncation

BRIR Binaural Room Impulse Response

DFE Diffuse Field Equalisation

DTW Dynamic Time Warping

FFE Free Field Equalisation

HOA Higher Order Ambisonics

HRIR Head Related Impulse Response

HRTF Head Related Transfer Function

IACC Interaural Cross Correlation Coefficient

IACF Interaural Cross Correlation Function

ICA Independent Component Analysis

ITD Interaural Time Difference

ILD Interaural Level Difference

IID Interaural Intensity Difference

IPD Interaural Phase Difference

JND Just Noticeable Difference

KLE Karhunen Loeve Expansion

MAA Minimum Audible Angle

MAMA Minimum Audible Movement Angle

MLS Maximum Length Sequence

PCA Principal Component Analysis

RIR Room Impulse Response

SVD Singular Value Decomposition

VAE Virtual Auditory Environment

VLA Virtual Loudspeaker Approach

WFS Wave Field Synthesis

1 Introduction

1.1 Introduction

This thesis is concerned with the efficient reproduction of immersive and realistic 3D virtual auditory environments. The growing recognition of the importance of accurately spatialised audio to the creation of a convincing multimodal display has led to its increased demand in applications such as gaming, conferencing and e-learning. However, such auditory scene synthesis demands that the soundfield incident at the ears, presented either over headphones or loudspeakers, must be identical, or as close as possible, to signals resulting from real world stimuli. Various spatial audio reproduction techniques have been devised in an attempt to achieve this and impart an equivalent sense of depth and directionality to a virtual reproduction.

Spatial audio experimentation began in the 19th century, an example of this being Ader's remote display of stereophonic sound at the 1881 Paris Expedition of Electricity [52]. Headphones were fed by telephone signals coming from microphones placed at different locations across the stage at the Paris Opera, several miles away. The invention and adaption of spatial audio began in earnest with Blumlein's far reaching work in the early 1930s on stereo reproduction and recording which rightly positions him as the father of spatial audio [19]. Bell Labs also completed important work in the area at this time including adding a third centre channel to the two channel stereo apparatus [167]. Another important addition to the field was Gerzon's introduction of Ambisonic theory in early 70's [60]. This theory is a major influence on this thesis and, as with Blumlein's work on stereo, proved to be significantly ahead of its time. It offers a method for recording and reproduction of the complete soundfield and is independent of the layout of

the reproduction array. While there was some sporadic adaptation of spatial audio in cinema in the decades following Blumlein's seminal work, most notably in Disney's Fantasia, it was not until Star Wars release in the late 70's that cinematic spatial audio, implemented by Dolby Laboratories, gained mainstream acceptance. 5.1 channel surround followed and this and stereo have become a standard formats for domestic media consumption. For a more comprehensive overview of the history of spatial audio see Davis's 2003 paper on the topic [43].

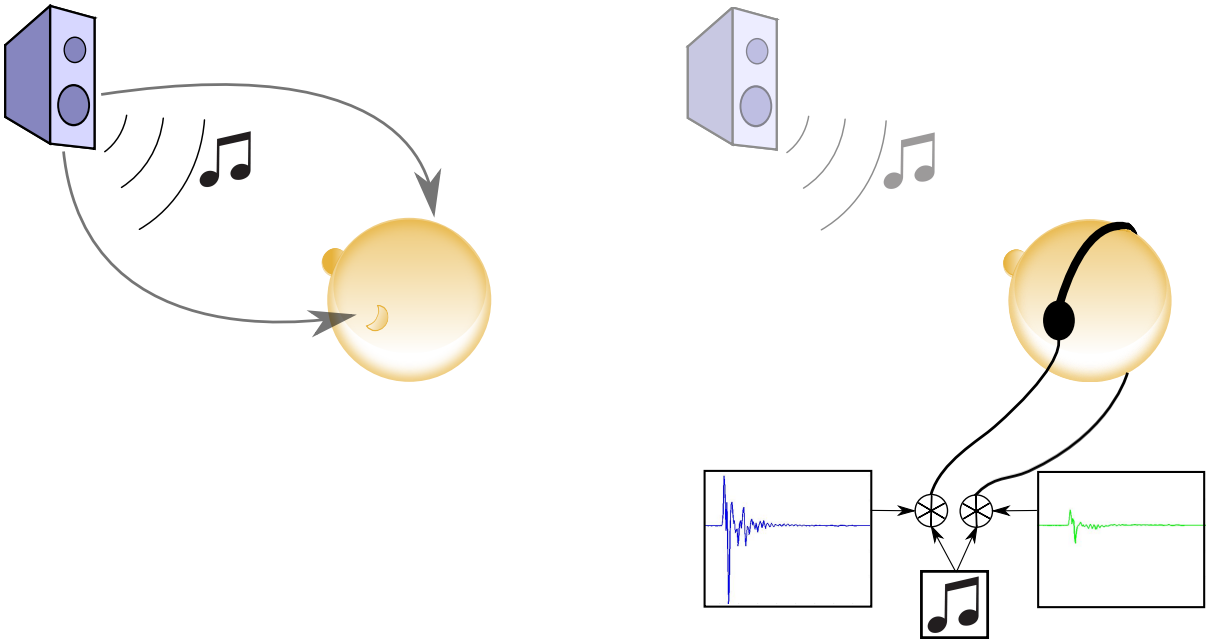
Spatial audio reproduction over loudspeakers has been further extended from the 5.1 channel system to 7.1 channel, 10.2 channel and even Wave Field Synthesis (WFS) which aims to recreate the correct sound wavefront using large arrays of speakers. However there has been increasing interest in personal audio presentation via headphone listening in the past number of years. Headphone reproduction offers increased privacy and immersiveness, is more immune to external interruption in noisy environments and does not suffer from the 'sweet spot' limitations of most loudspeaker reproduction techniques. Simple stereo panning remains popular for this medium but there is an increasing shift towards the use of dummy head microphones (see Figure 1.1) for recording and the filtering of mono audio with Head Related Impulse Responses (HRIRs) (see Figure 1.2) to offer more spatially accurate and immersive headphone feeds. Møller characterises this head related filtering well in this extract from his seminal paper 'Fundamentals of Binaural Technology' [124].

> "A sound wave coming from a particular direction and distance results in two sound pressures, one at each eardrum. The transmissions are described in terms of two transfer functions that include any linear distortion, such as coloration and interaural time and spectral differences. The task of a binaural recording and playback system is to present the correct inputs to the hearing, that is to reproduce the eardrum signals correctly. In this connection, it is not important how the hearing extracts information from the eardrum signals about distance and direction." Møller, H., Fundamentals of binaural technology, Applied Acoustics, 36(3&4):171-128.

A disadvantage of this medium over loudspeaker reproduction is that the fact that the headphones move with the listener and, as such, the virtual source position does not remain stable and constant in real world co-ordinates. One solution to this problem is to track the motion of the head and compensate the audio feed suitably. This solution is explored in Chapter 6 of this thesis.

In the real world however, humans rarely experience an anechoic environment. As such there is further filtering of sound, other than that caused by the ear, head and body, to be considered. The environment, in which the source and receiver are positioned, influences the sound field depending on its particular geometry and the material composition of reflecting surfaces. For a given source-receiver position and orientation in an environment this filtering is termed the Room Impulse Response (RIR). The modulation of an anechoic recording with a room impulse response is referred to as 'auralisation' (see Figure 1.3). The room response may be obtained

Figure 1.1: Neumann dummy head microphone.



(a) Real listening.

(b) Artificially spatialised audio using HRIRs.

Figure 1.2: Binaural Reproduction of Spatial Audio.

by measurement or synthesis. The term is introduced by Kleiner et al. [99] who coin it to be analogous to visualisation and describe it as the method of simulating the correct binaural rendering of a soundfield for a given source and receiver position in a given environment through the use of "*physical or mathematical modeling*". Vorländer [174] extends and generalises this definition somewhat and refers to auralisation as

> "the technique of creating audible sound files from numerical (simulated, **measured**, or synthesized) data." Vorländer, M., Auralization: Fundamentals of acoustics,

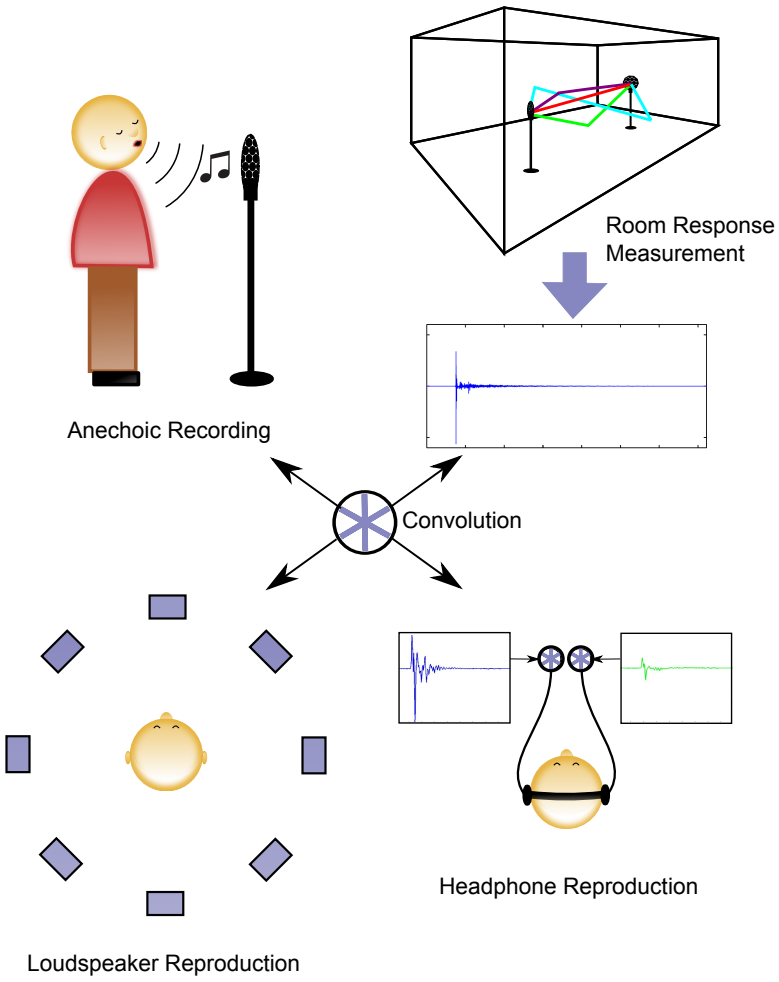


Figure 1.3: Auralisation.

modelling, simulation, algorithms and acoustic virtual reality.

The inclusion of measured responses as an auralisation technique is noteworthy and this is the form in which auralisation is considered in this thesis.

The real time implementations of such an auralisation system vary greatly in their complexity and hardware requirements. In 2009 Okamoto et al. [139] implemented an Ambisonics based system using a 3D grid of 157 loudspeakers (with speakers on the walls and ceiling of a damped room), a listener tracking system and 5 computers. Obviously such a setup is beyond the reach of most consumers! Wave field synthesis similarly requires dense loudspeaker arrays in order to produce compelling results [20]. The minimal hardware outlay required for headphone reproduction makes it an attractive alternative to such loudspeaker based systems. Of course tracking of head movement is required in order to compensate the audio feed and maintain a stable auralisation. Unfortunately the naive use of head related impulse responses or binaural room impulse responses directly in conjunction with head tracking can often lead to audible

glitches or clicks in the audio as the filters are changed due to a relative movement between the virtual source and listener. An interesting approach to addressing this problem is through the creation a number of virtual loudspeaker feeds using a small number of head related impulse responses and the use of Ambisonics decoding and to rotate the soundfield, instead of switching filters. This approach will be examined in this thesis.

There is a wide spectrum of applications for spatial audio in the real world which are only beginning to be examined and implemented. Firstly there is the obvious and lucrative interactive gaming market. In the past year a product called 'MyEars' [133] has come on the market offering personalised 3D headphone audio to gamers. It claims to offer perfect 7.1 surround reproduction over headphones with an initial personalisation stage based on user responses to test signals. Teleconferencing is also a popular application of spatial audio as it is thought to improve the intelligibility and comprehension of speech. VSpace [175] is a recent offering in this area and offers an online VOIP client which is compatible with Windows, Mac and Linux as well as portable devices like the iPhone or Nokia N-series. It allows the listener to hear the other participants' high quality audio feeds in spatially separate locations around the head, which the developers contend allows for a much more natural discourse than normal phone conversations. While the proprietary nature of the product dictates that details of the system are not available, it would appear that the increase of the maximal transmission frequency to beyond 7kHz (from 3.5kHz in similar telephony applications), the use of head related filtering and the addition of room reverberation are important factors in this implementation. Japanese telecommunications company NTT DoCoMo are developing a similar product for mobile phone users, allowing unique directions to be assigned to different speakers [44]. Auditory displays of information have considerable applicability to the visually impaired. A recent paper by Nagasaka et al. [134] demonstrates the implementation of the game Reversi with information being presented through auditory and tactile media. White et al. [180] explore some techniques, both audio and haptic related, that can be used to create accessible 3D virtual environments for the visually impaired.

The work in this thesis is motivated by the need to make the auralisation of audio more efficient, both in its real time computational and memory requirements and in the preparation time required to acquire necessary data and measurements.

1.2 Thesis outline

The remainder of this thesis is organised as follows.

Chapter 2: Background

This chapter introduces the auditory perception basis for the research including a brief description of the human auditory system and of the main spatial localisation cues. The neurological processing of these cues and the spatial resolution they offer is explored and the influence of the

room environment on localisation is examined.

Chapter 3: Head Related Impulse Response

Here head related impulse responses are explored in detail with a comprehensive review of existing literature on their physiological basis and techniques for their measurement and synthesis. The individual and varying nature of these filters is examined and existing methods for their interpolation and compression are investigated.

Chapter 4: HRIR Order Reduction using Approximate Factorisation

In this chapter, a novel method is proposed that allows for the extraction of a common component from large datasets of head related impulse responses. This technique is extended to include two regularisation options applicable to either minimum phase or non minimum phase datasets and which allow for a more robust, initial condition independent factorisation process. These algorithms are extensively tested on HRIR data from several datasets and comprehensive results are shown.

Chapter 5: Room Response Modelling

The composition of the room impulse response is reviewed and techniques for its division into its constituent early reflection and diffuse decay parts are explored. A novel approach for the interpolation of the early reflection components using a Dynamic Time Warping (DTW) is introduced and an implementation of this technique in conjunction with diffuse tail synthesis is proposed in order to offer a complete interpolation method. Results of the application of this method to measured responses are shown, as are the results of extensive perceptual tests in which the interpolated responses are used to implement a wave field synthesis reproduction.

Chapter 6: Virtual Auditory Environments

In this chapter the concepts of the anechoic HRIR and the environment dependent RIR are merged with the introduction of the binaural room impulse response. A brief overview to Ambisonic theory is offered as a precursor to the introduction to the virtual loudspeaker approach. The feasibility of this approach in conjunction with head-tracking driven soundfield rotation is explored and a real time implementation using Fmod game development middleware is described.

Chapter 7: Application to E-learning

The advantages of the application of spatial audio in environments such as e-learning and teleconferencing is examined. The implementation described in Chapter 6 is applied to an interactive, e-learning motivated, storybook player named the ReciTell player.

Chapter 8: Conclusions

The final chapter assesses the contributions of this thesis and outlines some directions for future work.

1.3 Contributions of this Thesis

The new work described in this thesis can be summarised by the following list:

- A novel least squares based algorithm for the factorisation of large HRIR datasets.
- A further refinement of this algorithm to include two regularisation techniques, one applicable to minimum phase data, the other to full, ITD inclusive HRIRs.
- A novel technique for the spatial interpolation of the early reflection component of room impulse responses using Dynamic Time Warping.
- The development of a real time implementation of the Virtual Loudspeaker Approach with head tracking using the Fmod API and the application of this to a particular e-learning scenario.

1.4 Publications

Portions of the work described in this thesis have appeared in the following publications:

- "HRIR Factorisation: A Regularised Approach" by Claire Masterson, Gavin Kearney and Frank Boland in *Proceedings of the 18th European Signal Processing Conference (EUSIPCO), Aalborg*, August 2010.
- "Optimised Virtual Loudspeaker Reproduction" by Claire Masterson, Gavin Kearney, Marcin Gorzel, Henry Rice and Frank Boland in *Proceedings of the Irish Signals and Systems (ISSC) Conference, Cork*, June 2010.
- "A Method for Head Related Impulse Response Simplification" by Claire Masterson, Stephen Adams, Gavin Kearney and Frank Boland in *Proceedings of the 17th European Signal Processing Conference (EUSIPCO), Glasgow*, August 2009.
- "Dynamic Time Warping for Acoustic Response Interpolation: Possibilities and Limitations" by Gavin Kearney, Claire Masterson, Stephen Adams and Frank Boland in *Proceedings of the 17th European Signal Processing Conference (EUSIPCO), Glasgow*, August 2009.

- "Approximation of Binaural Room Impulse Responses" by Gavin Kearney, Claire Masterson, Stephen Adams and Frank Boland in *Proceedings of the Irish Signals and Systems (ISSC) Conference, Dublin*, June 2009. (Winner of the best student paper award)
- "Towards Efficient Binaural Room Impulse Response Synthesis" by Gavin Kearney, Claire Masterson, Stephen Adams and Frank Boland in *Proceedings of the EAA Symposium on Auralization, Helsinki, Finland*, June 2009.
- "Acoustic Impulse Response Interpolation for Multichannel Systems Using Dynamic Time Warping" by Claire Masterson, Gavin Kearney and Frank Boland, in *Proceedings of the 35th International Audio Engineering Society Conference on Audio for Games, London*, February 2009
- "Head Related Impulse Response Simplification by Deconvolution" by Claire Masterson and Frank Boland, in *Proceedings of the 8th IMA International Conference on Mathematics in Signal Processing, Cirencester*, November 2008.

2

Background

In this chapter, a brief introduction to sound perception in humans will be presented. Firstly an introduction to the anatomy of the auditory system is presented. This forms a necessary basis for the investigation of the main sound localisation cues used by humans, including both monaural and interaural cues as well as less obvious cues related to head movement and source knowledge. The neurological processing of these cues and the resolution they offer to the listener is explored. Finally the theory of localisation is extended to include the influence of the listening environment and techniques for the quantification and synthesis of this effect are examined.

2.1 Human Auditory system

The human auditory system is made up of three main components: the outer, middle and inner ear, as illustrated in Figure 2.2. The outer ear consists of the pinna (the external flaps, see Figure 2.1) as well as the ear canal. Its function is to collect and focus sound waves and transmit them to the middle ear. The pinna also imposes elevation dependent modulations on the incoming sound. The ear canal boosts frequencies in the 3 kHz region, the frequency range used for speech communication. The middle ear consists of the eardrum (or tympanic membrane) and the malleus, incus and stapes ossicles (tiny bones). The ossicles transmit the vibrations of the eardrum to the inner ear and act like a mechanical amplifier.

The inner ear contains the cochlea and the vestibular system. The cochlea is made up of three fluid chambers (scala vestibuli, scala tympani and scala media) separated by two membranes. The basilar membrane separates the scala media and scala tympani and is the location of tiny hair

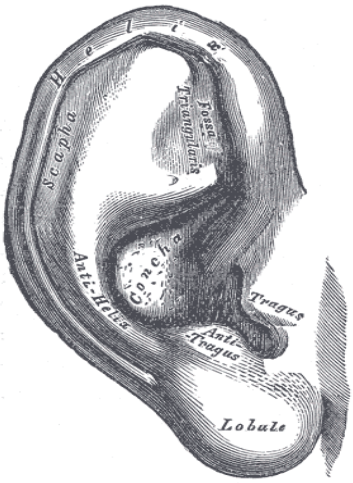


Figure 2.1: Human Pinna. (Taken from 'Anatomy of the human body' by H. Gray [66]).

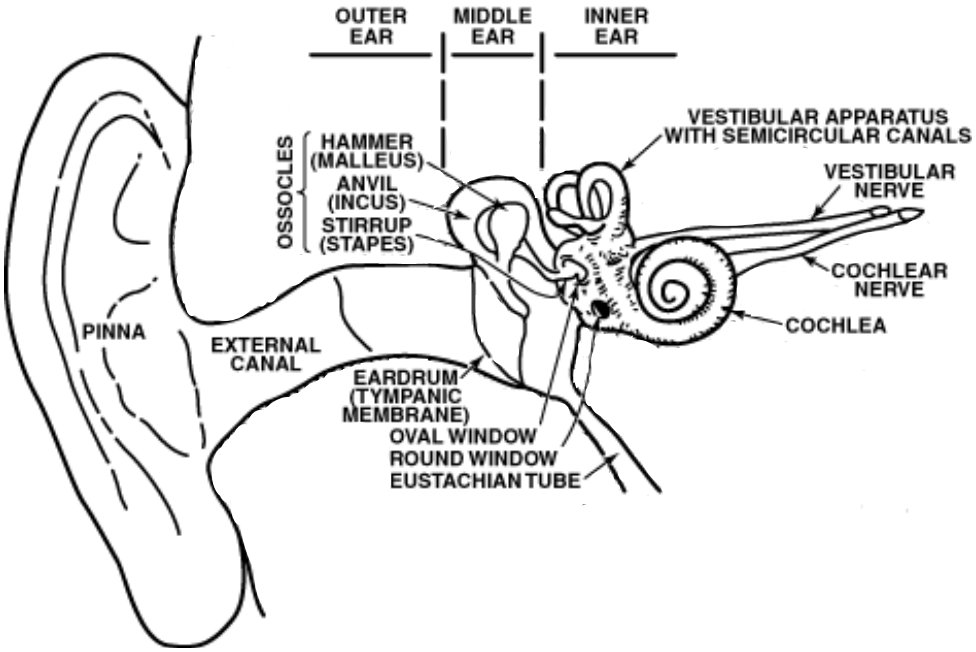


Figure 2.2: Human Auditory System. (Adapted from 'Speech Analysis and Perception' by JL Flanagan [51]).

cells. The membrane varies significantly along its length. At its base it is stiffer and narrower, while near its apex it is more flexible and wider. The mechanical vibrations of the stapes bone are transmitted to the fluid of the scala vestibuli through a small window in the cochlea. This vibration in the scala vestibuli causes vibration in the scala media and scala tympani which, in turn, results in movements of the sensory hairs of the basilar membrane. Due to the variation along its length, when presented with a pure tone, the membrane will respond maximally at a

given position depending on the frequency of the stimulus. These hair movements are converted to electrical signals which are transmitted by the auditory nerve to the brain. The membrane acts as a bank of bandpass filters with different stretches dedicated to different frequency bands. These frequency bands are referred to as critical bands. Examples of scales which aim to describe this filtering are the Bark scale [197] and the Equivalent Rectangular Bandwidth (ERB) [127] scale. Frequency, f, is related to Barks, z, as follows

$$z = \frac{26.8}{1 + \frac{1.96}{f}} - 0.53 \tag{2.1}$$

The ERBs, E are related to frequency as follows

$$E = 21.4 \log_{10}(1 + 4.37f) \tag{2.2}$$

Equations 2.1 and 2.2 are taken from [126]. The nature of critical band filtering means that if multiple tones are presented to the ear which are within the same frequency band it may not be possible to resolve them. This is called spectral masking.

The vestibular system consists primarily of three semicircular canals which sense rotational movement in three planes and two otolith organs which sense linear acceleration. A more detailed examination of the human auditory system may be found in [123].

2.2 Sound localisation

Humans use a number of cues, both monaural and binaural, to allow for the spatial localisation of sound sources. The three main cues used are interaural time difference (ITD), interaural level difference (ILD) and spectral filtering cues caused by the shape of the pinna and the body. 'Duplex theory' was developed by Lord Rayleigh in the late 19th and early 20th century [150] and this work was the first to provide a comprehensive theory to explain the mechanism of binaural sound localisation cues (i.e. ITD used at low frequencies and ILD used at high frequencies). While Rayleigh's theory is lacking in some aspects its basic tenets remain as the fundamentals of sound localisation theory.

2.2.1 Interaural Time Difference

Interaural time difference (ITD) is caused by the different transmission path lengths from the sound source to each ear and is dependent on the angle of incidence of the sound source. It takes longer for a sound wave to travel from the source to the contralateral ear than to the ipsilateral ear[1]. For pure sinusoidal tones it can be considered as an interaural phase difference (IPD). This cue is dominant in the low frequency region (<700Hz). Above this frequency the cue becomes ambiguous as the wavelength of the sound is comparable to head size. However high

[1]Ipsilateral means to be located on, or affecting, the same side of the body. Contralateral means to be located on, or affecting, the opposite side of the body.

frequency signals can also provide ITD information if the signal is modulated by a low frequency envelope. This is one aspect that Rayleigh's theory did not include. Henning [76] showed that the detection of ITD is equally good when a 3.9kHz carrier modulated with a 300Hz envelope is used as when just the envelope is used. Woodworth and Schlosberg [188] developed a formula

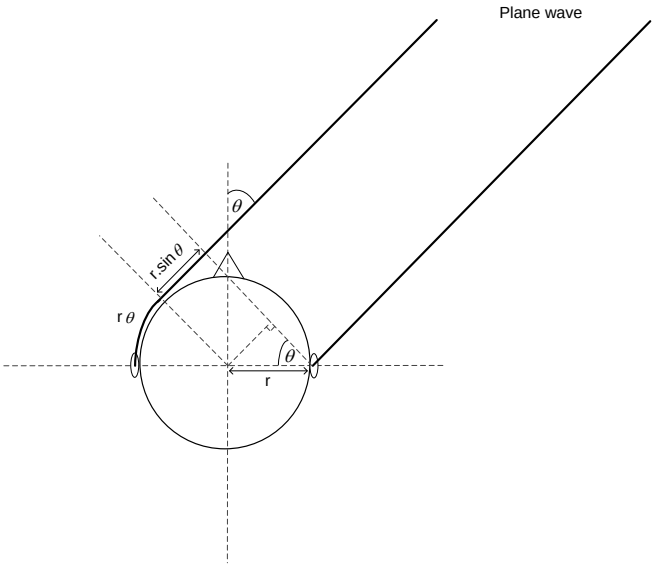


Figure 2.3: Source of Woodworth's formula.

to calculate the ITD for the simple case of a spherical head with ears at ± 90°.

$$t_{ITD} = t_{contra} - t_{ipsi} = \frac{r}{c}(\theta + \sin\theta) \tag{2.3}$$

where θ is the angle of incidence in radians, r is the radius of the sphere and c is the speed of sound in air. Figure 2.3 shows how this equation is arrived at. The equation takes into account the direct path and the low frequency diffraction of the sound wave around the sphere. Larcher et al. [105] extended this to include elevation.

$$t_{ITD} = \frac{r}{c}(\sin^{-1}(\cos\phi\sin\theta) + \cos\phi\sin\theta) \tag{2.4}$$

Busson [25] has extended the spherical model formulation to take into account the offset of ear position from the midline. This more complex arrangement requires four formulas to define the solution for all incident angles. These four cases are illustrated in Figure 2.4.

$$\text{Case A: } t_{ITD} = -\frac{r}{c}(\overrightarrow{U}_{inc}.\overrightarrow{U}_l + \overrightarrow{U}_{inc}.\overrightarrow{U}_r) \tag{2.5}$$

$$\text{Case B: } t_{ITD} = -\frac{r}{c}(\frac{\pi}{2} - arccos(\overrightarrow{U}_{inc}.\overrightarrow{U}_l) - \overrightarrow{U}_{inc}.\overrightarrow{U}_r) \tag{2.6}$$

$$\text{Case C: } t_{ITD} = -\frac{r}{c}(\frac{\pi}{2} - \overrightarrow{U}_{inc}.\overrightarrow{U}_l - arccos(\overrightarrow{U}_{inc}.\overrightarrow{U}_r)) \tag{2.7}$$

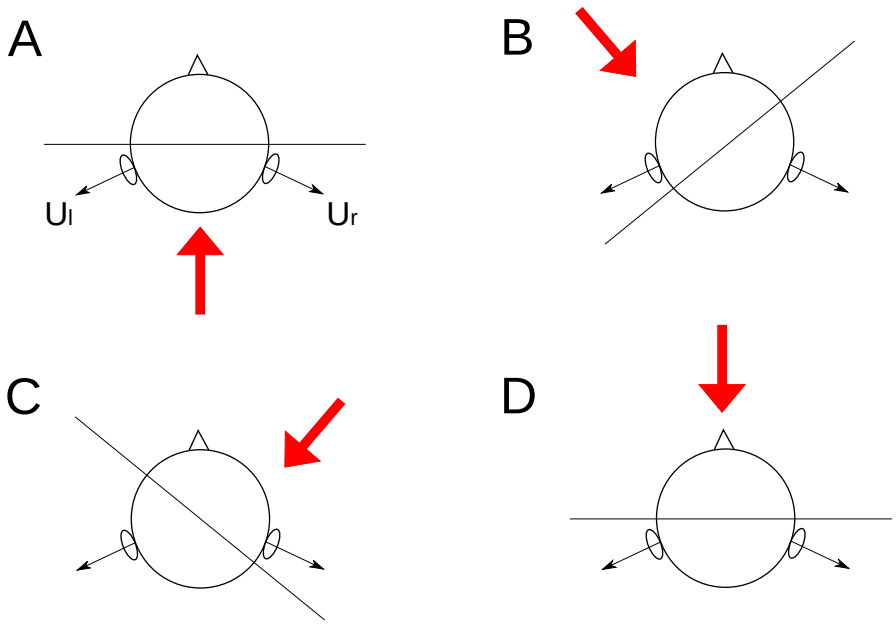


Figure 2.4: 4 cases of Busson's ITD formulae.

$$\text{Case D: } t_{ITD} = \frac{r}{c}(arccos(\overrightarrow{U}_{inc}.\overrightarrow{U}_l) - arccos(\overrightarrow{U}_{inc}.\overrightarrow{U}_r)) \tag{2.8}$$

$\overrightarrow{U}_l$ and $\overrightarrow{U}_r$ are the unit left ear and right ear vectors and both are a function of the azimuthal and elevation position and $\overrightarrow{U}_{inc}$ is the unit incidence vector. As the vectors are of unit length, the dot product expressions in each equation simplify to the cosine of the angle between the two vectors. Figure 2.5 demonstates Busson's formulae and shows the effect that moving the ear has on the ITD. The ears are offset 5°,10° and 15° from the midline towards the back of the head and the ITD is shown for the horizontal plane.

Onset detection and interaural cross correlation on left and right ear HRIRs as well as the calculation of the interaural group delay difference at 0Hz are examples of other common techniques used to calculate ITD. An explanation and comparative study of many of the techniques can be found in [122]. The authors indicate that most methods produce incorrect values in the 90° to 110° region in the azimuth. The interaural cross correlation technique is used in this thesis and as such a brief description will be given here. The Interaural Cross Correlation Function (IACF) is a measure of the correlation between the received left and right ear signals within the integration limits t_1 to t_2 as a function of the time delay τ.

$$IACF\tau = \frac{\int_{t_1}^{t_2} x_1(t)x_2(t+\tau)dt}{\sqrt{\int_{t_1}^{t_2} x_1^2(t)dt \int_{t_1}^{t_2} x_2^2(t)dt}} \tag{2.9}$$

The point at which the function yields its maximum is known as the Interaural Cross Correlation

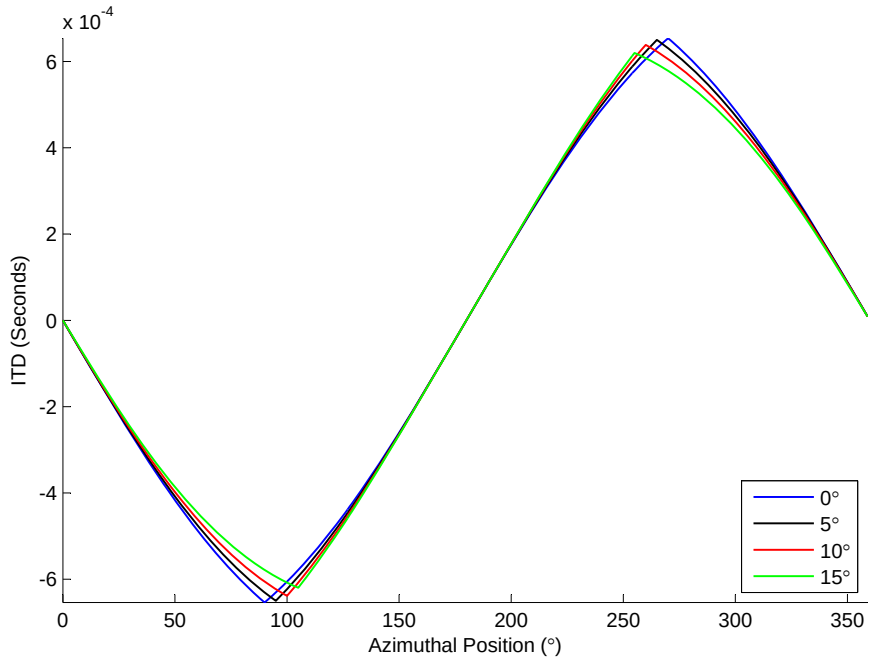


Figure 2.5: Effect of ear displacement on ITD according to Busson's formulae.

Coefficient (IACC) and the value of τ at this point gives the ITD.

$$ITD = arg(max_{\tau}|IACF\tau|) \tag{2.10}$$

ITD varies in a similar way amongst subjects for difference source positions and is relatively constant over frequency.

2.2.2 Interaural Level Difference

Interaural Level Difference (ILD), also known as Interaural Intersity Difference (IID), is produced by the shadowing of the head which reduces the sound intensity at the contralateral ear for high frequency sounds (>1.5kHz), see Figure 2.6. 700kHz is the level where the wavelength of the sound becomes comparable to head size. For lower frequencies (or longer wavelengths) the sound begins to defract around the head and ILD is minimised. Between 700Hz and 1.5kHz there is a transition range where both ITD and ILD are used as well as spectral cues which will be discussed later. Unlike ITD, ILD varies in a much more subject dependent way for different source positions and is a complex function of frequency.

2.2.3 Spectral cues

It becomes apparent that simple ITD and ILD cues alone result in ambiguity in source position. ILD and ITD cues result in front back confusions and provide no method for distinguishing

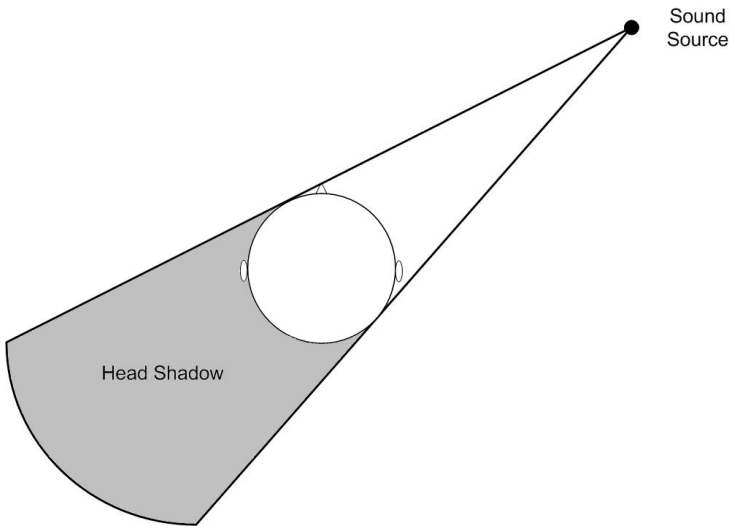


Figure 2.6: Interaural level difference.

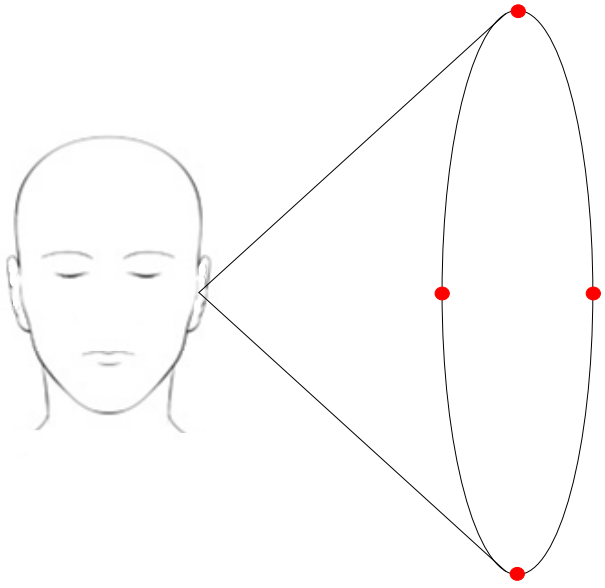
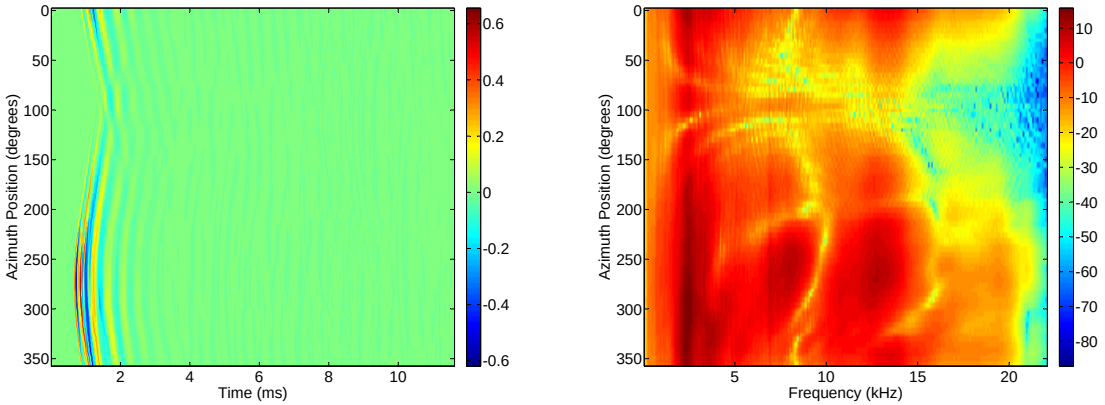


Figure 2.7: Cone of confusion.

elevation positions. This is known as the cone of confusion. A sound source placed at any point on the surface of the cone shown in Figure 2.7 would result in the same ITD and ILD.

The shape of the pinna, head and torso impose direction dependent spectral colourations onto incoming sound signals. Instead of travelling directly into the ear canal, some sound is reflected in various ways by the pinna before either entering the ear canal or being reflected away. This aids in reducing the cone of confusion ambiguity. This spectral information, as



(a) HRIR (b) HRTF

Figure 2.8: Left ear HRIRs and HRTFs for 0° elevation from Gardner and Martin's KEMAR dataset [58].

well as the ITD and ILD are all encompassed by the Head Related Impulse Response function (HRIR). The frequency domain equivalent of the HRIR is referred to as the Head Related Transfer Function (HRTF). Roffler and Butler [153] carried out an experiment where the pinnae of test subjects were covered by plexiglass plates attached to a head band. They found that the vertical localisation of sound sources became nearly impossible in this instance. This important paper demonstrates the key role spectral cues play in the localisation of sounds in the vertical plane. An older experiment by Musicant and Butler [132] also noted the importance of pinna induced spectral colouring (as well as high frequency content in the source spectrum) to vertical localisation as well as highlighting their importance to resolving front back confusions. HRIRs will be introduced in more detail in Chapter 3.

2.2.4 Dynamic Cues

In order to achieve convincing virtual auditory environments with externalisation[2] of source position and minimal front back confusions it is necessary that the audio processing system respond to and update with head movements of the listener and changes in source position.

Wallach [177] hypothesised that small head movements are used by a listener to get rid of front back confusions. Young [191] performed an interesting experiment where the pinnae of the test subjects where bypassed by funnels. When the head movements were prevented the participants were unable to localise vertically or resolve front back confusions. However when head movement was permitted localisation accuracy returned to relative normality which

[2]Externalisation refers to the 'out of head' localisation of a virtual sound source. It is the opposite of 'in head' localisation, which is also referred to as lateralisation. This occurs when the sound source is perceived to be positioned on the axis between the two ears and is a result of missing, ambiguous or conflicting localisation cues [111].

suggests that head movements can offset the unavailability of individual pinnae. The key role of head movements in the resolution of front back confusions was demonstrated by Wightman et al. [183]. In their experiments when listeners were encouraged to move their heads or to move the source position front back confusions were almost eliminated. Further confirmation can be found in [12]. Rosiles [154] demonstrates the link between dynamic spatial localisation cues and spatial orientation information received from the vestibular system.

Wightman and Kistler [182] hypothesise that monaural sound localisation is heavily influenced by small head movements. While studying the effects of occluding one ear in sound localisation the authors noted a significant difference in localisation performance between free field listening and a virtual source scenario (using HRTFs). Using a head tracker the authors found that listeners who had been asked to keep their heads stationary during the experiments still moved them slightly (by up to 2°) and this aided them in the free field case.

Movement of the sound source also provides dynamic localisation cues. The Doppler shift results in a change in the pitch of a sound for the listener due to the motion of the source (see Collins' 2010 book titled 'Introduction to Computer Music' [36] for more detail on this).

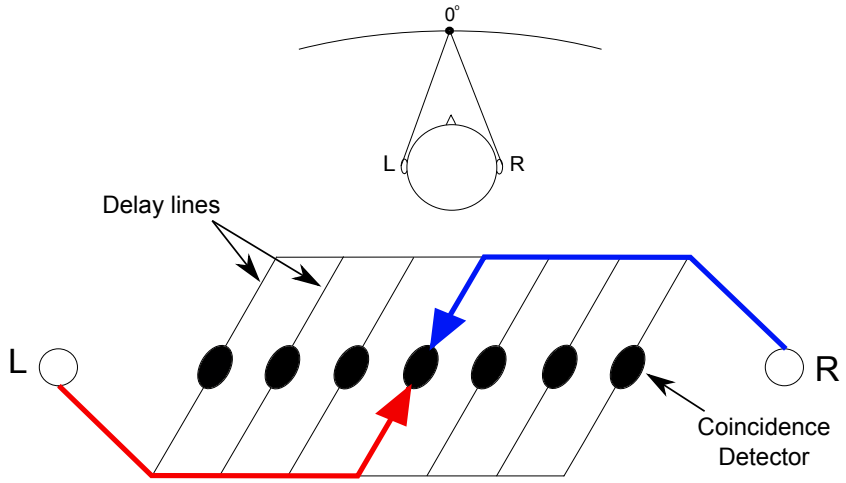
2.2.5 Other cues

Knowledge of the sound source, from either past experience or from optical information is an important cue for localisation. One can appreciate the importance of knowledge of the spectral properties when one considers the monaural localisation of the sound. In order for the spectral filtering of the pinna to be extracted prior knowledge of the spectrum of the sound source is necessary.

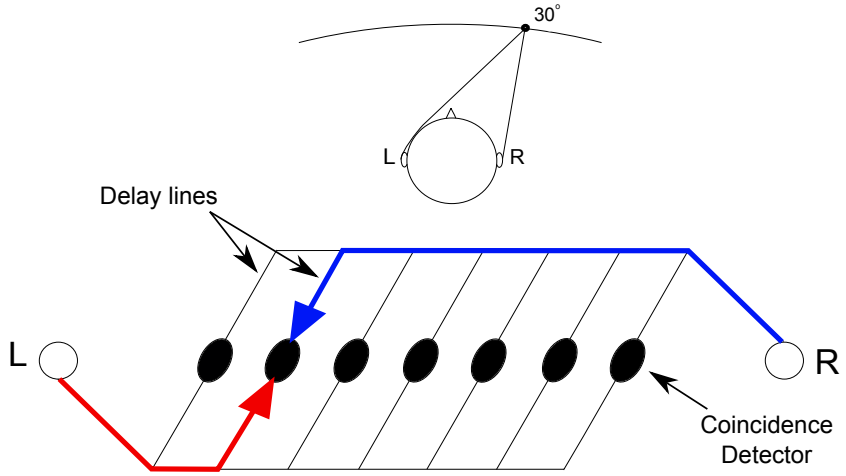
Wightman and Kistler [182] provide a comprehensive study of monaural localisation and the effect of source knowledge on this. They examined the effect of scrambling the sound source spectrum by randomising the magnitude of six critical bands and comparing localisation to that of a flat spectrum source. The authors show that scrambling the spectrum significantly degrades both frontback and updown perception and conclude that this is most likely as a result of the effect such scrambling has on monaural cues. Without knowledge of the source spectrum the listener is unable to successfully deconvolve the useful location based spectral filtering. They also note however that flat spectrum source sounds are an ideal case and that real world listening conditions would undoubtedly have a less regular source spectrum.

2.3 Neurological Processing of Localisation Cues

Jeffress [86] introduced his theory on the neural mechanism for ITD detection over 60 years ago and it still remains at the centre of literature on the topic. His model proposes that a series of axonal delay lines and coincidence detectors compute an ITD between the left and right auditory channels as shown in Figure 2.9. As such this system can be viewed as a bank of cross



(a) Source straight ahead



(b) Source offset by 30° from straight ahead

Figure 2.9: Demonstration of Jeffress model.

correlators, each receiving a slightly delayed version of the signal. The ITD is determined by finding the maximally responsive cross correlator.

This model remained theoretical for a considerable period of time but has been supported in recent decades by studies carried out on the barn owl [28], [29]. However more recent work suggests that the Jeffress model may not be as applicable to mammals as it is to birds [68], [142]. Hancock et al. [72] demonstrates that ITD accuracy gets worse as ITD increases in small test mammals. This is at odds with the Jeffress model which would require ITD accuracy to be a constant function of ITD. The authors suggest that a neural pooling model is more appropriate

to ITD discrimination than one where each neuron is tuned to a different ITD (lower envelope hypothesis). It has been shown that glycinergic inhibition is key to mammalian ITD coding [24], [143]. Glycine is an inhibitory neurotransmitter, meaning that it blocks the tendency of a neuron to fire. Dramatic shifts in ITD tuning curves shown in these papers following the suppression of inhibition demonstrate this. One key result of the study of Brand et al. [24] was that the blocking of inhibitory inputs caused all the neurons tested to respond maximally to ITDs close to 0μs. This indicates that the systems of varying length delay lines suggested by Jeffress do not exist and instead that inhibition and its timing with respect to the excitation of neurons is responsible for the extraction of ITD information. The medial superior olive (MSO) is the part of the brain believed to be responsible for ITD capture.

Excellent recent overviews of the state of the art in this area can be found in [148], [26], [115] and [161] if further reading is required. The inherent elegance of the Jeffress model dictates that it will continue to dominate literature on the neurological auditory processing, however, it is becoming increasingly clear in light of recent research that it is at best incomplete when applied to mammals.

The lateral superior olive (LSO) is the part of the brain thought to process ILD cues in mammals. ILD-sensitive neurons are excited by input from the ipsilateral ear and inhibited by input from the contralateral ear [69].

2.4 Spatial Resolution of Localisation

The resolution of our localising ability is dictated by our adeptness in detecting small changes in interaural cues. The just noticeable difference (JND) has been found to be as low as 10μs for ITD and 0.5dB for ILD under optimal conditions [121]. The Minimum Audible Angle (MAA) encompassed both these differences and is described as the smallest detectable difference between the azimuths off two identical sound sources. Mills [121] shows this value to be approximately 1° for sources directly in front of the listener. However this value is shown to dramatically increase as sources divert from straight ahead of the listener. Mills finds that there is a gradual increase in MAA from 1° to 2° as the source moves from directly in front of the listener to approximately 60° to the right. After this the MAA exponentially increases to above 40° for an angular position of 90° (i.e. directly right of the listener). MAA resolution is measured for different frequency tones and when the MAAs for each frequency where compared a significant increase is found between 1kHz and 3KHz, peaking between 1.5 and 2kHz. Mills hypothesises this is due to a crossover in dominancy between ITD and ILD.

The Minimum Audible Movement Angle (MAMA) is defined as the minimum angle of travel required for detection of the direction of sound movement. Perrott and Musicant [144] found it to increase as source velocity increased. They use a 500Hz sine tone as the stimulus. At a velocity of 90°/s the MAMA was found to be 8.3°, increasing to 21.2° for 360°/s. Perrott and Tucker [146] expand on these results by examining the effect of different stimulus frequencies

(ranging from 500Hz to 3.7kHz) on the findings. They found the best resolution performance for frequencies in the range 700-800Hz and the worst for the 1.3-2kHz range. Saberi and Perrott [158] report more comprehensive results for MAMA for movement velocities ranging from 1.8°/s to 320°/s and find the MAMA to vary from approximately 1.7° to 10° in this range. Chandler and Grantham [33] confirm the inverse relationship between velocity and dynamic spatial acuity and conclude that this is an indication that there is a minimum integration time needed by the auditory system to resolve position. They also show both MAA and MAMA exhibit a similar dependence on signal frequency and bandwidth.

This discussion so far has considered spatial resolution in the horizontal plane. Our ability to resolve source location vertically is poorer than horizontally. Perrott and Saberi [145] find the mean MAA threshold in the vertical plane to be 4 times that of the horizontal plane. However at large azimuths ($\sim 90°$) resolution in the vertical plane is comparable to that of the horizontal plane [157]. Grantham et al. [64] show that performance is as good in the diagonal plane as it is in the horizontal. This is because localisation in this plane is based contributions from both interaural differences and spectral cues.

Depth (or distance) resolution is not as well documented. Gardner [54] completed a study where speech stimulus was used in an anechoic environment. He found for a source a 0° the ability of the listener to judge the distance of the source was very poor. He finds the level of the sound source to be an important factor. Shouting generally lead to an overestimation of distance while whispering lead to underestimation. Deviation from the 0° orientation improves perception. Of course the anechoic case is not one that is frequently encountered in real world environments. More recently Zahorik [192] discusses the cues used in naturally occurring situations and their relative importance. The four main cues are sound intensity, direct to reverberant energy ratio, spectrum and binaural differences. The direct to reverberant energy ratio is the ratio of sound energy reaching the listener directly to that reaching the listener as a result of reflections. The sound absorption of air alters the spectrum of a signal for large distance (>15m) by attenuating high frequencies. The spectrum of sound may also be altered by the acoustic properties of any surface it reflects off. Binaural cues are only relevant for near field sources. Intensity and direct to reverberant energy ratio were found to be the most important cues with intensity being weighted highest for speech signals and direct to reverberant energy ratio being more reliable for unfamiliar signals.

2.5 Environmental Cues

The size and shape of the room an auditory event occurs in, the materials its surfaces are comprised of and the position of the sound source and the listener in the room are factors which effect sound transmission from source to receiver. These effects can be approximated as linear and can be comprised in the room impulse response (RIR). This is effectively an acoustical footprint of the room. The RIR can be split into three component parts: the direct sound,

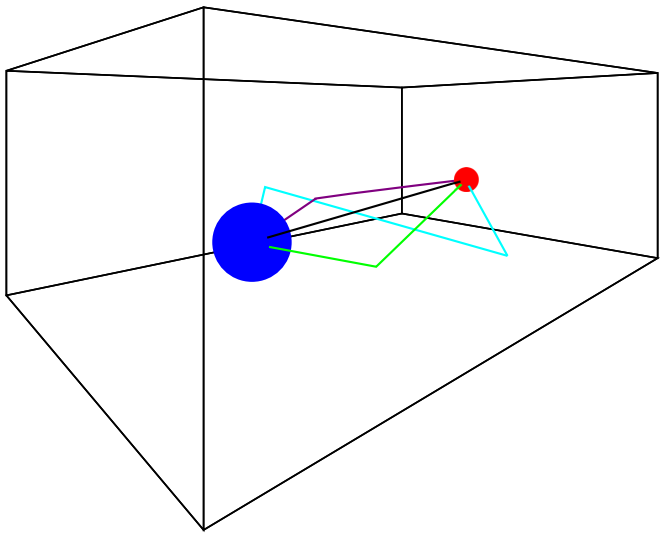


Figure 2.10: Demonstration of various reflection paths between source and receiver.

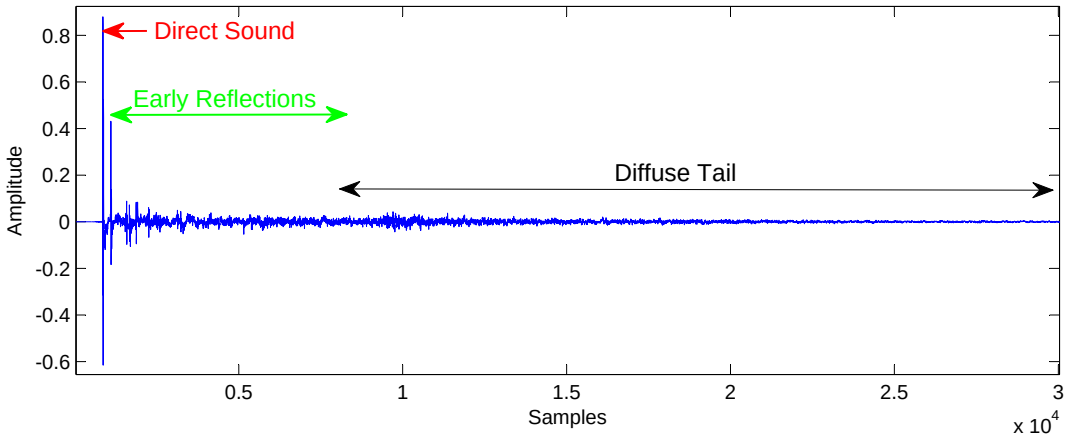


Figure 2.11: A sample RIR.

early reflections and the diffuse tail (see Figure 2.11). If there is a line of sight between the source and receiver the direct sound will manifest as the first peak in the response. If there is not a direct path due to the presence of some obstacle it is still likely that due to the diffraction of low frequency sounds around the obstacle there will be some activity in the RIR before the reflections. The early reflections are quite sparse in nature. These play an important role in the perception of distance. For sources that are near to the listener the ratio of the direct sound magnitude to that of the early reflections is a key distance cue. The time difference between the arrival of the direct sound and the early reflections is also of great importance. The closer the

sound source to the listener the larger the time delay between the direct sound and the early reflections. The late reflections or the diffuse tail component can generally be considered to be stochastic in nature. The transition between the early and late reflection sections of the RIR is more of a gradual one than a specific single point. There are numerous methods of determining where this transition occurs. This will be discussed further in Chapter 5.

The RIR conveys considerable information to the listener regarding its capture environment aside from localisation cues. However it is not always clear how objective properties of the RIR can be linked to the subjective quality rating of acoustic environments. Authors such as Okano et al. [140] and Cerda et al. [32] have contributed to this area by examining RIR measurements along with subjective ratings of concert hall type venues. They find parameters such as interaural cross correlation index, reverberation time, lateral force, early decay time and the time between the first direct sound and the first reflection useful in characterising the quality of an acoustic space. Okano et al. [140] term this subjective assessment as 'spatial impression' and relate it to two factors: Apparent Source Width (ASW) and Listener Envelopment (LEV). ASW is associated with the early relation component of the room response and describes, as its name suggests, the apparent auditory width of the sound field created by a sound source in a room at a given listener position. LEV refers to the listener's sense of being surrounded by sound and is largely related to the late reflections or diffuse component of the impulse response. The diffuse reverberation is key to enhancing the timbre and realism of an auralised reproduction. Barron and Marshall [10] have shown that listeners have a preference for lateral reverberation energy which produces uncorrelated signals at the two ears.

2.5.1 Measurement and Synthesis of RIRs

The most commonly used techniques for the measurement of RIRs are maximum length sequences (MLS) [162] and logarithmic sine sweeps [50]. The MLS technique involves exciting the room in question with a pseudo random binary signal. The impulse response is obtained by cross correlating the excitation signal and the recorded signal. This can be done efficiently using the fast Hadamard transform. However MLS based measurement is based on the assumption of a linear time invariant system and as such does not quite live up to its theoretical promise in practical applications. Factors such as local variations in air temperature and small vibrations of the surfaces of the environment may cause such time variance and result in noise spikes. It is for this reason that Farina's sine sweep method is preferred. This technique involves using a non-periodic logarithmic sweep stimulus to excite the room and convolving the resulting recorded signal with an inverse of the original stimulus. This produces a time delayed impulse response with the responses of the harmonic distortion orders separated out at the start of the signal. Since Farina's seminal paper in 2000 [50] this technique has become the standard in RIR measurement due to its robustness to time variance and non linearity.

If actual physical RIR measurement is not feasible (perhaps due to the building not yet

being in existence!) then synthetic RIRs can be acquired based on a geometrical model of the environment as well as knowledge of the materials the environment is made up of. These modelling techniques fall into two main categories: Wave equation based techniques and ray tracing based techniques. Wave equation based techniques include Boundary Element Method (BEM), Finite Element Method (FEM) [101] and Finite Difference Time Domain (FDTD) methods such as Digital Waveguide Mesh (DWM) [131]. FEM involves obtaining a numerical solution to the wave equation by dividing the space into small volumes, while BEM divides the boundaries of the space into small areas. FDTD involves using finite differences in place of the derivatives in the wave equation. While these techniques are based on rigorous wave acoustics and produce accurate results they are unfortunately computationally cumbersome and as such are only applied for low frequencies. Ray tracing techniques include the image source method [7], [21] and pure ray tracing as well as hybrid techniques. These techniques neglect the wave nature of sound and consider it to propagate only as rays. The image source method (see Figure 2.12) is most suited to small polyhedral shaped rooms but has been extended to incorporate diffraction [169] and as such can be used in more irregular spaces.

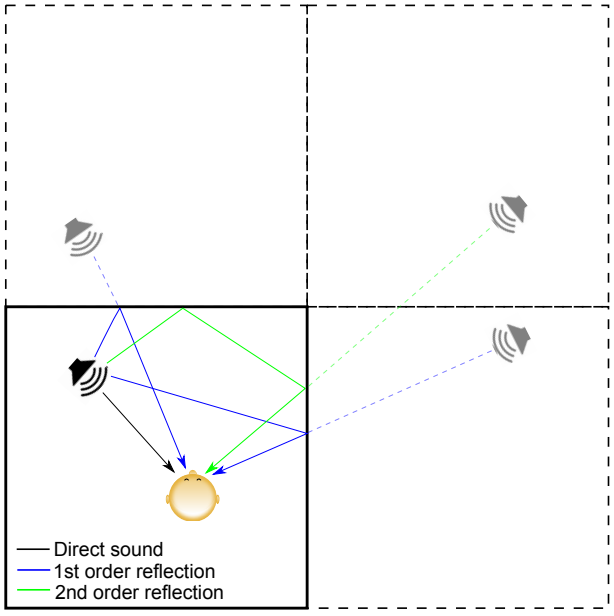


Figure 2.12: Simple demonstration of image source technique.

Alternatively scale models of the space under study can be constructed. This kind of measurement has its own restrictions however. The frequencies used in the measurement stimulus must be scaled upwards by the scaling factor used to scale down the room. For true reproduction the absorption coefficients of all materials used in the model should be scaled upwards from their real world equivalents. Air absorption is also a factor which must either be compensated for or avoided by replacing the air by an inert gas such as nitrogen. For further reading on this topic see [31]. This technique is, however, slowly falling out of favour as published work in the area

focuses more on the increasingly powerful computational modelling techniques discussed in the previous paragraph.

2.6 Conclusion

This chapter has provided an overview of the main aspects of human sound localisation. It began with an examination of the anatomy of the auditory system and proceeded to explore the principal sound localisation cues, the spatial resolution they allow for and the neurological processing used in their comprehension. The importance of the influence of the room or environment in which the sound source and listener are located has been discussed and the RIR has been introduced to quantify this influence for a particular source receiver position. Techniques for measurement, or alternatively synthesis, of the RIR have also been detailed.

3

Head Related Impulse Response

As was briefly introduced in Section 2.2.3 the spectral cues introduced to a signal due to the filtering effects of the pinna, head and torso, as well as the ITD and ILD cues introduced by the displacement of the ears, are contained in the Head Related Impulse Response (HRIR). The Head Related Transfer Function (HRTF) is simply the Fourier transform of the HRIR. By convolving a mono sound source with the left and right ear HRIRs for a given spatial position and playing the resulting two channels binaurally, a virtual sound source at that spatial position can be created. To allow for convincing, artifact free reproduction in real time dynamic virtual audio environments it is necessary that filters be as short as possible while still retaining their perceptual integrity. Interpolation is also necessary to allow for smooth transition between filters as the source or listener position changes. As well as discussing these factors, the physiological basis of the HRIR will be explored in this chapter and measurement and modelling techniques will also be examined.

3.1 Physiological Basis

The head, torso and pinna shape the spectrum of the HRTF for a given source position. As pinna configuration differs from person to person, this spectral shaping is highly individual dependent.

The torso imposes some relatively low frequency perturbations to the HRTF spectrum. The wavelength of sound does not become comparable with pinna size until frequencies are greater than 3.5kHz but spectral shaping is present below this level. Gardner [55] concludes that the torso imposes cues in the 0.7 - 3.5 kHz range. Gardner examined the difference response of

HRTF data measured at ±18° for a dummy head with and without occlusion of the pinna and with and without the torso piece. He finds that below approximately 3.5kHz occlusion of the pinna has minimal effect. However when the torso piece is removed from under the mannequin the broad dip between 0.7 and 2 kHz is smoothed. Kuhn [102] also demonstrated that the torso had a spectral effect at low frequencies. Algazi et al. [5] establishes a similar result. The authors conclude that removal of the torso from the measurement apparatus results in the loss of large arch shaped notches below 3kHz.

The head is also an important contributory localisation factor especially for the contralateral ear. Algazi et al. [5] show that when the impulse responses for the contralateral ear are examined as a function of elevation, the ITD is not constant as a function of elevation as would be expected within to the cone of confusion. This is caused by the sound diffracting around the head from both sides, the non central positioning of the ear and asymmetry of the head. The interaction between the direct and diffracted waves leads to comb filtering in the spectrum.

As mentioned above the pinna imposes spectral information on the HRTF for frequencies greater than 3.5kHz. Bloom [18] reports that applying a notch to the frequency spectrum of sound can apply the illusion of elevation to the sound source. Shaw et al. [164] found a spectral notch whose position increased in frequency from 6 to 10kHz as elevation was increased from -45° to 45°. Hebrank and Wright [75] reported a number of spectral features in the HRTF and suggested what their function in localisation might be. Frontal cues are said to be as a result of a notch between 4 and 8 kHz. An increase in frontal elevation is mirrored by an increase in the frequency of the notch and the authors suggest that this is caused by the reflection of sound of the posterior concha. A peak between 7 and 9kHz is hypothesised to be the result of a source above the head and a peak between 10 and 12kHz to be the result of a source behind the listener. Moore et al. [128] confirm that the cue described in [75] are detectable by listeners. They also state that spectral peaks should be easier to detect than notches. Raykar et al. [149] provide a good overview of the composition of the HRIR as well as proposing a signal processing method to determine the frequencies spectral notches occur at and relating these positions to pinna shape.

These cues are undoubtedly important. Gardner and Gardner [56] showed that progressive occlusion of the pinna leads to increasing confusion when localising vertically. It has also been shown [173] that covering subjects ears with other pinnae decreases localisation performance but that the subject gradually adjusted to the new pinnae and localisation ability returned to original levels. Interestingly when the fake pinnae where removed the subjects retained high localisation accuracy. In Section 2.2.4 the use of dynamic cues for sound localisation was discussed briefly. Wightman and Kistler [182] suggest that head movements are an important factor in monaural localisation.

3.2 Measurement and Modelling

HRIRs are generally measured with the measurement microphone located at the entrance to, or slightly inside, a blocked ear canal (see Figure 3.1). It is argued that as the response of the ear canal is largely direction independent it can be excluded from the measurement process [163]. Probe microphones can be used to take open ear canal measurements alternatively but the resonance of the ear canal decreases the signal to noise ratio of the measurement as well as reducing the repeatability of the measurement. The measurement process should ideally be completed in an anechoic environment but pseudo-anechoic HRIRs can be acquired by measuring in a non anechoic environment and choosing a cut of point that removes the environmental reflections. An interesting alternative to this is a method proposed by Duraiswami et al. [46] based on the principle of reciprocity. The authors reverse the positioning of the loudspeaker and microphone.

Figure 3.1: HRIR measurement apparatus.

Currently the most widely used measurement stimulus is the logarithmic sine sweep as proposed by Farina [50] (refer to Section 2.5.1 where this is discussed in relation to RIR measurements). Following the measurements the impulse responses are obtained by a deconvolution operation and any harmonic distortion artifacts, which the deconvolution operation separates out, are removed. Maximum length sequences (MLS) and other stimuli can also be used.

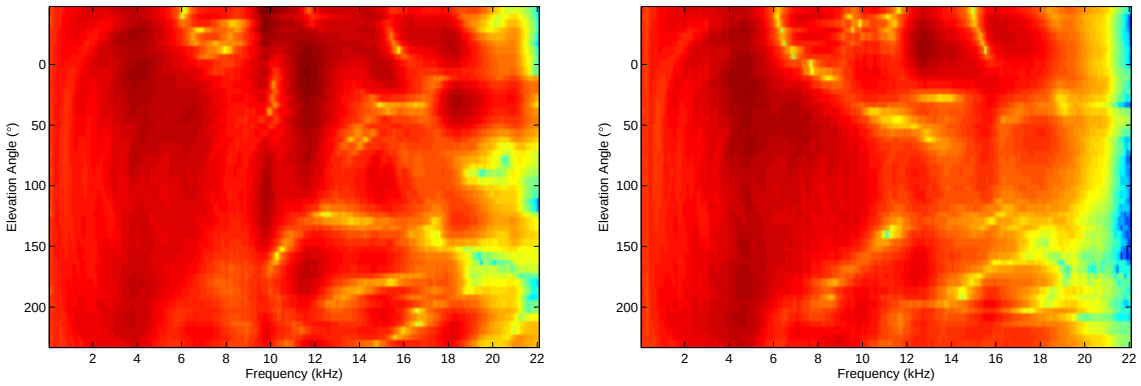
A loudspeaker is moved through a grid of measurement positions which are located on a sphere of fixed radius centred on the subject head. High spatial resolution is required for convincing reproduction and as such the measurement process can be a lengthy one. If one considers 5° angular increments around the measurement sphere, the number of measurements can soon reach into the thousands. Considering that the subject must remain as still as possible

for the measurement procedure it becomes clear that this measurement process can become tedious and very dense grids of measurements may be unrealisable for live subjects.

The time consuming and invasive nature of measurement motivates a need to simulate HRTFs with a high degree of accuracy. Kahana [90,91] successfully models HRIRS using BEM solution of the wave equation based on high resolution 3D laser scans of the ear. This work builds on the work of Katz [93] who, due to computation restrictions at the time, could only work on frequencies up to 5kHz. In [91] (published in 2007) Kahana successfully investigates heads with pinnae attached up to 10kHz and baffled pinnae up to 20kHz.

3.2.1 Individual Nature of the HRTF

One of the main issues holding back binaural reproduction of spatial audio is the individual nature of sound localisation. Differences in head shape and diameter, distance from torso to ear, ear position on the head and of course the individual nature of the pinna all effect the nature of the HRTF. Recently Treeby et al. [172] found that even hair has an effect by producing asymmetric perturbations in the HRTF and interaural cues. This individual nature is demonstrated in Figure 3.2 where the HRTFs in the sagittal plane are shown for two subjects (21 and 165) from the CIPIC database. These subjects are the KEMAR mannequin head with large and small pinnae respectively. There is a clear difference in the spectra evident when the the subjects are compared. The fact that the responses are from a mannequin head eliminates head movements as the cause of this differential. Using non individual HRTF data has been shown to reduce externalisation [96]. Wenzel et al. [179] show an increase in front back confusions and errors in vertical localisation as a result of using non individual HRTFs. Moller et al. [125] report similar results.



(a) Subject 21 (b) Subject 165

Figure 3.2: HRTFs at different elevation positions for two subjects from the CIPIC database.

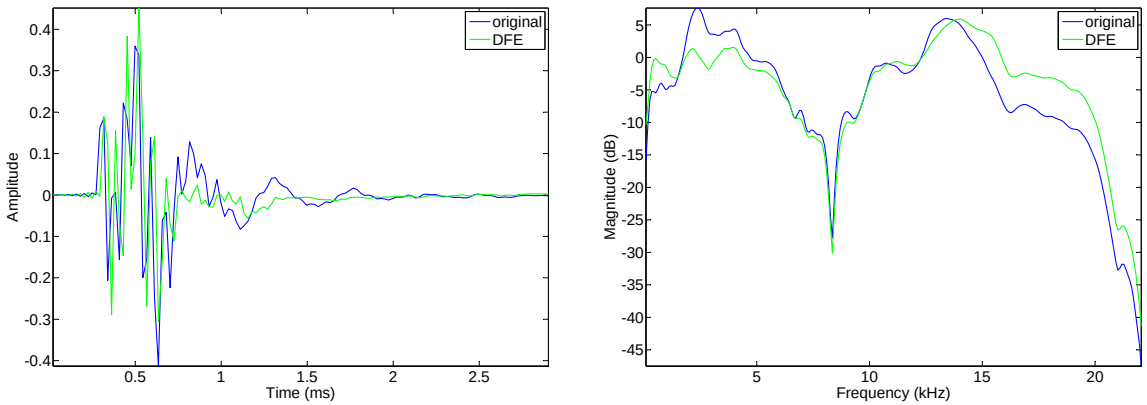
Due to the unfeasibility of HRTF measurement of the mass population, it is necessary to attempt to model this individuality. In Section 2.2.1 Woodworth's formula for ITD calculation based on a spherical model for the head was introduced. This allows for the head radius to be

used as a parameter for ITD calculation. More recently Busson has adapted this formula to take into account the fact that the ears are generally offset from the midline. This allows for two parameters to be customised based on listener specific information, both the head radius and ear position.

While adapting the ITD to individual characteristics is a good start, modelling the spectral detail based on listener morphology is also necessary and is considerably more difficult. The brute force approach is to model a correct HRTF based on high resolution optical scans of the pinna, head and torso using the BEM or a similar technique as described in the previous section. As has been discussed this is extremely computationally demanding and requires time consuming and expensive scanning techniques. Alternatively it may be possible to choose the closest HRTFs for a listener from a large dataset representative of a wide range of the population based on certain quick morphological measurements [196] or by listening tests [85]. It may also be possible to adapt more generic HRTFs to an individual. Middlebrooks et al. [120] describe a method of scaling HRTFs in frequency. The scaling factor is decided by a fast psychophysical procedure which assessed listeners preferences for different scale factors. Guillon et al. [70] extend this further by introducing rotation of the coordinate system as another degree of freedom. This rotation technique was first introduced by Maki and Furukawa [113] for HRTF customisation of gerbils. Tan and Gan [170] employ a technique where the user chooses the most suitable HRTF set from a dataset and then fine tunes it with spectral manipulation and filtering.

3.2.2 Equalisation

When a HRTF is captured the transfer functions of the loudspeaker and microphone used in the measurement process are included in the HRTF. It is possible to capture the transfer function of the measurement apparatus and apply its inverse to each HRTF. This response could be measured with a microphone positioned at the centre of the head with no head present. Alternatively it is possible to equalise the measured HRTF set. The two main techniques are free field equalisation (FFE) and diffuse field equalisation (DFE). [57, 89, 106, 124] provide detailed descriptions of these techniques. Free field equalisation involves equalising the entire dataset of HRIRs with respect to one HRIR measurement, usually a frontal one. An inverse filter is created based on this frontal reference HRTF and is applied to the dataset. This technique works well if the sound source is in front or in the diffuse field but, if this is not the case, it may alter the sound spectrum in an asymmetric manner. Diffuse field equalisation negates the need to privilege a particular direction, instead equalising with respect to the average power spectrum of the whole uniformly distributed dataset. The filter is assumed to be minimum phase. Larcher et al. [106] advise the use of DFE and describes it as a robust technique for removing non-directional idiosyncrasies from HRTF measurements.



(a) HRIR (b) HRTF

Figure 3.3: Demonstration of effect of DFE.

3.3 The Minimum Phase Assumption

Any rational system can be decomposed as follows [141]

$$H(z) = H_{min}(j\omega).H_{ap}(j\omega) \tag{3.1}$$

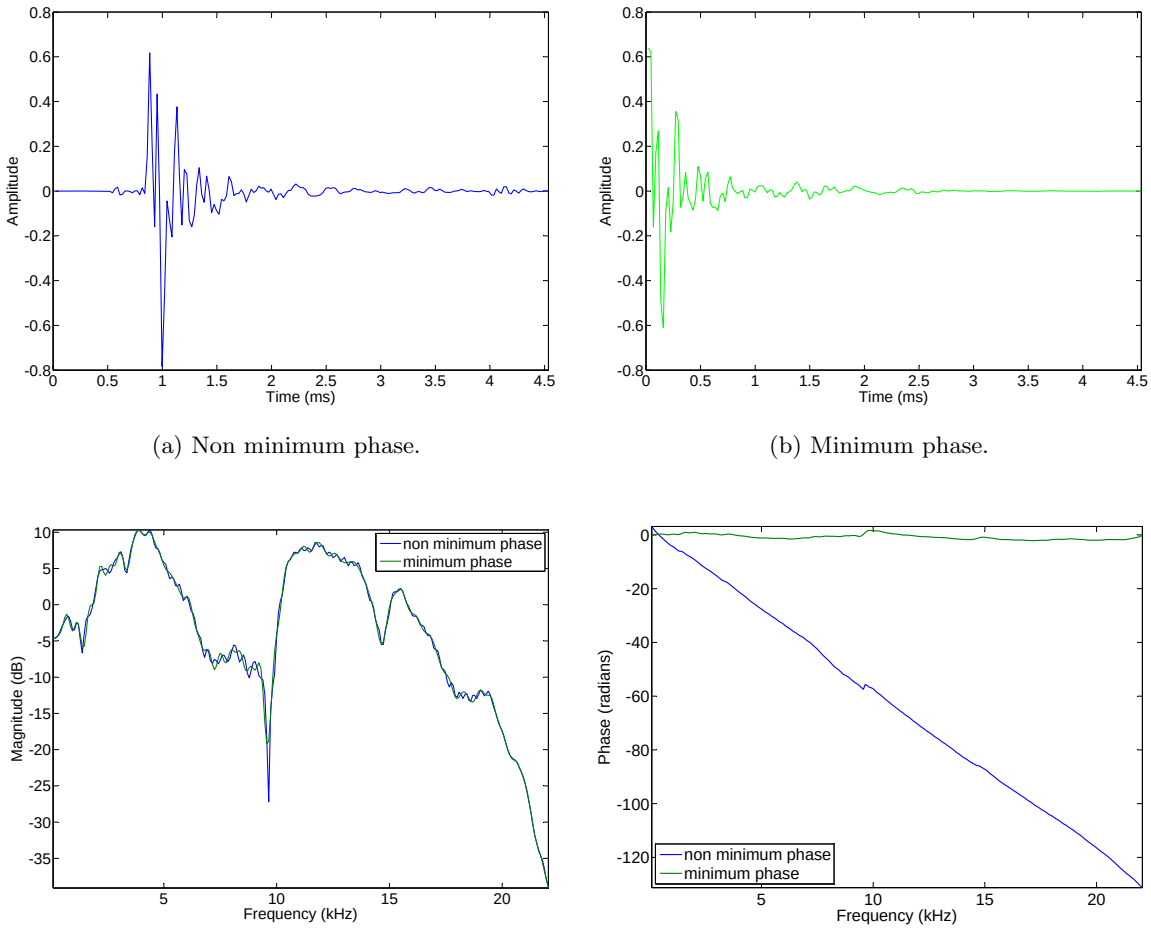
where $H_{min}(z)$ is a minimum phase system and $H_{ap}(z)$ is an all-pass system. The minimum phase system has the property that all its poles and zeros are inside the unit circle, i.e. both the system $H_{min}(z)$ and its inverse $1/H_{min}(z)$ are causal and stable. An all-pass system has a unity magnitude response (i.e. it does not apply a gain or attenuation to any frequencies) but it does change the phase relationship between frequencies.

A FIR filter can be made minimum phase by factorising it into its constituent poles and zeros and mirroring those which are outside the unit circle inside it to their conjugate reciprocal positions. Long filters (such as HRIRs) cannot be easily factorised into their component poles and zeros. The minimum phase condition however implies that the real and imaginary components of the frequency response are related by the Hilbert transform as follows:

$$arg[H(j\omega)] = -\mathcal{H}\{log(|H(j\omega)|)\} \tag{3.2}$$

As such the minimum phase version of a filter is generally calculated using real cepstrum analysis [141]. Matlab's rceps function implements this.

An interesting property of the minimum phase condition is that when applied, it shifts the energy towards the beginning of the impulse response. A non minimum phase filter and its minimum phase equivalent have the same magnitude response in the frequency domain but the time domain impulse response of the minimum phase filter will have its energy more concentrated at the start of the response than its non minimum phase counterpart. An impulse response whose energy profile is optimally compact can be truncated with less information loss than would otherwise be the case. For further reading on minimum phase see [141].



(a) Non minimum phase.

(b) Minimum phase.

(c) Frequency domain magnitude comparison.

(d) Frequency domain phase comparison.

Figure 3.4: Comparison of minimum phase and non minimum phase HRIR.

HRIRs are frequently approximated by their minimum phase representation. Figure 3.4 shows an example of the use of the minimum phase assumption on a HRIR (taken from Subject 3 in the CIPIC database [6]). Mehrgardt and Mellert [117] first introduced the concept of approximating the response of the external ear as a minimum phase system. The authors declare the all-pass component of the outer ear transfer function to be almost linear up to 10kHz. Kulkarni et al. [103] and Kistler et al. [98] found that the minimum phase version of the impulse response along with a separate linear delay component was a sufficiently accurate representation of the HRTF from a psychoacoustic viewpoint. Hence Equation 3.1 can be rewritten as

$$H(j\omega) = H_{min}(j\omega).e^{-j\omega\tau} \tag{3.3}$$

where τ is the delay replacing the all-pass component.

However Avendano et al. [8] state that contralateral HRIRs often exhibit non minimum phase properties. Plogsties et al. [147] propose that such contralateral HRIRs should include

an additional delay if the minimum phase approximation is used. Nam et al. [135] used the cross coherence between minimum phase and unprocessed HRTFs as a measure of the accuracy of the minimum phase assumption. The authors concluded that, in the majority of cases, the HRTFs they examined were well modelled by their minimum phase equivalents and a pure delay component. However they did note some non minimum phase behaviour among front and ipsilateral directions. The use of the minimum phase assumption on HRIRs also removes the ITD information and necessitates a separate method of ITD calculation. This can be done using techniques such as physical modelling of the head and torso, onset detection, IACC etc. See 2.2.1 for further details.

The ambiguity regarding how ITD is detected in the brain (see Section 2.3) and the limitations of the minimum phase approach described above are a motivation to consider leaving the HRIR intact, with no minimum phase approximation. There is a trade off here between the computational efficiencies offered by the minimum phase assumption and the possible loss of perceptually important information.

3.4 HRTF Interpolation

In the Section 3.2.1 the tedious nature of HRTF measurement was discussed as was the resulting effect on the angular resolution of HRTF measurement grids. Accurate modelling techniques are also extremely computationally expensive and require high quality digital scans. As was discussed in Section 2.4 the minimum audible angular source separation can be as low as 1° for frontal source positions. Hence, a high spatial resolution of HRIRs is necessary for convincing virtualisation and, as such, spatial interpolation is a useful tool. Even if a high resolution grid is available there may not be sufficient memory available to store this information. It may be preferable to store a less spatially dense grid and interpolate in real time.

A common approach to interpolation is linear interpolation. Time domain linear interpolation is probably the simplest and one of the most effective approaches. Nishino et al. [137] demonstrate that this technique performs better than a neural network and PCA based approaches. It is necessary to remove the initial onset from the HRIR by finding its minimum phase equivalent or a similar technique. The linear delay is calculated by cross correlation of upsampled versions of the raw and minimum phase HRIRs. Upsampling is necessary to take into account fractional delay. The time aligned HRIRs can then be linearly interpolated, as can the separate delay component, using a simple weighting based on the inverse distances between the new spatial position and the existing measurement positions. The weighted linear interpolation is shown in Equation 3.4 with the distances e and f as shown in Figure 3.5.

$$HRIR_X = \frac{f}{e+f} HRIR_A + \frac{e}{e+f} HRIR_B \tag{3.4}$$

This can be easily extended to interpolation between 3 or 4 measurements on a 2D grid as shown

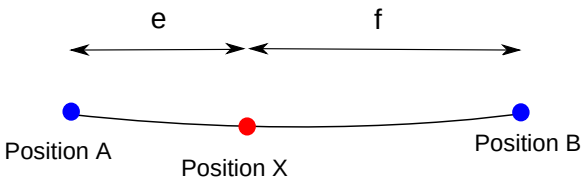


Figure 3.5: Interpolation on a line.

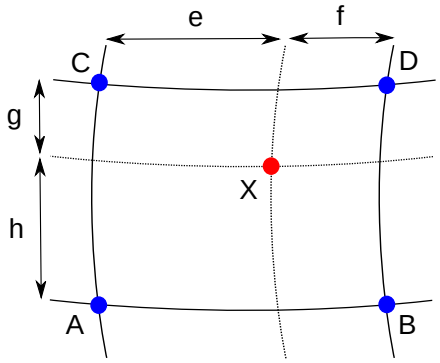


Figure 3.6: Interpolation on a grid.

in Figure 3.6 as opposed to 2 measurements on a line. This is called bilinear interpolation [160], [53] and is demonstrated in Equation 3.5.

$$HRIR_X = \frac{f}{e+f}.\frac{g}{g+h}HRIR_A + \frac{e}{e+f}.\frac{g}{g+h}HRIR_B + \frac{f}{e+f}.\frac{h}{g+h}HRIR_C + \frac{e}{e+f}.\frac{h}{g+h}HRIR_D \tag{3.5}$$

These distance based weights are applied in the same way to interpolate the onset delays calculated via cross correlation. Linear interpolation is also often applied in the frequency domain to the magnitude response [30]. The minimum phase assumption is generally applied in this case and a separate pure delay component, which is replacing the all-pass component, must be separately interpolated in order to maintain the ITD. The minimum phase component can also be applied before time domain interpolation in the same way.

Kistler and Wightman [98] demonstrate that 90% of the variance in the HRTF magnitude functions can be modelled by a linear combination of 5 basis functions obtained by principal component analysis (PCA). It is possible to interpolate across the weights attributed to the basis functions in order to create new HRTF data. Chen et al. [34] implement such an technique on complex valued HRTF data. The authors use the Karhunen-Loeve Expansion (KLE), a more generalised form of PCA where complex valued data can be used. The weights obtained are termed 'spatial characteristic functions' (SCFs) and a two dimensional spline is fitted to them to perform the interpolation. Results on HRTF data measured for the KEMAR dummy head and a cat generally show errors of less than 1%. The exception to this is the contralateral HRTFs of

the dummy head and the lower back area of the cat where larger errors occur. Wu et al. [189] also use the KLE but apply it to the time domain HRIRs and use a more simple linear interpolation technique instead of spherical splines. The HRIR data is normalised by removing time and level differences before the KLE is applied. Wu et al. present their results in the same form as Chen et al. allowing for easy comparison of the two. The linear and spline based approaches produce broadly similar errors with the linear method giving lower error in the rear lower elevation area. Carlile et al. [27] apply splines in a similar way to PCA weights. PCA was applied solely to the frequency domain magnitude response in this case. However for large datasets there was no discernable improvement over the simple nearest neighbour linear interpolation technique.

Further detail on the use of spherical thin plate splines for interpolation can be found in the following: [95, 152, 176]. Spherical splines combine the idea of periodic splines on a circle and 2D thin plate surface splines. A periodic spline is a closed seamless shape that has no beginning or end point. The theory behind this form of interpolation is very mathematically intensive and while it may seem like an elegant solution to the problem as it takes into account the spherical nature of the grid of data collected, it is quite complex and questions must be raised as to whether it is beneficial to take the entire grid into account and not just data local to the HRTF we are trying to find.

More recently Wang et al. [178] proposed another method for PCA weight interpolation. The authors propose a multivariate polynomial fitting technique to approximate the weights with a bivariate polynomial of azimuth and elevation angles. This is then optimised by a sphere partitioning optimisation scheme. This technique is compared to both linear and spline interpolation of the weights and is found to give better results, especially for contralateral positions where both linear and spline interpolation throw up significant errors.

The plenacoustic function is another option for HRTF interpolation [4]. Consider the case of a loudspeaker and a circular array of uniformly spaced microphones. The plenacoustic function on the circle is a continuous function that encaptures all the room impulses responses from the position of the source to all possible microphone positions. Knowledge of this function would allows for interpolation between the microphone positions. The HRTF setup is the reciprocal of this i.e. the microphones are replaced by speakers and the speaker by an in ear microphone (left and right ears are considered separately). The use of the plenacoustic function to interpolate HRTFs is explained in detail in [3].

Discrete Fourier transform (DFT) was used by Matsumoto et al. [114] to interpolate HRIRs. The interpolation is carried out samplewise across the dataset. Consider the case where the HRIRs are inserted as the columns of a matrix. Columns of zeros are inserted between the HRIRs and the DFT is applied to each row. The inverse DFT of each row is then found and the matrix columns which previously contained zeroes now contain spatially interpolated HRIRs for positions between those that where known. However the authors found linear interpolation with arrival time correction to perform better than DFT interpolation when compared.

Keyrouz et al. [95] use a rational state space approach to do the interpolation. Their method

is based on the factorization of a block Loewner matrix into a product of generalized observability and controllability matrices. From this a minimal state space realisation of an interpolating matrix transfer function is obtained. Like the DFT approach this method takes into account the whole dataset and not just the local data when interpolating. The authors find this method to perform better than DFT and spline based techniques.

Pole-zero modelling also allows for the interpolation of HRTFs. Runkle et al. [156] implements interpolation of the poles and zeros both linearly and along poles and zeros found by a gradient search algorithm. Zhang et al. [194] use continuous Fourier-Bessel functions to model the HRTF in the horizontal plane and report an error of less than 2% when comparing measured and modelled HRTFs. Spherical harmonic fitting has recently gained popularity as a HRTF interpolation technique. Evans et al. [49] use a surface spherical harmonic representation for a Gaussian grid of HRTF measurements and find that interpolated data is close fit to measured data at the interpolated positions. More recently Zotkin et al. [195] present two methods of finding the spherical harmonic representation: direct integration and a least squares fitting method. The least squares method allowed for the use of arbitrary grid layouts which allows for its use with existing datasets.

3.5 HRTF Reduction Techniques

HRTF reduction is motivated by the need for efficient real time dynamic virtual audio reproduction. A reduction in HRIR length implies less computational processing resources are required for a spatial audio implementation. The need for large, dense spatial grids of HRTFs has been demonstrated in Section 2.4. This has consequences for the memory capabilities of the system requiring this data which can also be eased by the reduction or simplification of HRIRs.

As was discussed in Section 3.4 Karhunen-Loeve expansion and principal component analysis are frequently used reduction techniques for HRTFs [27,34,98,189]. PCA is the decomposition of data to a linear combination of orthogonal functions. It is optimal for normally distributed data as a Gaussian assumption is implicit in the algorithm. Leung and Carlile [110] compare different HRTF formats with regard to the application of PCA and find the frequency domain representation with linearly scaled amplitude to be optimal. Kapralos et al. [92] implement HRTF data reduction using non linear techniques such as local linear embedding (LLE) and Isomap and find it to be superior to PCA. These techniques utilise local neighbourhood information in order to form a nonlinear basis for the HRTF data being examined. LLE identifies local neighbourhood distance relationships and computes a lower dimensional mapping that preserves them. A set of weights is determined at each input point which approximate it in the least-squares sense from its neighbouring points. It uses an eigenvector-based optimization technique to find the low-dimensional embedding of the points, such that the relationship between each point and its neighbours is preserved. Isomap is based on the assumption of isometry of geodesic distances in the manifold. Each point is connected to its nearest neighbours to form a graph and pairwise

geodesic distances are approximated on the graph. Metric multidimensional scaling is used to recover a low dimensional isometric embedding.

Grindlay et al. [67] applied tensor singular value decomposition (Tensor-SVD) to HRTF reduction. Singular value decomposition (SVD) forms the basis for PCA. However in this case finer control of the reduction is allowed as the basis vectors are truncated separately. Grindlay found this technique to significantly outperform PCA. Rothbucher et al. [155] apply Tensor-SVD and generalised low rank approximations of matrices (GLRAM) to HRTF reduction and compare their perform with that of PCA. GLRAM is a simplified, more computationally efficient form of Tensor-SVD. They found both these techniques to significantly outperform PCA reduction and the performance of optimised GLRAM was on par with Tensor-SVD.

Balanced model truncation (BMT) is another technique used for reduction. Mackenzie et al. [112] use BMT to design low order IIR models to HRIRs. They apply an algorithm formulated by Beliczynski [13] specifically to HRIRs. The original lengthy FIR filter is converted into state space form and a transformation matrix is obtained such that the controlability and observability grammians are equal and diagonal (i.e. the system is balanced). The states are then ordered by the magnitude of their contribution, which is quantified by their hankel singular values, and the least important are disregarded. The authors reduce the order of 512 tap HRIRs to 128 taps initially using the minimum phase assumption and truncation of the impulse responses. These 128 tap filters are subsequently reduced to 10th order IIR filters using the BMT technique. However Mackenzie et al. use only the minimum phase versions of HRIRs and also apply critical band smoothing before the BMT is conducted. Both of these steps result in an inherent loss of some of the information contained in the HRIR before the reduction technique is applied.

The state space approaches of Beliczynski and Mackenzie et al. described in the previous paragraph are applied only to single HRTFs. Hence the computational cost still increases linearly with the number of filters in the dataset. Haneda et al. [74] propose a method for extracting common acoustic poles from HRIR datasets, such that each HRIR is represented by an IIR filter and a FIR filter. They model HRIRs using common poles which are independent of source direction and zeros which are dependent on direction. The common poles are considered to represent a resonance system in the pinna and ear canal and are estimated as the autoregressive coefficients for a HRTF set. Georgiou and Kyriakakis [59] elaborate on this work through the use of a MISO (Multiple Input Single Output) state space system to create a combined model of the HRTFs for all directions simultaneously for each ear. Grantham et al. [65] further elaborate on this by modelling full sets of horizontal HRTFs for both ears using high order state space systems. Adams and Wakefield [1] compare these approaches and find the MISO architecture proposed by Georgiou and Kyriakakis to be the best compromise as it allows ITD to be modelled separately and multiple monaural sources to be rendered together.

Huang and Liu [80] use independent component analysis (ICA) and BMT to achieve significant length reductions. ICA is used to extract independent spatial features from the HRIRs. The independent basis vectors are obtained using the second-order blind identification (SOBI)

algorithm [15]. BMT is then implemented on the ICA reduced FIR HRIRs to produce shorter IIR filters.

The wavelet transform is used as a spectral smoothing tool by Hacihabiboglu et al. [71]. The authors obtained the minimum phase version of the wavelet smoothed filters and designed IIR filters for each HRTF using Prony's method. Torres et al. [171] also apply the wavelet transform to HRIR data. After initially reducing the HRIRs from 512 to 128 samples by removing the initial delay and truncating them, the Daubechies wavelet is applied and an analysis of the energy content is conducted to determine how the model can be reduced to limit the error introduced to the responses.

3.6 Conclusion

This chapter has examined the concept of the HRIR in detail. The physiological basis for these filters and techniques for their measurement and synthesis have been introduced. Two methods to equalise HRTF datasets were described and compared. The exploration of the minimum phase assumption raised questions regarding the suitability of its application to HRIR simplification which will be of particular interest in the proceeding chapter. A review of existing HRIR interpolation and reduction techniques was undertaken.

4

HRIR Order Reduction using Approximate Factorisation

4.1 Introduction

Long HRIRs can be problematic in real time virtual auditory environments. They impose significant computational and memory load on the system. Shorter filters would ease this computational burden and also allow for easier switching of HRIRs due to relative position change between the source and listener. To this end a novel factorisation technique has been developed which allows for a direction independent component to be extracted from sets of HRIRs. Each individual filter is split into two filters which, when convolved together, result in a close approximation to the original filter. The convolution with the common factor could be completed offline and stored leaving a shorter HRIR that would change with a relative movement between source and receiver.

Two regularisations will also detailed which allow for a more robust, initial condition independent process. As the approximate factorisation does not yield a unique common component, the regularisations are aimed at constraining this solution. One of these is suitable for minimum phase HRIR data and will allow for very short direction dependent components to be obtained. The other is suited to raw HRIR data and will allow for the initial time delay to be maintained in the direction dependent components. The base algorithm will be described in the following section and results provided for its application to HRIR data from the CIPIC database. The regularised derivatives of the algorithm will be described in Section 4.3 and their application to

data from the CIPIC dataset and Gardner and Martin's Kemar dataset will be examined.

4.2 The Algorithm

Here it is proposed that a set of HRIRs (denote h^ϕ) be simplified by factoring each filter into the convolution of a direction independent subsystem (denote f) which is common to the whole set and a direction dependent residual (denote g^ϕ). The algorithm used in finding this common subsystem of a HRIR dataset is equivalent to finding the approximate greatest common divisor (AGCD) of the HRTF z-domain set. The task of finding the AGCD is formulated as a non linear optimisation problem:

$$\min_{f,(g^1,\cdots,g^N)} \sum_{\phi=1}^{N} \|h^\phi - (f * g^\phi)\|^2 \tag{4.1}$$

$$\begin{aligned} \text{where} \qquad h^\phi &= [h_0^\phi, ..., h_{m-1}^\phi]^T, \\ g^\phi &= [g_0^\phi, \ldots, g_{j-1}^\phi]^T \\ f &= [f_0, f_1, \ldots, f_{k-1}]^T \end{aligned}$$

The divisor-quotient iteration method is a variant of the well-known Gauss-Newton non-linear least squares algorithm with the exception that the usual step of linearisation around the current guess is already done, as the system is bilinear i.e. by holding f constant, the system is linear in g^ϕ, and vice-versa. Given an initial guess for f, standard least squares can be used to find the residues, g^ϕ, which minimise the error between $f * g^\phi$ and h^ϕ. This g^ϕ can then be used to generate a refined f again using least squares and hence a recursive process is defined.

Divisor-Quotient iteration

i=iteration count

1. Guess f_0 $(i = 0)$
2. Solve for each residual, g^ϕ, as follows:

$$g_{i+1}^\phi = F_i^\dagger h^\phi \tag{4.2}$$

Where F_i is the convolution matrix formed from f_i and † denotes the Moore-Penrose pseudoinverse.

3. Solve for f_{i+1} using

$$f_{i+1} = \begin{pmatrix} G_{i+1}^1 \\ \vdots \\ G_{i+1}^N \end{pmatrix}^\dagger \underline{h} \quad \text{where} \quad \underline{h} = \begin{pmatrix} h^1 \\ \vdots \\ h^N \end{pmatrix} \tag{4.3}$$

G_{i+1}^ϕ is the convolution matrix formed from g_{i+1}^ϕ

4. Set $i = i + 1$ and repeat steps 2 and 3 until there is convergence.

4.2.1 Convergence

Authors such as Zeng [193], Corless et al. [37] and Chin et al. [35] have published extensively in the area of determining AGCDs of polynomial sets. They establish methods of approaching this problem for small numbers of polynomials of relatively short order (generally less than tenth order). Even in these limited circumstances there is no guarantee of convergence to a global minimum. In this thesis such methods are applied to polynomials of order 200 or greater of which there can be hundreds, even thousands and, as such, finding a global minimum is very unlikely. Our proposed iterative least squares method is equivalent to the divisor quotient method described by Chin et al [35] and Corless et al. [37] wherein they provide a proof of convergence to a point on the mean square error surface with gradient zero, i.e. a local minimum or maximum.

4.2.2 Results of Tests on CIPIC Database

The CIPIC database [6] is a public domain HRIR database which consists of 1250 HRIR measurements for each of 45 subjects. Each 200 sample long HRIR is measured at a location on a sphere of radius one metre centred on the subject head and is sampled at 44.1kHz. The results displayed below are using the left ear HRIRs from subjects 3 and 21. Subject 3 is a human subject while subject 21 is the KEMAR dummy head with large pinna.

Due to the size of the dataset, results for each spatial position will not be shown. The results for a representative group of responses are reproduced here instead. Figures 4.1 and 4.2 show reconstructed HRIRs, at a variety of positions on the azimuth, where a 100 sample long common subsystem has been extracted from the whole dataset (1250 measurements) and compares them to the original unfactorised HRIRs. There are three different initial guesses used in these examples. The first initial guess is all ones. The second initial guess is the first 100 samples of an average taken over the entire HRIR dataset for the relevant subject and ear and the third is a 100 sample long random vector generated by Matlab's rand function. The elements of the vector are drawn from the standard uniform distribution on the open interval (0,1). The same random vector is used in each case. The comparison is shown in both the time and frequency domain (magnitude (dB)). It is evident from the fact that only the blue

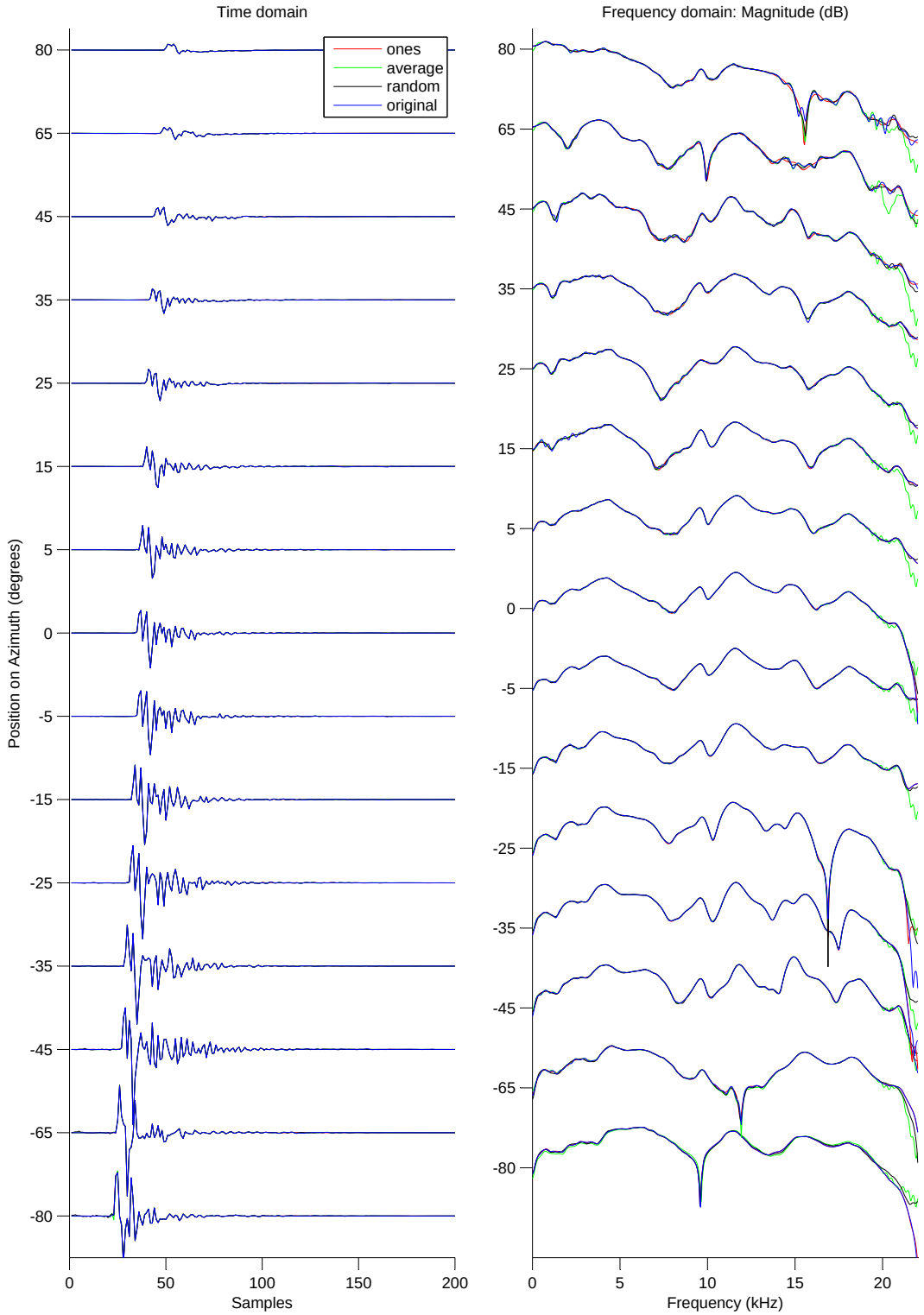


Figure 4.1: Subject 21. Comparison of original HRIR to reconvolved HRIRs with different initial f_0 guesses.

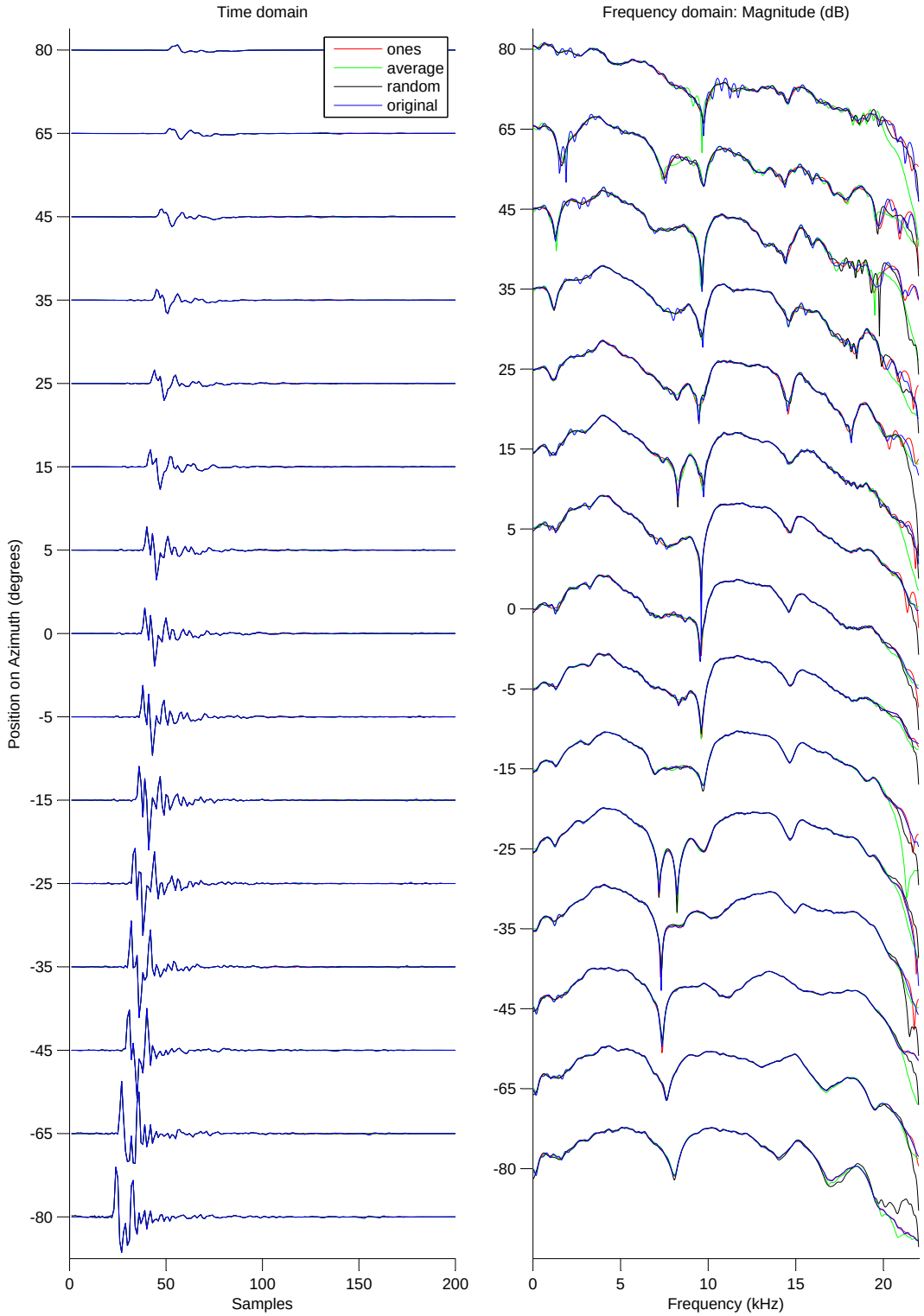


Figure 4.2: Subject 3. Comparison of original HRIR to reconvolved HRIRs with different initial f_0 guesses.

line (which denotes the original HRIRs/HRTFs) is visible for most cases, that there is negligible difference between the original and the reconvolved HRIR for each initial condition, given a 100 sample long common component has been extracted. In Figure 4.1 there is a small mismatch visible in the frequency spectrum for each position on the azimuth at high frequencies (>18kHz), especially when the average initial condition is used. There is also some distortion visible at the 16kHz notch for the 80° HRIR. Figure 4.2 shows the same high frequency distortion for subject 3. It also appears that the reconstruction is better for all initial conditions for the ipsilateral HRIRs for both subjects i.e. those HRIR which are on the left side of the head ($-80°$ to 0° in the azimuth). Nonetheless the reconstructed HRIRs/HRTFs are still very similar to the original ones. As there is no blurring or significant distortion of the initial peak of the HRIRs the ITD will be maintained between the left and right ear responses.

Figure 4.3 shows the HRTFs for uniformly spaced postions in elevation from -45° to 230.625° at 0° azimuth for Subject 21. Figure 4.4 shows the reconstructed HRTFs again after a 100 sample long direction independent component had been extracted using the average initial condition. Figure 4.5 shows the difference in dB between the two. There is no significant distortion of the HRTFs as a result of the factorisation process except at very high frequencies (mostly >20kHz). Figures 4.6, 4.7 and 4.8 show the original HRTF, the reconstructed HRTFs and the error for Subject 3. As was the case for Subject 21 there is no significant distortion introduced by the factorisation process.

Rather than seek to extract a common subsystem of a given length from the entire measurement sphere, it is also reasonable to consider the extraction on a quadrant or octant of the sphere. Also consideration can be given to the length of the common subsystem. In the above examples a length of 100 samples is used and yields good results. The choice of length is a trade off between computational capacity available for real-time convolution and the error in the reconstructed HRIR.

Figures 4.9 and 4.10 plot the mean squared error versus the length of the common subsystem for subjects 21 and 3 respectively. The mean square error measure describes the difference between the entire original HRIR dataset (1250 HRIRs, each 200 samples in length) and the reconvolved HRIRs for each initial condition. One would expect the error to be monotonically increasing with the length of the common component and the average and random initial conditions generally follow this profile for both subjects. There is however a pronounced notch in the error profile for the ones initial condition which occurs in the region of 140 to 180 samples for each case.

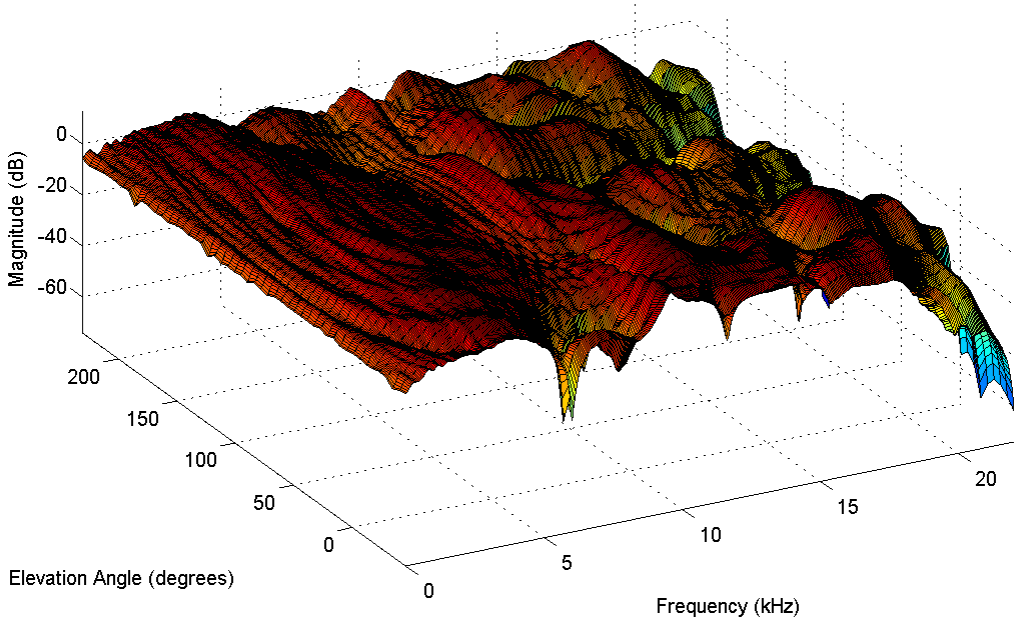


Figure 4.3: Subject 21: HRTFs, varying elevation, 0° azimuth.

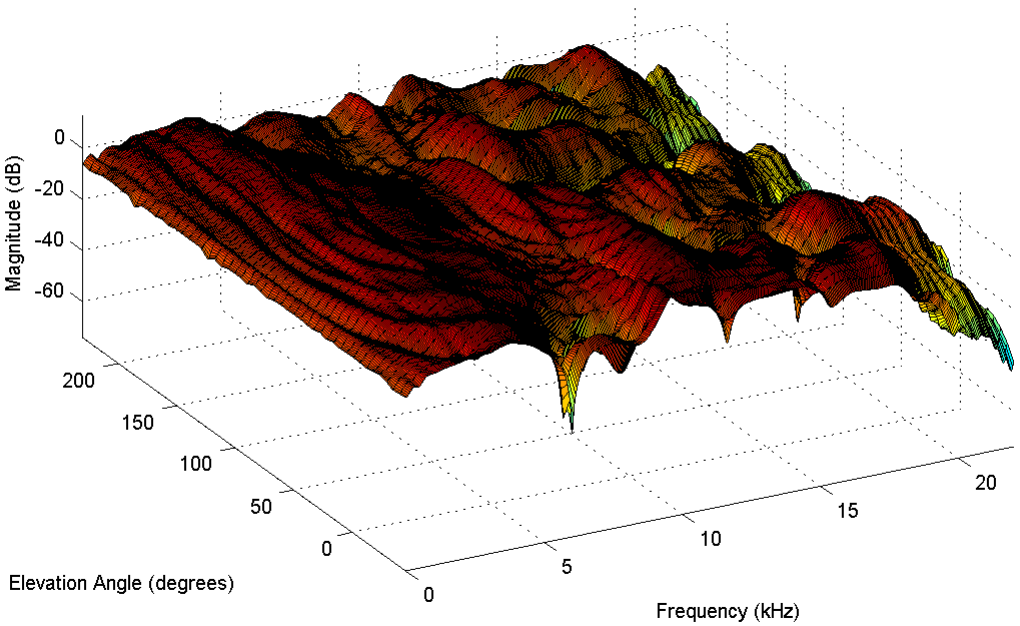


Figure 4.4: Subject 21: Reconstructed HRTFs.

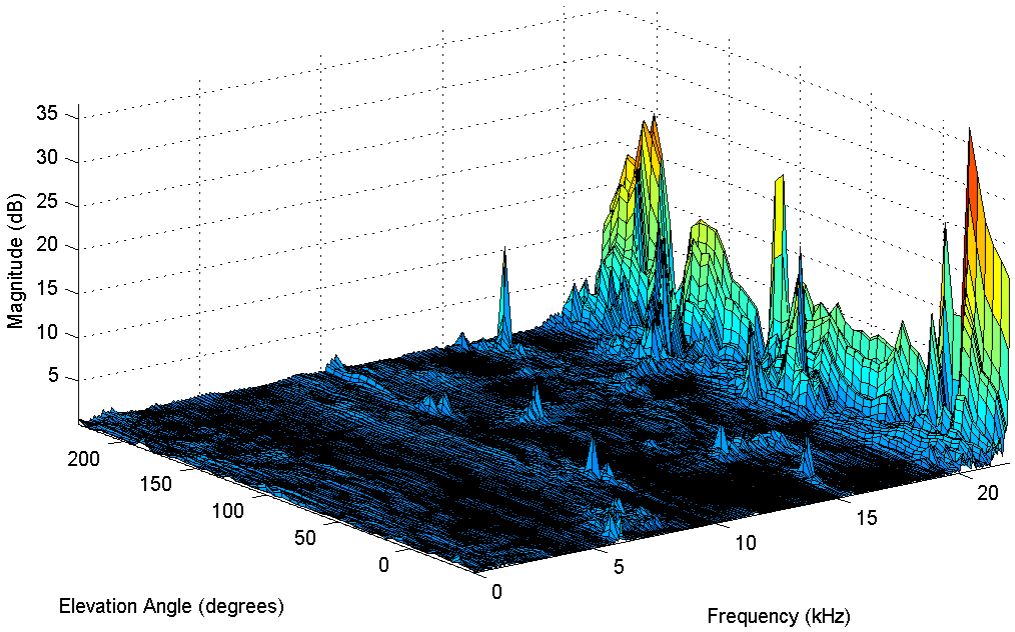


Figure 4.5: Subject 21: Error between original and reconstructed HRTFs.

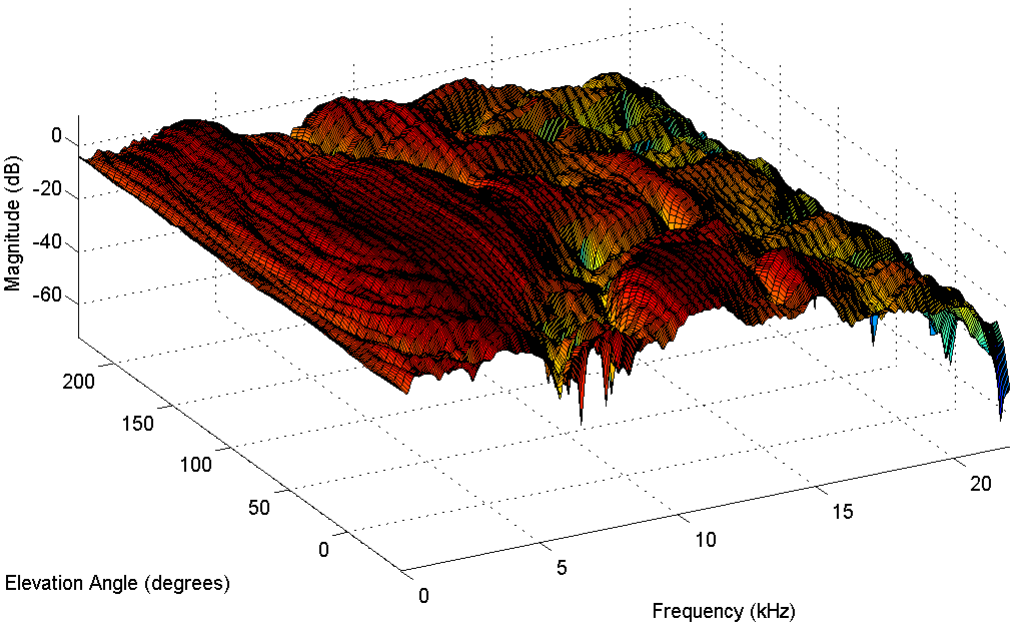


Figure 4.6: Subject 3: HRTFs, varying elevation, 0° azimuth.

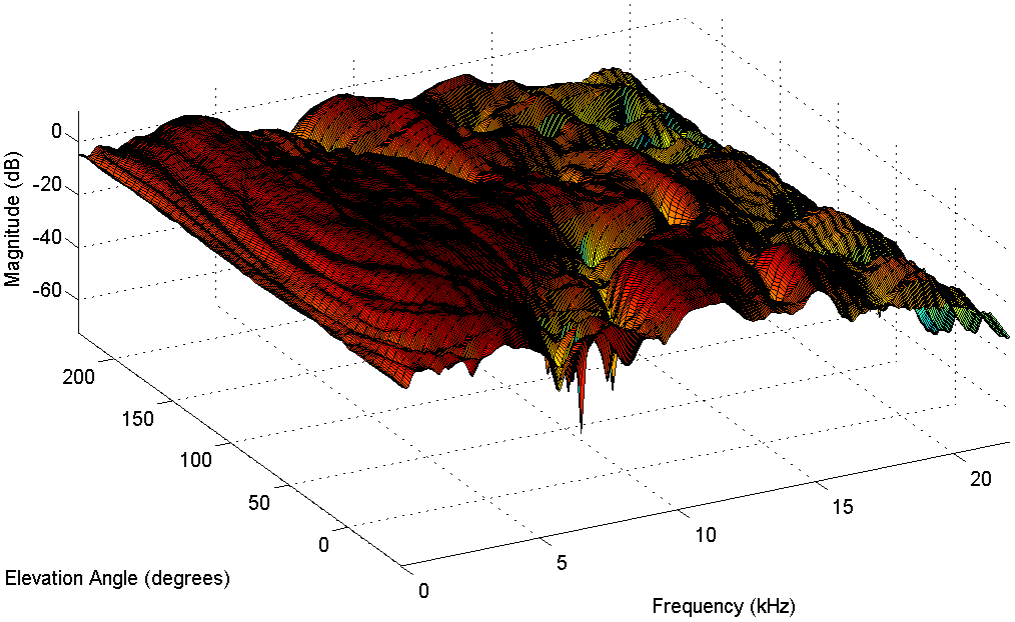


Figure 4.7: Subject 3: Reconstructed HRTFs.

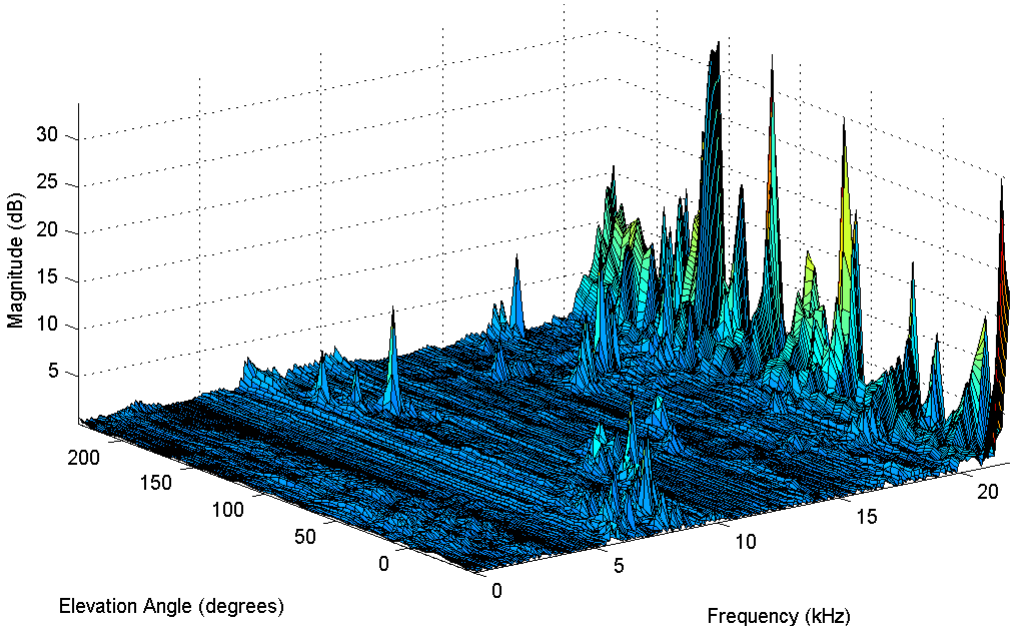


Figure 4.8: Subject 3: Error between original and reconstructed HRTFs.

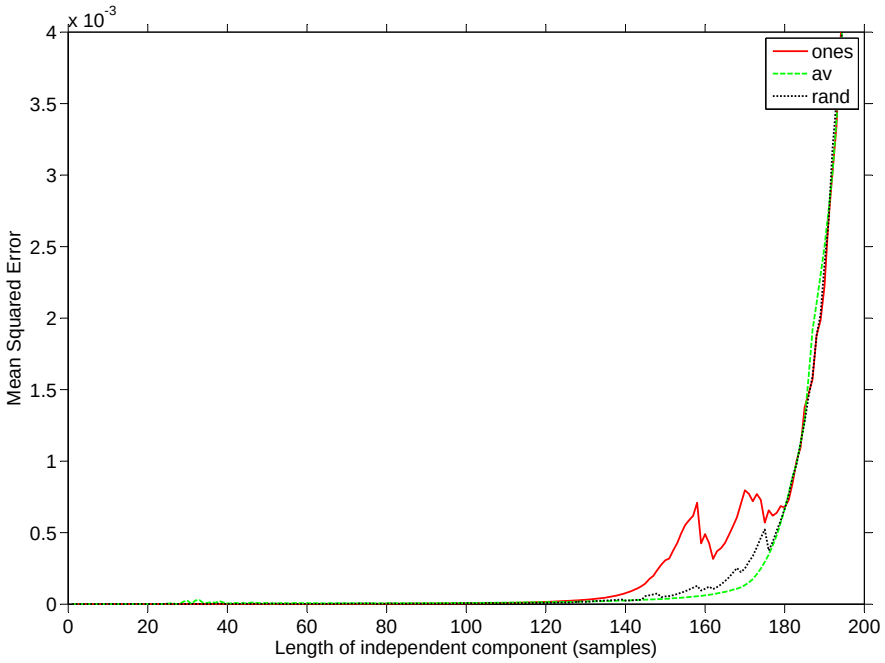


Figure 4.9: Subject 21: MSE profile for different initial f_0 guesses.

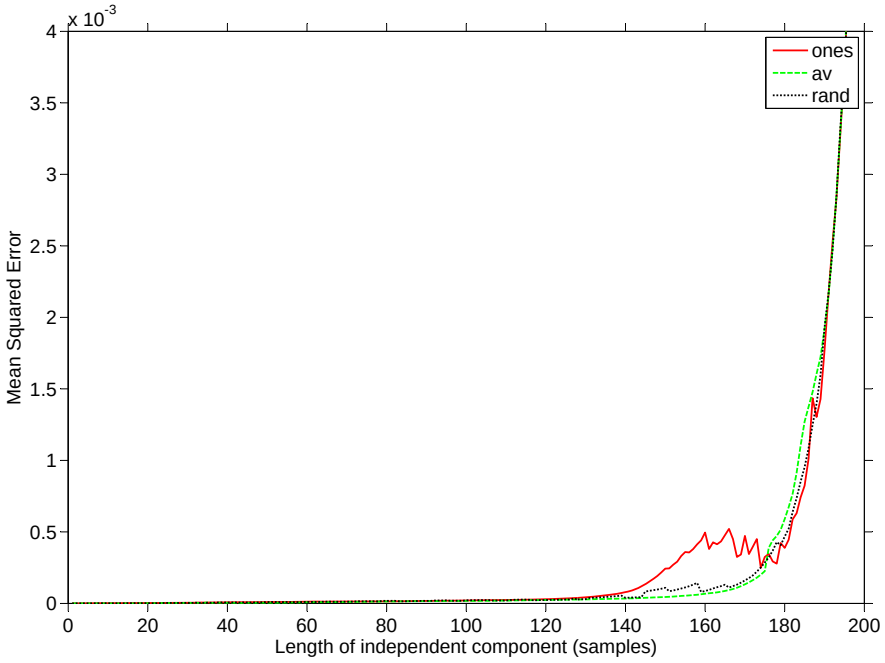


Figure 4.10: Subject 3: MSE profile for different initial f_0 guesses.

4.3 Regularisation

The factorisation algorithm, as proposed in the previous section, offers low levels of reconstruction error. However different initial guesses result in drastically different direction independent and dependent components and when the factorised components it produces are examined in isolation, they are meaningless from a psychoacoustic viewpoint. It would appear the problem is one with multiple local minima. As such, techniques to refine the factorisation process through the use of regularisation techniques are examined in this section. The aim is to find more meaningful factorised components while maintaining the low reconstruction error.

Recall the original criterion used for the optimisation (see Equation 4.1). If, instead of just minimising the difference between the original and reconstructed HRIRs, a second regularising term is added which allows for desirable properties to be imposed on f, then the criterion to be satisfied becomes as follows:

$$\min[\|\underline{h} - \underline{G}f\|^2 + \lambda\|f - f_p\|^2] \tag{4.4}$$

where f_p is the filter that is regularising the optimisation process and λ is the weighting which controls the importance or effectiveness of the regularisation. To solve this for f the modified cost function, $C(f)$, is defined and its first derivative with respect to f is set to equal zero, as follows:

$$C(f) = \|\underline{h} - \underline{G}f\|^2 + \lambda\|f - f_p\|^2 \tag{4.5}$$

$$\nabla C(f) = -2\underline{G}^T\underline{h} + 2\underline{G}^T\underline{G}f + 2\lambda f - 2\lambda f_p = 0 \tag{4.6}$$

$$(\underline{G}^T\underline{G} + \lambda I)f = \underline{G}^T\underline{h} + \lambda f_p$$

$$f = (\underline{G}^T\underline{G} + \lambda I)^{-1}(\underline{G}^T\underline{h} + \lambda f_p) \tag{4.7}$$

I denotes a $k \times k$ identity matrix. So we redefine Equation 4.3 in the iterative process as follows.

$$f_{i+1} = (\underline{G}_{i+1}^T\underline{G}_{i+1} + \lambda I)^{-1}(\underline{G}_{i+1}^T\underline{h} + \lambda f_p) \tag{4.8}$$

Conversely, it is also possible to regularise with respect to the direction dependent component.

$$\min[\|h^\phi - Fg^\phi\|^2 + \lambda\|g^\phi - g_p^\phi\|^2] \quad \text{for} \quad \phi = 1 \ \text{ to } \ N \tag{4.9}$$

In this case Equation 4.2 in the iterative process is redefined as follows:

$$g_{i+1}^\phi = (F_{i+1}^T F_{i+1} + \lambda I)^{-1}(F_{i+1}^T\underline{h}^\phi + \lambda g_p^\phi) \tag{4.10}$$

The different regularisation strategies introduced above will satisfy situations with different factorisation requirements. If it is acceptable for minimum phase HRIRs to be used then factorising the minimum phase set with regularisation on f is most effective. f_p is set to the first k samples of the average of the minimum phase HRIR set. This gives very low reconstruction error

and studies suggest the solution is independent of the initial condition used. If minimum phase HRIRs are not acceptable due to the non minimum phase behaviour of contralateral HRIRs and the uncertainty regarding the detection of ITD in the brain as discussed in Section 3.3, then it is necessary for the delay for each spatial measurement position to be maintained in the direction dependent component. Hence it is preferable to use regularisation on g^{ϕ}. For each position ϕ, g_p^{ϕ} is set as an impulse occurring at the maximum of the HRIR. The impulse's magnitude is this maximum value. For each regularisation case λ, the weighting applied to the regularisation, is varied from a very large value at the start of the iterative process ($\sim 10^3$) to a very small value ($\sim 10^{-3}$) at the end. This allows for the correct local minimum to be established at the beginning of the process and the original least squares criterion to be given priority at the end.

4.3.1 Results when Applied to KEMAR HRTF Data

The regularised factorisation techniques described above were applied to a sample set of KEMAR HRIRs [58]. A set of 72 HRIRs is taken from the left ear dataset. Each HRIR is 512 samples long and sampled at 44.1kHz. The HRIRs used were those for zero degree elevation with a uniform 5° spacing in the azimuth from 0° (straight ahead) to 355°. For both factorisation cases 20 iterations of the algorithm are run.

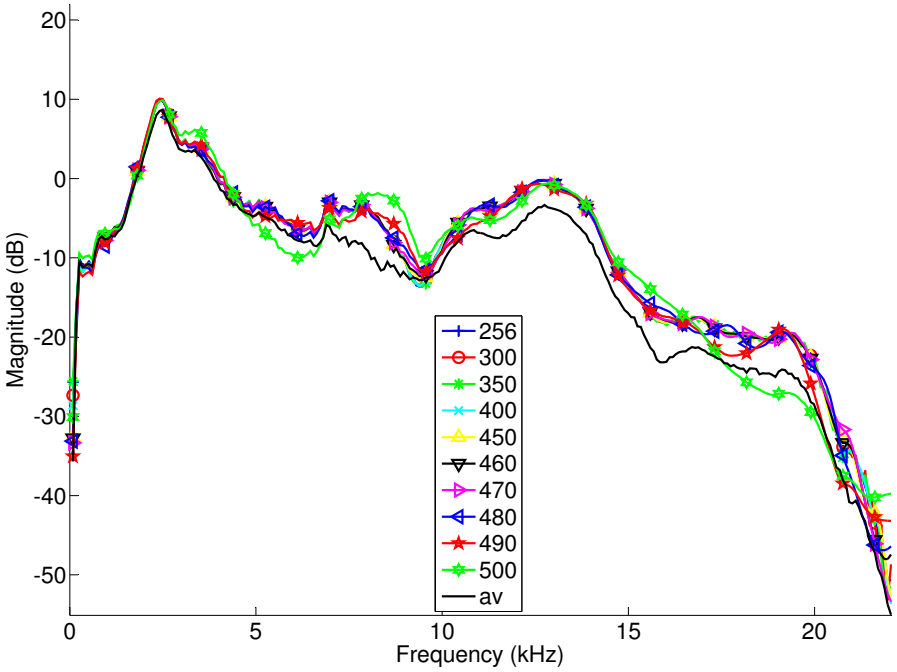


Figure 4.11: Different length f components for regularisation on f of minimum phase data.

Figure 4.11 shows the magnitude frequency response of different length common components, f, extracted from the minimum phase dataset using weighted regularisation on f in the factorisation process, while Figure 4.12 shows the same when weighted regularisation on g^{ϕ} is

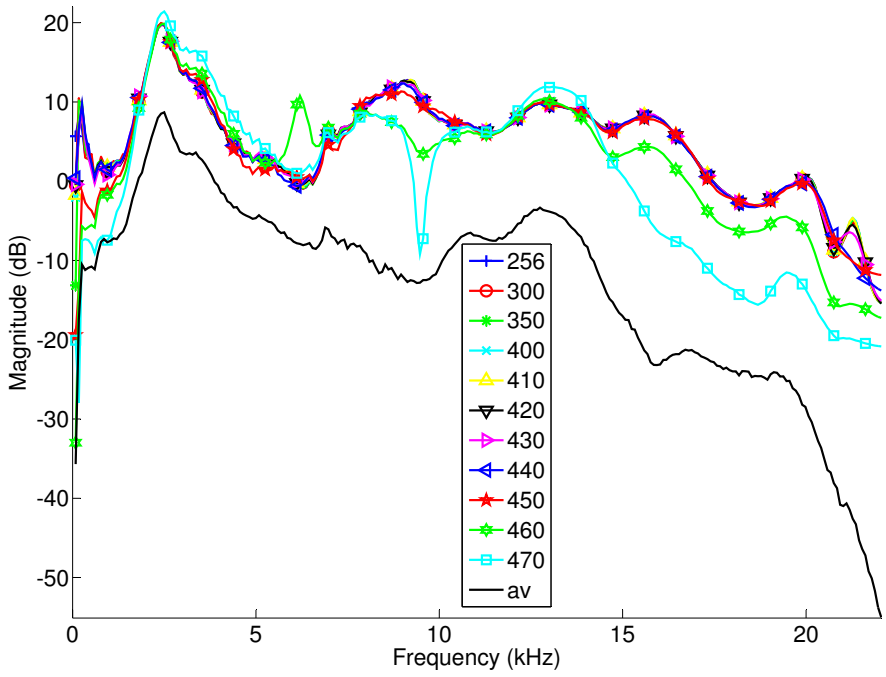


Figure 4.12: Different length f components for regularisation on g^{ϕ} of non minimum phase data.

employed on the non minimum phase dataset. The black line with no markers on each graph indicates the averaged minimum phase HRTF of the set. It can be seen that different length f components maintain broadly the same spectral information and that this spectrum is very close to the average spectrum, which is the regularising agent. In Figure 4.12 only lengths of up to 470 samples for f are shown, as after this length the factorisation gives non consistent f. When one considers that regularisation on g^{ϕ} requires that g^{ϕ} be long enough to encompass the initial delay and main peak it is understandable that the length of f for this case is more limited than in the minimum phase case. Figure 4.13 shows the full HRTF set. Figures 4.14 and 4.15 show the reconstructed HRIR set after a 256 and 470 sample long f have been extracted respectively. The regularisation on f is employed in the factorisation process. Figures 4.13 and 4.14 are almost identical showing that extracting a 256 sample component causes, as expected, minimal loss of information. When one examines a more extreme case where a 470 sample long independent component is used in Figure 4.15 the reconstruction is still very good but there is a definite smoothing of the spectrum. In the time domain this manifests itself as a loss of late reflection information. However the first 100-150 samples which contain most of the key information of the HRIR are still accurate. Figure 4.16 and 4.17 show the phase of the full minimum phase set and the reconstructed set after a 470 sample long f is extracted. There is some smoothing of the phase spectrum but there is no significant distortion introduced by the factorisation process.

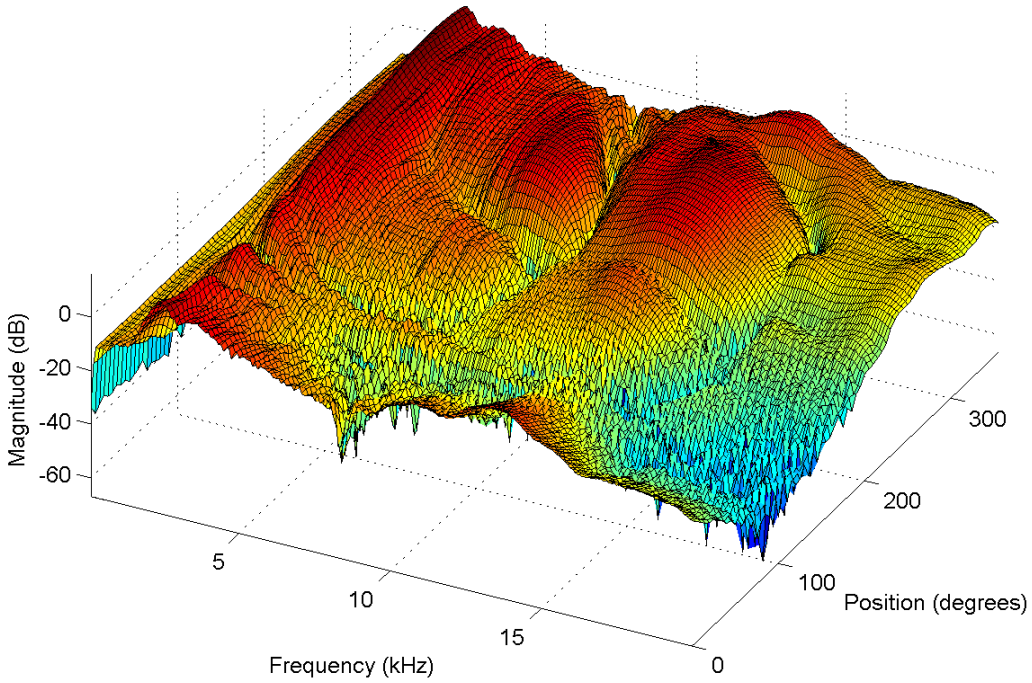


Figure 4.13: Original HRTFs.

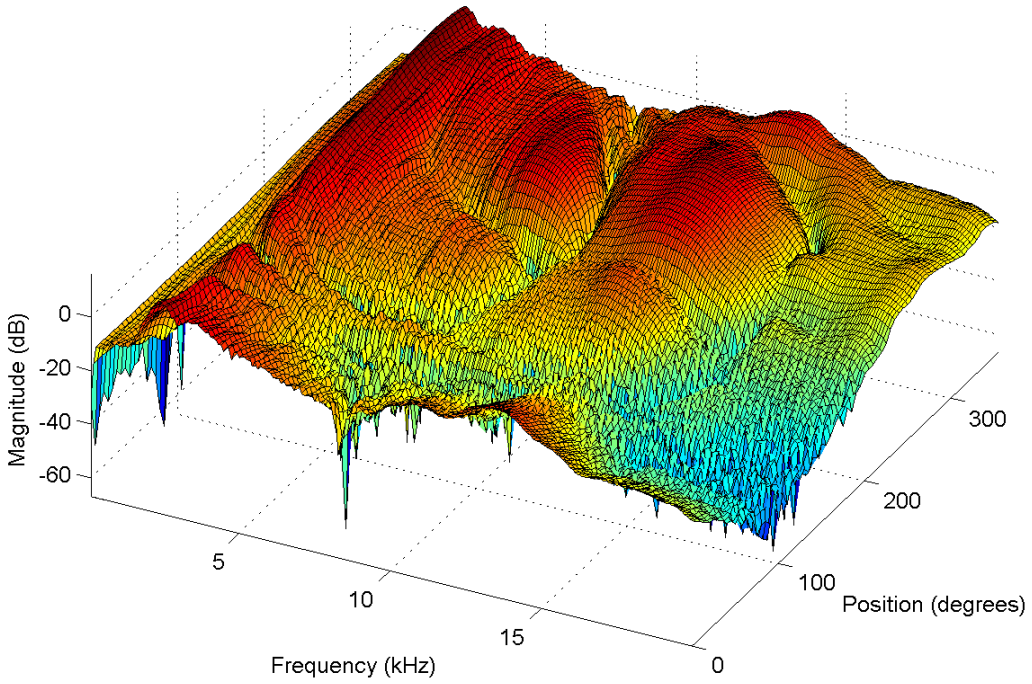


Figure 4.14: Reconstructed HRTFs. Length $f = 256$.

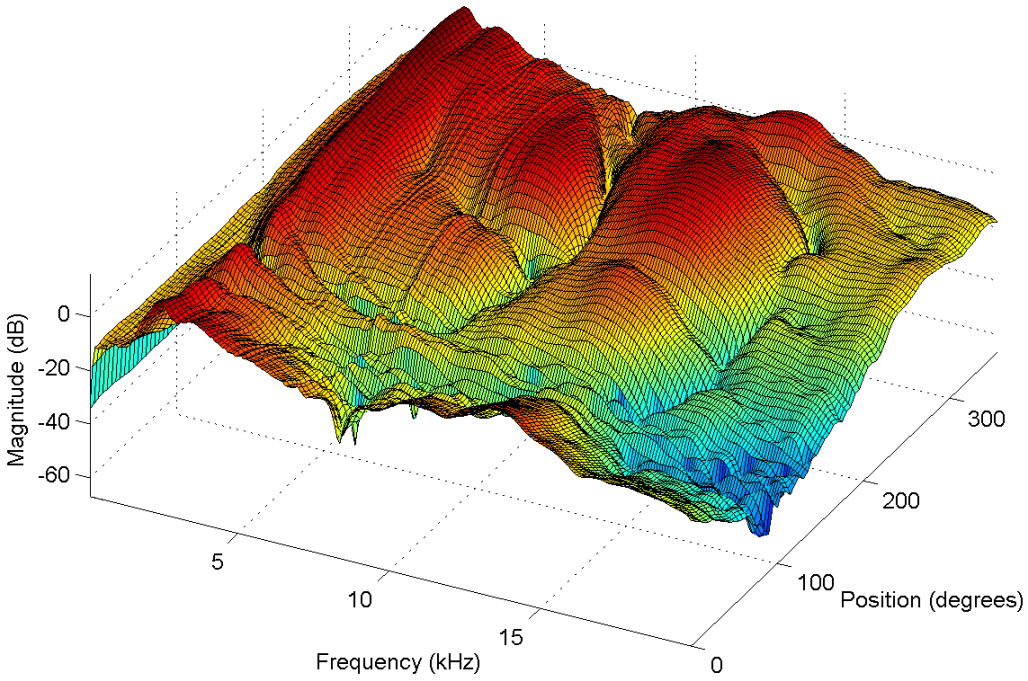


Figure 4.15: Reconstructed HRTFs. Length $f = 470$.

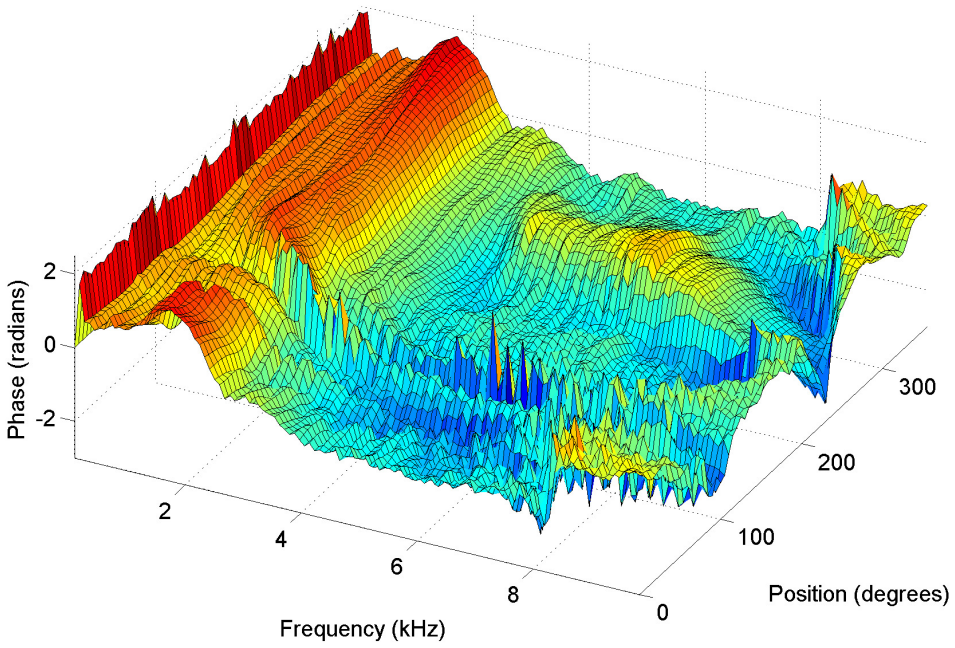


Figure 4.16: Phase of original minimum phase set.

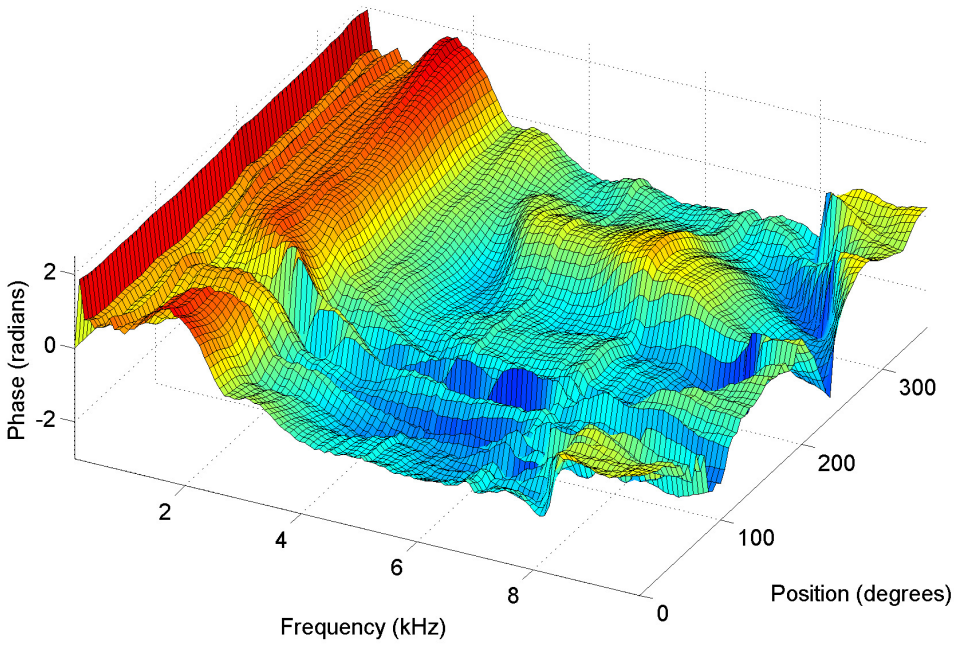


Figure 4.17: Phase of reconvolved set. Length $f = 470$.

A study was conducted to explore the capability of the factorised system to preserve important interaural time differences. The full non minimum phase HRIR dataset, as shown in Figure 4.18 was used. Figures 4.19 and 4.20 show the reconvolved HRIRs when regularisation on g^{ϕ} is implemented, with a 256 sample and a 430 sample long f extracted. The 256 sample long case shows near perfect reconstruction when compared to the original HRIRs in Figure 4.18. For the 430 sample long case the reconstruction again is good and the ITD is maintained as can be seen in Figure 4.21. ITD is calculated using interaural cross correlation of the left and right ear HRIRs (upsampled by a factor of ten) at each position (see Section 2.2.1). Figure 4.22 is the frequency domain equivalent to Figure 4.20 and it shows there is no significant spectral distortion of the HRTFs. If longer lengths (>450 samples) for f are used, significant activity appears in the reconstructed HRIRs before the expected onset point and there is also some distortion of the main peak and reflections for some positions. This distortion is to be expected when one considers that the maximum initial delay before the onset of the HRIR in the set is 55 samples. The largest delay before the maximum peak of the HRIR in the set is longer again at 86 samples.

Figure 4.23 shows the frequency domain representation of the direction dependent components after a 430 sample long f component has been extracted using the regularisation on g^{ϕ} technique. It can be seen when comparing this to the original HRTFs in Figure 4.13, that the pronounced notches in the 7-10kHz range and 15-17kHz range have been maintained in the direction dependent component.

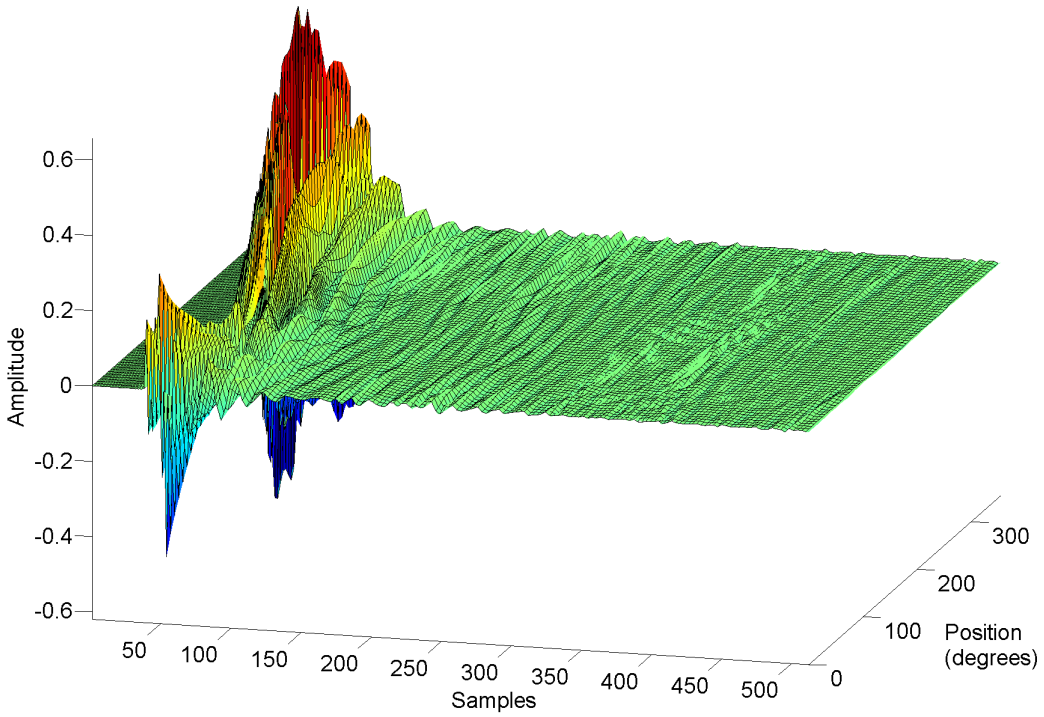


Figure 4.18: Original HRIR dataset.

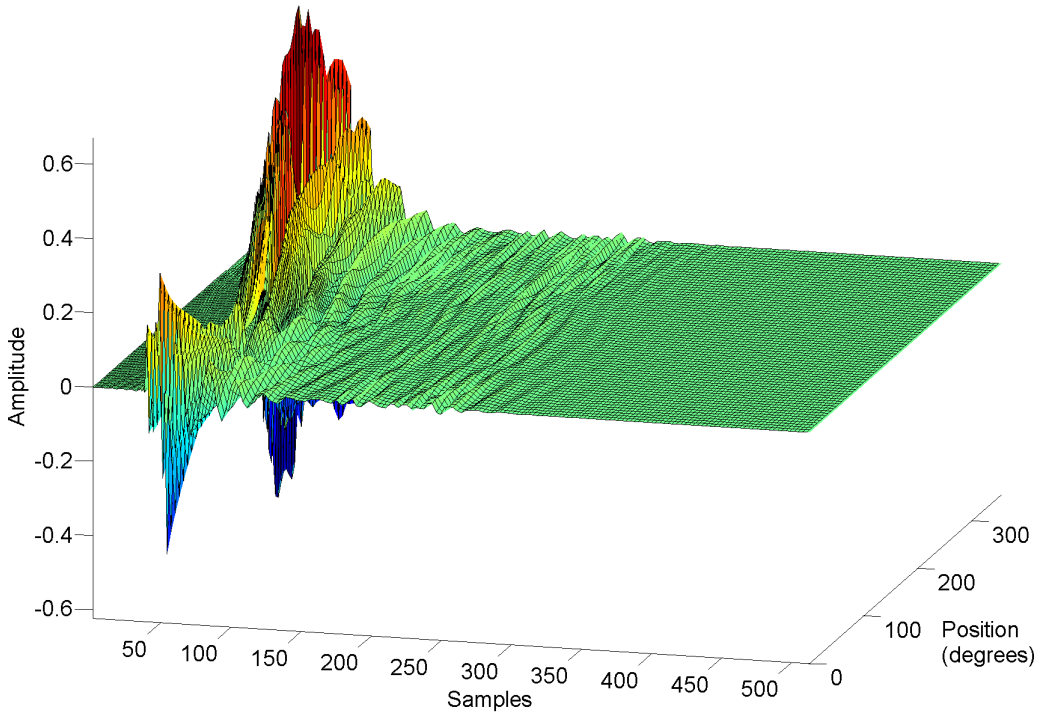


Figure 4.19: Reconstructed HRIRs. Length $f = 256$.

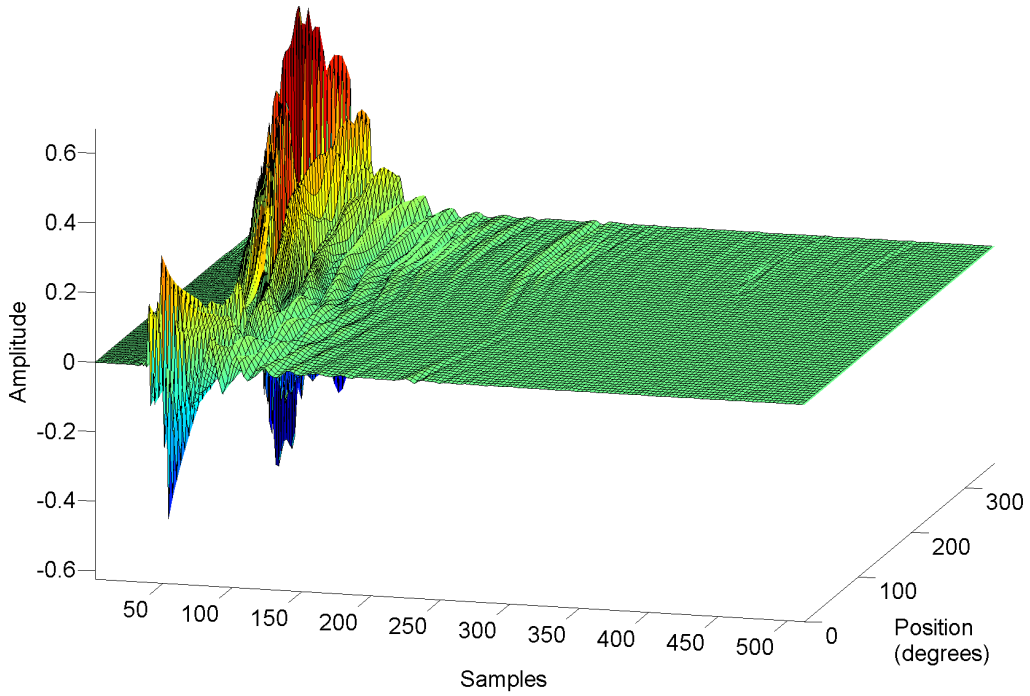


Figure 4.20: Reconstructed HRIRs. Length $f = 430$.

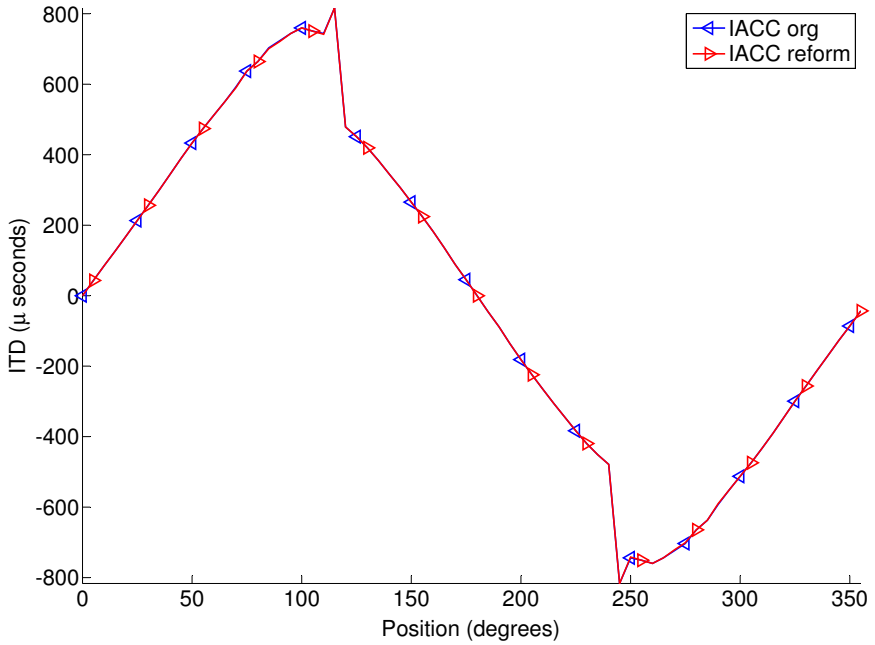


Figure 4.21: ITD for original and reformed HRIRs. Length $f = 430$.

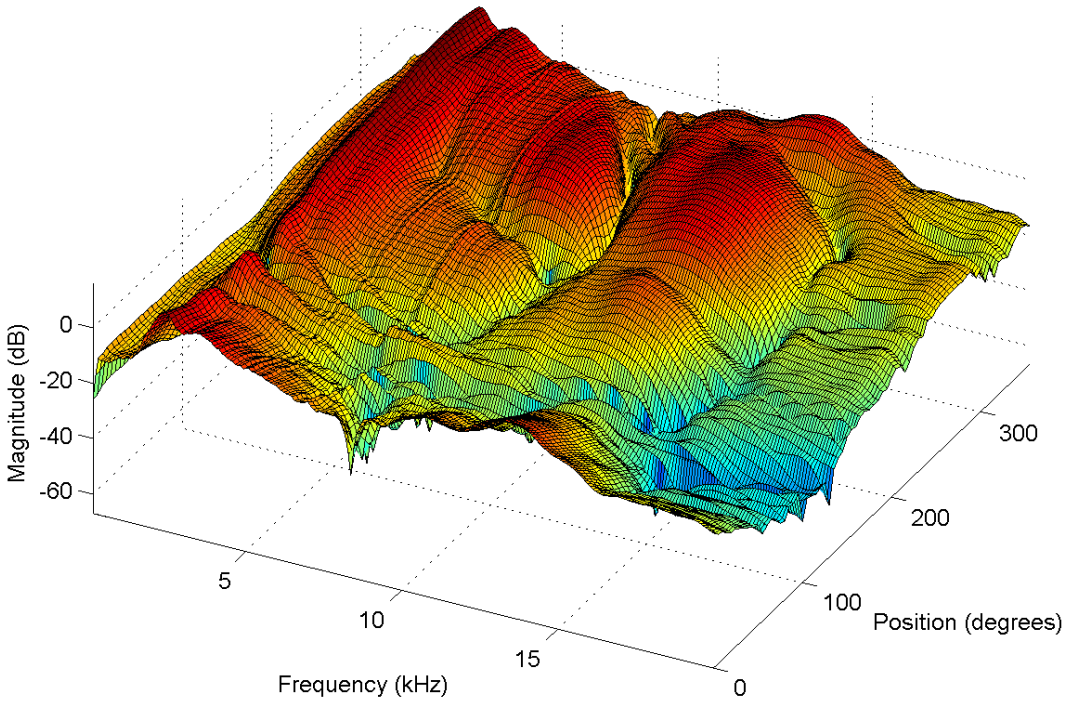


Figure 4.22: Reconstructed HRTFs. Length $f = 430$.

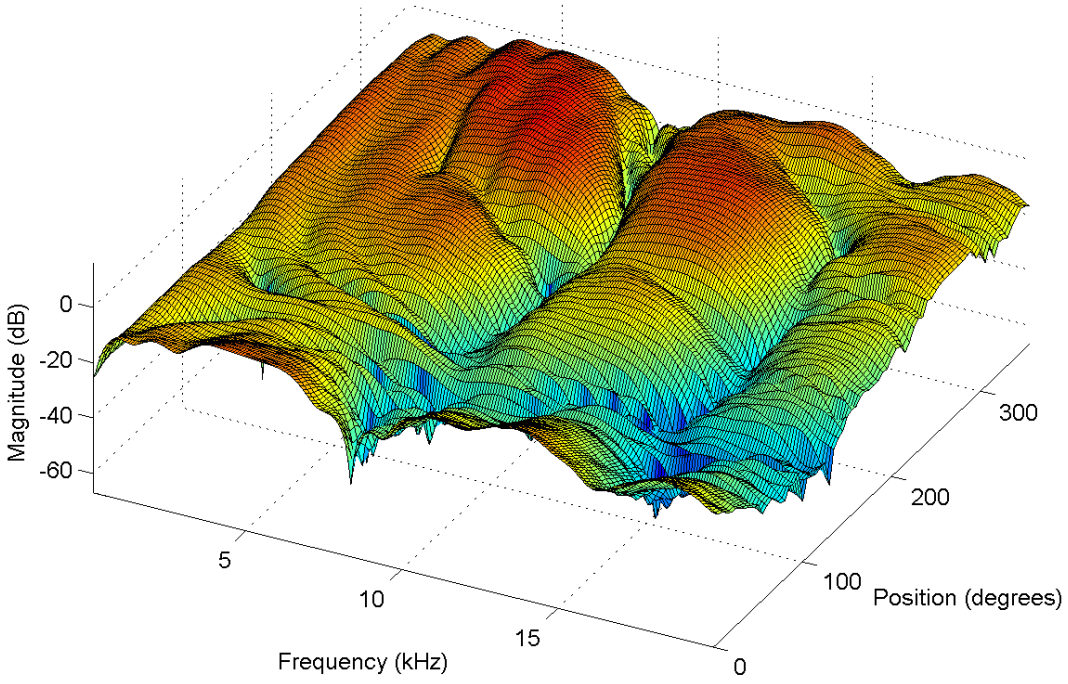


Figure 4.23: g^{ϕ} components in frequency domain. Length $f = 430$.

4.3.2 Investigation of Regularisation using Human Data from CIPIC Database

HRIRs from Subject 3 (a human subject) in the CIPIC database [6] were used to further test the algorithms. Both forms of regularised factorisation was applied to the 0° elevation HRIRs for the left and right ear. Hence the dataset contained 100 HRIRs (50 for each ear). Again 20 iterations of the algorithm were run in each case. Figures 4.24 and 4.25 show the magnitude responses

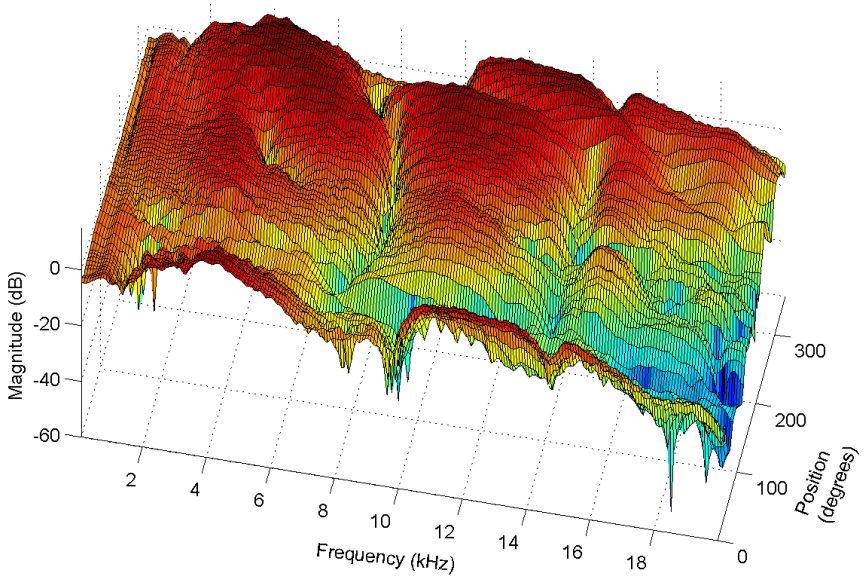


Figure 4.24: Left ear HRTFs.

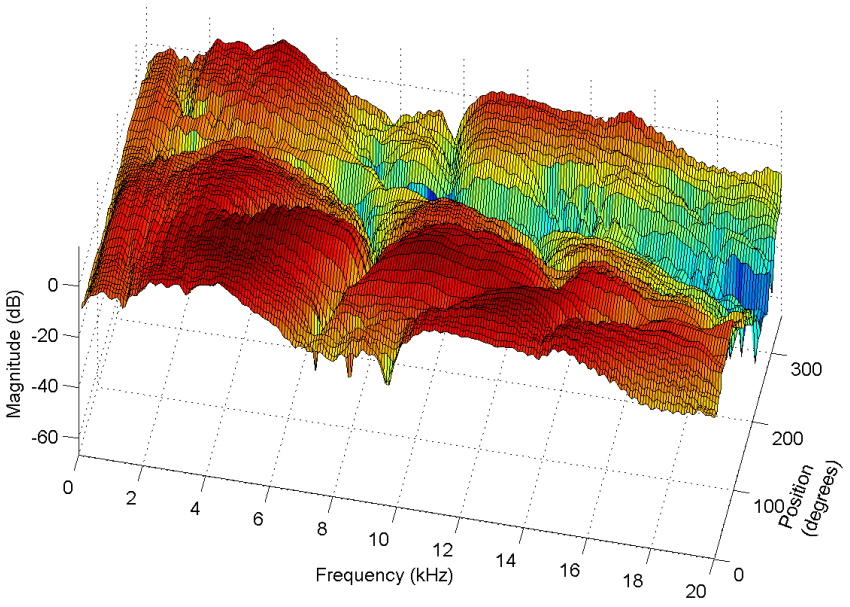


Figure 4.25: Right ear HRTFs.

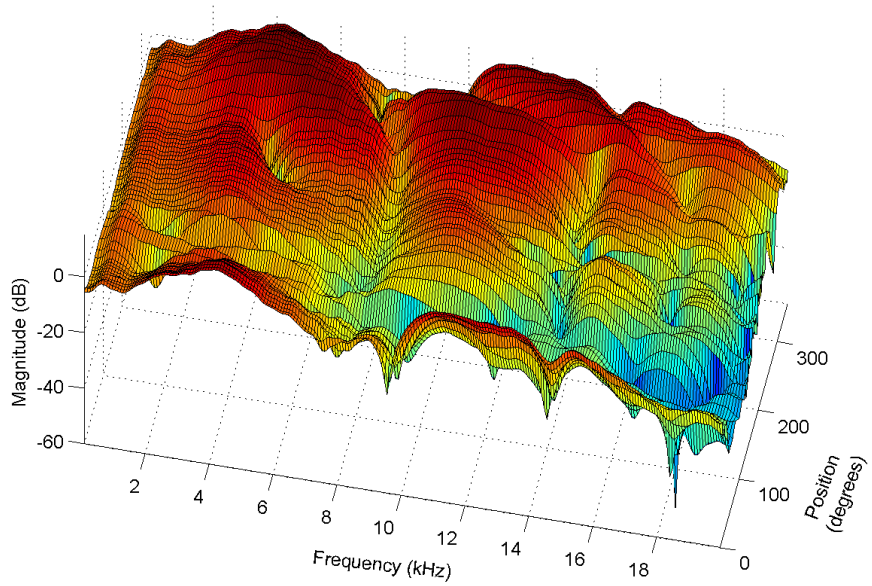


Figure 4.26: Reconstructed left ear HRTFs. f regularisation. Length f =180.

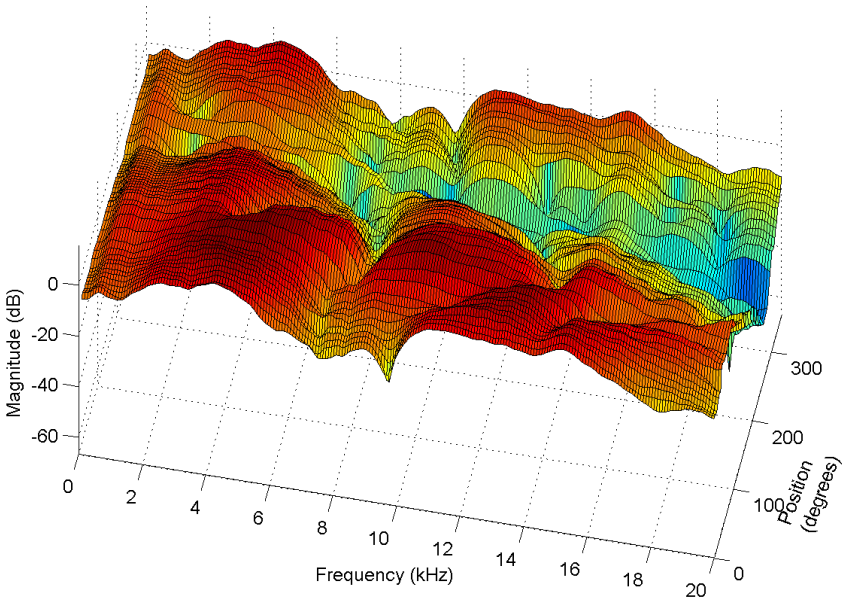


Figure 4.27: Reconstructed right ear HRTFs. f regularisation. Length f =180.

for the original left and right ear HRTFs. Figures 4.26 and 4.27 show the magnitude responses of the reconstructed HRTFs after a 180 sample long direction independent component has been extracted using regularisation on f from the minimum phase equivalent of the dataset. This means that the direction dependent components are only 21 samples in length. The magnitude responses of the reconstructed HRTFs show a close similarity to the original HRTFs.

Figures 4.28 and 4.29 shows the minimum phase response of the original HRTFs while Figures 4.30 and 4.31 show the same for the reconstructed HRTFs. As with the magnitude response there is some smoothing of the spectrum in the case of the reconstructed HRTF but the main features remain. There is a visible discrepancy in Figure 4.31 at high frequencies that is a result of phase wrapping.

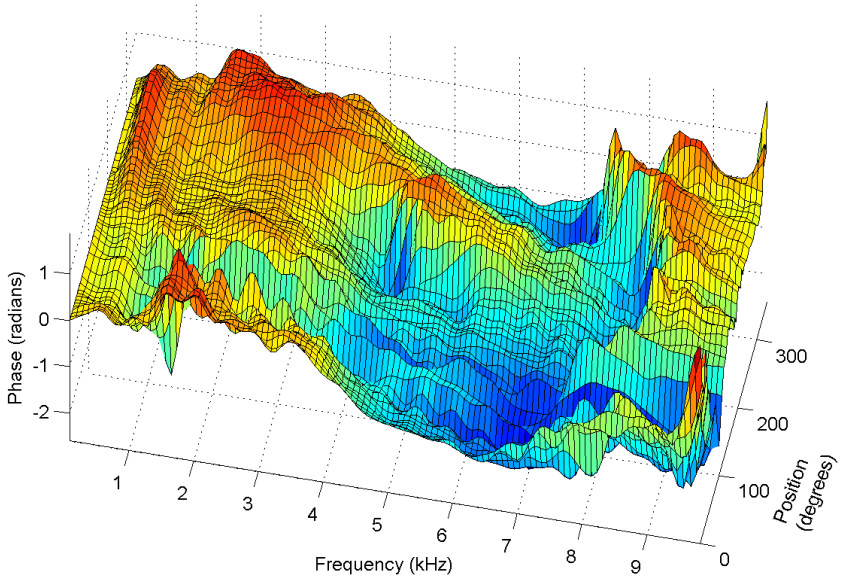


Figure 4.28: Phase. Left ear HRTFs.

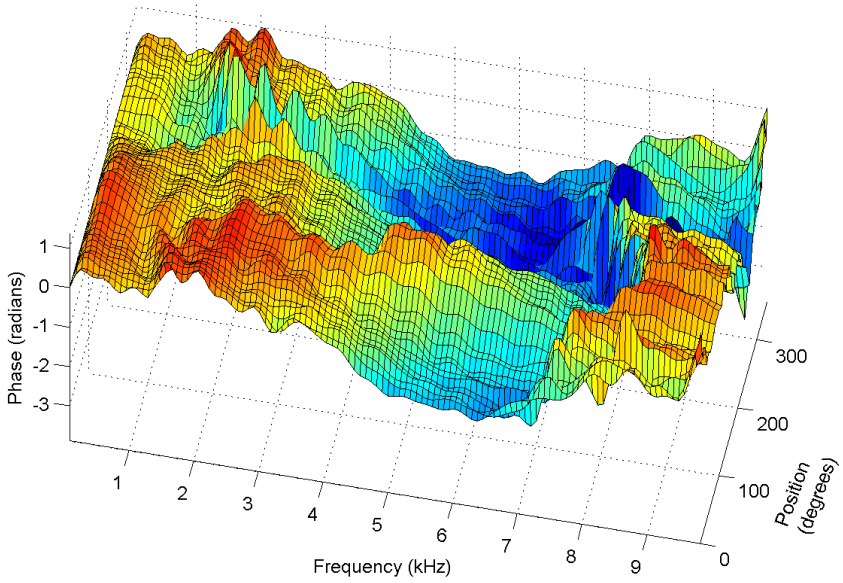


Figure 4.29: Phase. Right ear HRTFs.

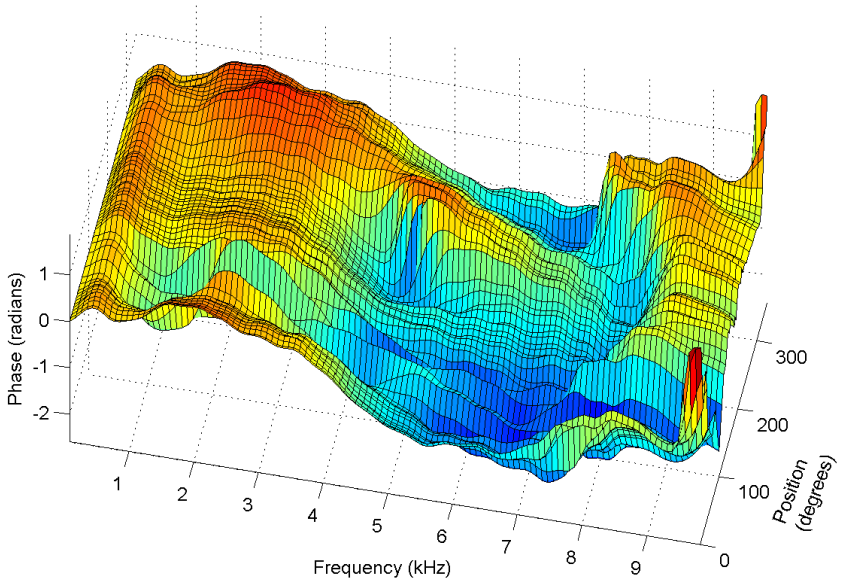


Figure 4.30: Phase. Left ear. f regularisation. Length f =180.

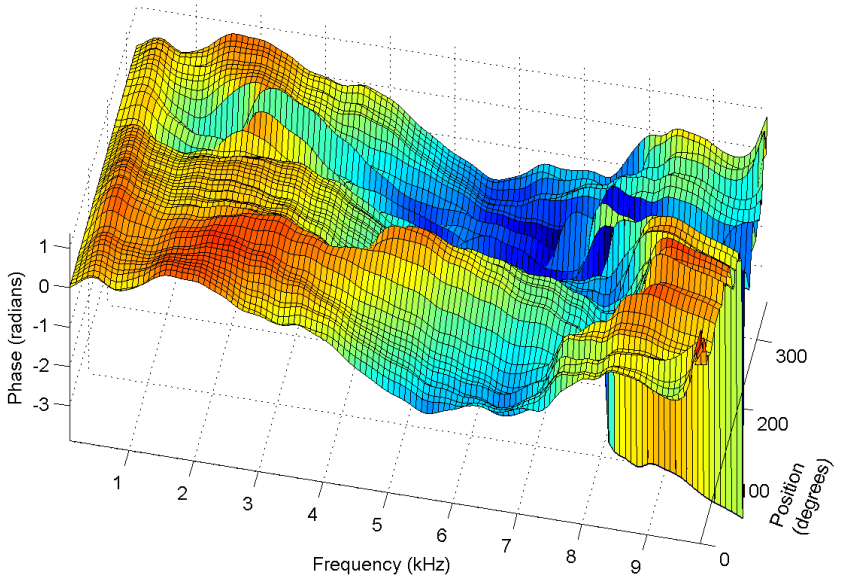


Figure 4.31: Phase. Right ear. f regularisation. Length f =180.

Figures 4.32 and 4.33 show the left and right ear HRIRs in the time domain while Figures 4.34 and 4.35 show the reconstructed HRIRs after a 130 sample long direction independent component has been extracted using regularisation on g^{ϕ}. The agreement between the original and reconstructed responses is excellent. Figure 4.36 shows the ITD calculated by cross correlation of the left and right ear HRIRs for both the original and reconstructed HRIRs. The plot shows that in this case the factorisation of a 130 sample long direction independent component

causes no distortion of the ITD, one of the main cues used for sound localisation.

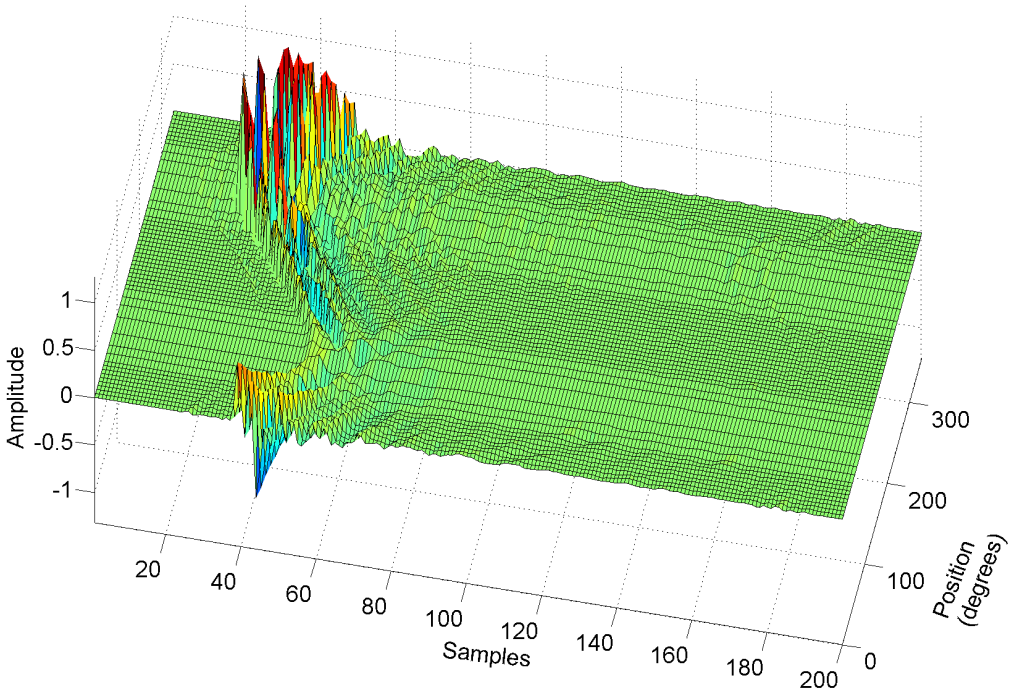


Figure 4.32: Left ear HRIRs.

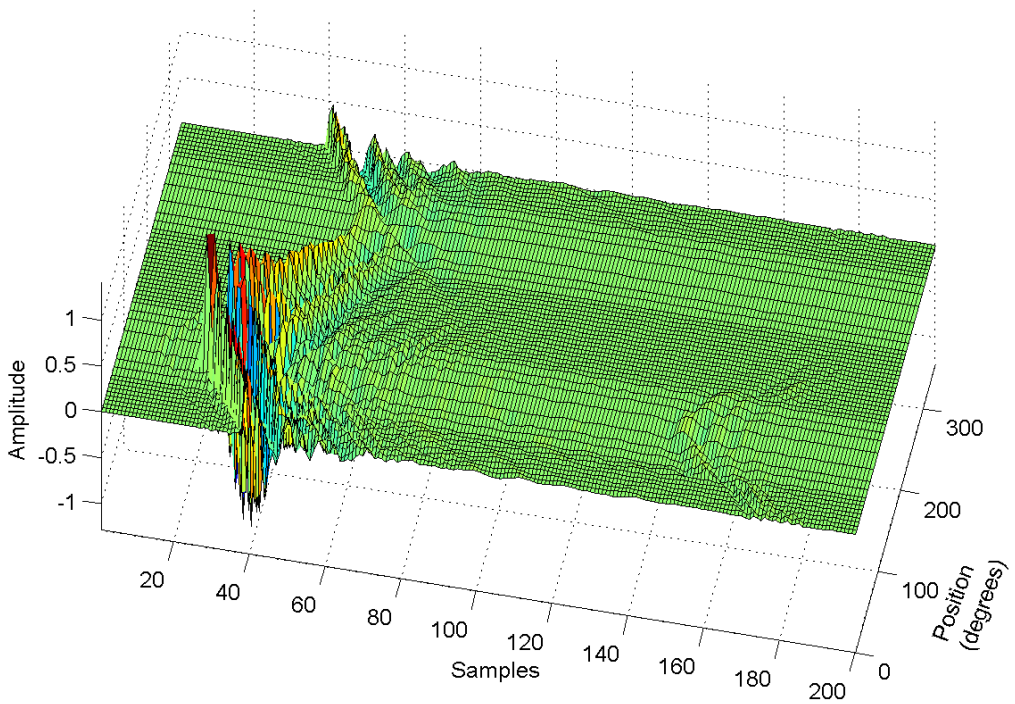


Figure 4.33: Right ear HRIRs.

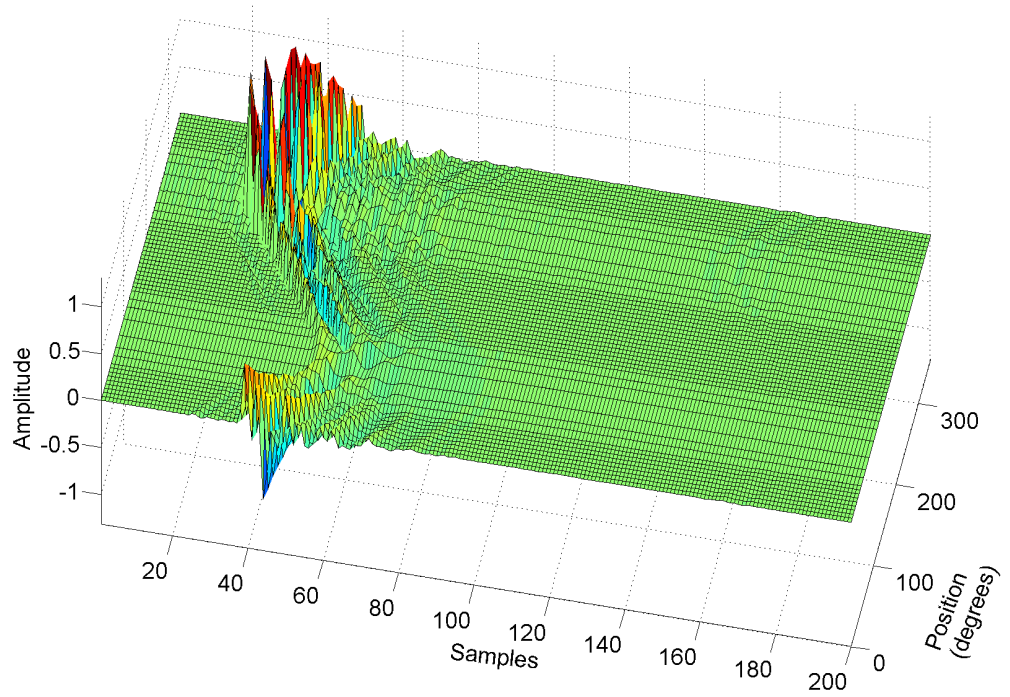


Figure 4.34: Reconstructed left ear HRIRs. g^{ϕ} regularisation. Length f =130.

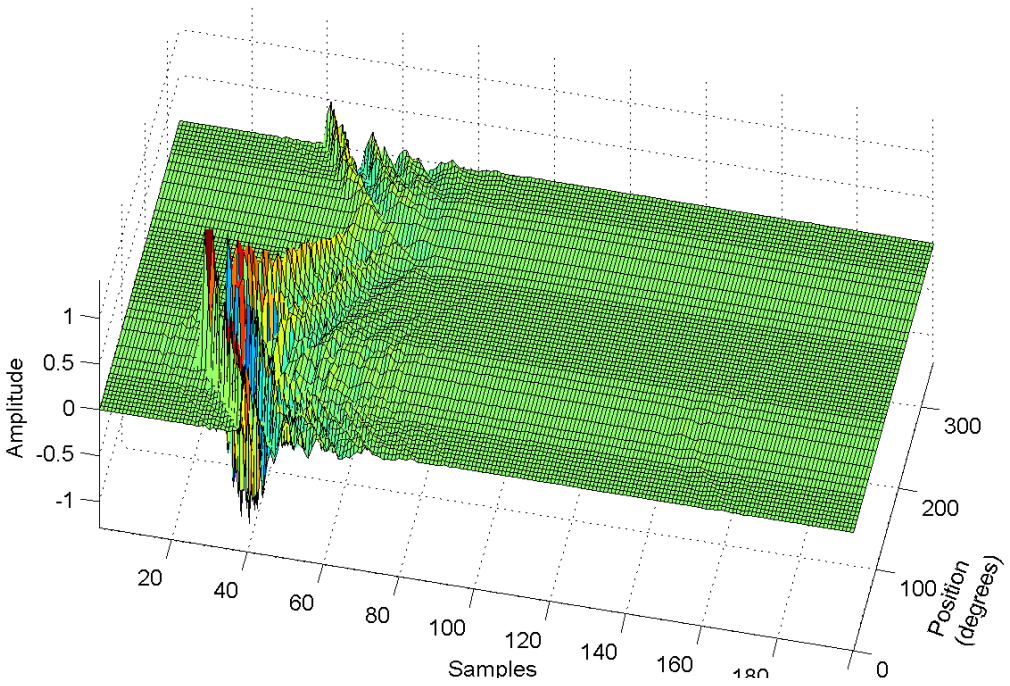


Figure 4.35: Reconstructed right ear HRIRs. g^{ϕ} regularisation. Length f =130.

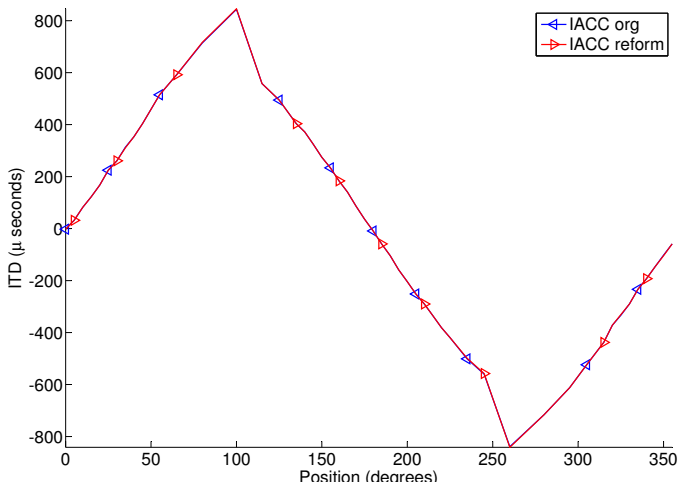


Figure 4.36: ITD. g^{ϕ} regularisation. Length f =130.

4.4 Conclusion

The factorisation technique detailed in this chapter allows for a direction independent component to be extracted from HRIR sets, leaving shorter direction dependent components. This offers significant advantages in real time dynamic virtual auditory environments as it enables quicker real time convolution, faster filter switching and lower memory usage. Application of the technique to HRIR data from the CIPIC dataset shows the effectiveness of the base algorithm. Two regularised derivatives of this base algorithm are also introduced which allow for more robust, initial condition independent factorisation. Regularisation on the direction independent component when factorising minimum phase HRIR datasets allows for long direction independent components to be extracted and for very low reconstruction error. However this technique necessitates obtaining accurate ITDs for reintroduction. Regularisation on the direction dependent component allows for the ITD to be maintained in the reconstructed HRIRs. However the length of the direction independent component that can be extracted is more limited than in the minimum phase case. These two techniques are applied to HRIR data from both the CIPIC dataset and Gardner and Martin's KEMAR dataset. Both show good reconstruction of the full length HRIRs after factorisation.

5

Room Response Modelling

To create convincing virtual auditory environments it is important to include the effect of the room on the sound that is received at the listener's ears. This can be done by convolving the source audio with the Room Impulse Response (RIR) for that particular source and receiver position in that particular environment. If the source or the listener moves then a different RIR is required. In a fully interactive 'walk through' situation with full freedom accorded to the listener, large grids of such RIRs are required. If source movement is also to be considered, the number of measurements is increased by a scalar multiple depending on the allowable range of source movement. The concept of the RIR has been introduced in Section 2.5, along with techniques for its measurement and synthesis.

In this chapter a novel technique for the spatial interpolation of RIR is devised. This technique involves the partitioning of the RIR into two components: an early reflection component and a diffuse tail component. The early reflection component is relatively sparse in nature containing peaks when reflections are incident. The diffuse decay component is dense and stochastic in nature. Figures 2.10 and 2.11 in Section 2 illustrate the formation and structure of a RIR and Figure 2.11 is reproduced below for clarity. The first challenge is to determine the transition point between the early reflection and diffuse decay components. Spatial interpolation of the early reflection components is then achieved using a novel dynamic time warping based approach while the diffuse component is synthesised through the decomposition of nearby tails into critical bands and randomly shifting them in time. The application of this technique to Wave Field Synthesis reproduction will be demonstrated and both objective and perceptual tests performed to compare the use of measured versus interpolated responses.

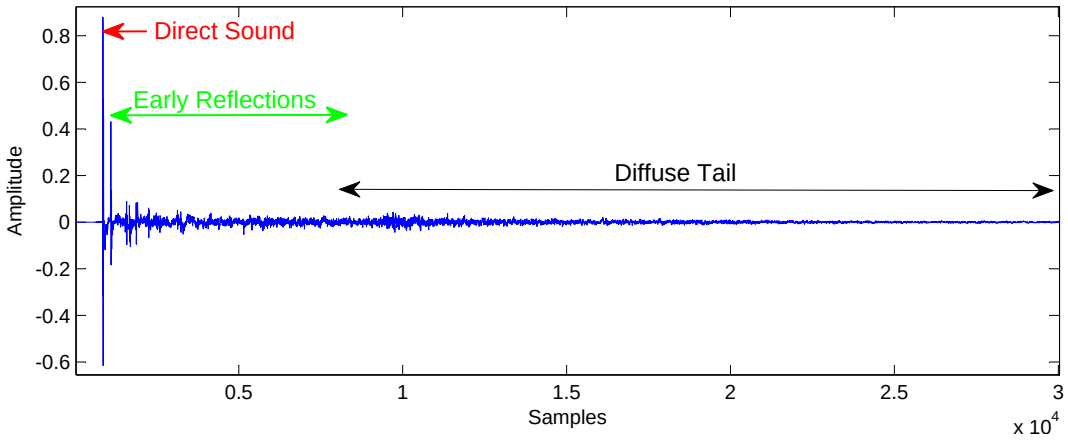


Figure 5.1: A sample RIR.

5.1 Existing RIR Interpolation Techniques

Interpolation of real measured room impulse responses is not widely used. Instead many prefer to synthesise RIR using computational models (see Section 2.5 for more detail). However synthesis techniques like these have their own drawbacks including the need for accurate geometrical models of the environment and precise knowledge of the absorption and diffusivity properties of its boundaries [22]. As such measurement based auralisation remains of central importance to field.

Haneda et al. [73] propose the use of a common pole and residue model to interpolate RIRs for rooms with simple geometry. The poles correspond to the resonance frequencies of the room which are independent of source and receiver position while the residues correspond to the eigenfunctions of the room which are position dependent. Thus interpolation of RIRs simplifies to interpolation of these eigenfunctions. The authors use either a simple cosine approximation or the linear prediction method to do this. This produces good results when compared with linear interpolation. Huzsty et al. [81] introduce a fuzzy modelling method for RIR interpolation. This technique does not require information regarding the room geometry or boundary materials. Their approach relies on accurate detection of significant reflections. They implement this manually and concentrate on the other parts of the technique in this paper. While this paper is not entirely complete in its approach it acknowledges some basic tenets that apply to RIR interpolation, namely that only the deterministic early reflections can be interpolated and some kind of temporal mapping is necessary to do this.

5.2 Transition Point Determination

Determining a suitable boundary point between early and late reflections is difficult as the transition is as much a perceptual event as it is a mathematically defined change. In the past

some authors have assigned a value to this transition point in a rather arbitrary way. 80ms is a commonly used value [11, 151] but this value varies between authors. Jot [87] places this transition point between 60 and 100ms while Hidaka et al. [77] describes the transition time as existing in the range from 50 to 200ms. Clearly a more systematic and definitive approach is needed in determining this transition point. Lehmann and Johansson [109] define the transition point as the time for which the overall acoustic energy in the RIR has decreased by a certain amount. Hidaka et al. [78] use short term correlation analysis between the direct and initial sound and the subsequent sound as a measure while Stewart and Sandler [168] explore the use of standard deviation as well as kurtosis values. These techniques do not require knowledge of the room geometry. Kuttruff [104] proposes the analysis of the temporal density of reflections to determine how diffuse the sound field is at a given point in time. The mean free path, which is determined by the volume and surface area of the room, is another technique used. Reflection order can be used as an indicator [136].

The approach to finding the transition time used in this thesis is based on the mean free path. Naylor and Rindel's method [136] suggests using fourth order reflections as the cut off point of the early reflection region. They define the time to reach a given reflection order, T_{ro}, as

$$T_{ro} = \frac{4V}{cS}(O_e + 1) \tag{5.1}$$

where V is the room volume, S is the surface area of the room, c is the speed of sound and O_e is the reflection order. The transition point, T_t is taken as T_{ro} with O_e equal to 4.

5.3 Dynamic Time Warping

Dynamic time warping (DTW) is an efficient algorithm for determining the similarity of two sequences which may be shifted or distorted in time. It allows for sequences to be optimally aligned by stretching their time axes relative to each other. First developed in the late 1950's [14], the technique became popular in the area of speech processing in the following decades [159].

Let us first examine the DTW algorithm in its simplest form. Consider the two sequences a and b shown in Figure 5.2. These sequences are similar in that they both contain two main features. However these features occur at different time instances and are slightly different in shape in each sequence. DTW can be used to align these main feature points. The first step is to create a distance matrix, D, which contains the squared difference between every element in a and every element in b.

$$D(i, j) = \|a(i) - b(j)\| \tag{5.2}$$

In order to find the best alignment path an accumulative distance matrix (or cost matrix), C,

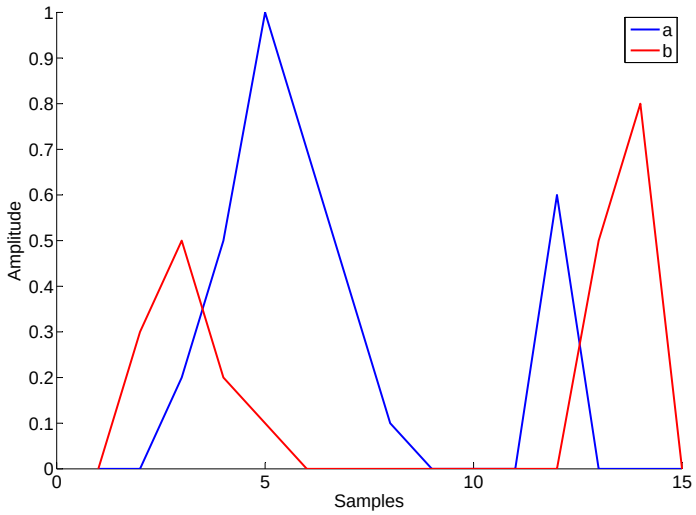


Figure 5.2: Two sequences.

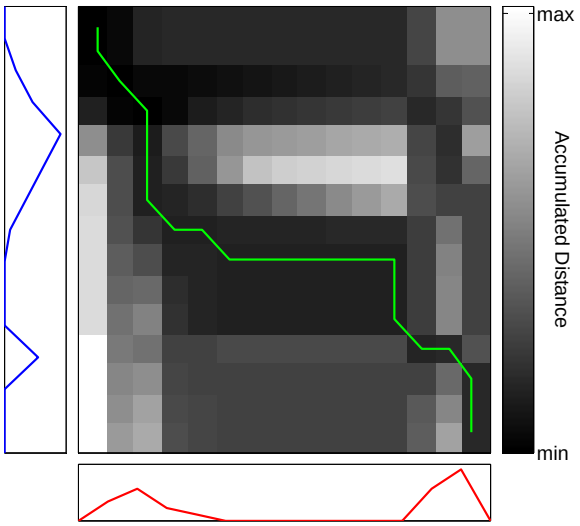


Figure 5.3: Accumulated distance matrix.

is formed from the distance matrix (see Figure 5.3) as follows:

$$C(i,1) = \sum_{k=1}^{i} D(k,1) \tag{5.3}$$

$$C(1,j) = \sum_{k=1}^{j} D(1,k) \tag{5.4}$$

$$C(i,j) = D(i,j) + min \begin{cases} C(i-1,j-1), \\ C(i-1,j), \\ C(i,j-1) \end{cases} \tag{5.5}$$

The optimal warping path, w, can be found by backtracking through the accumulated distance matrix, at each step choosing the minimum accumulated distance value and obeying the following conditions:

- Boundary condition: The start and end points of the warping path must be the start and end points of the two sequences
- Monotonic condition: The sample indexes of the warping path must incrementally increase or remain the same (i.e. the path cannot turn back on itself). This preserves the time ordering of features and prevents repetition of features.
- Continuity Condition: The warping path cannot have more than a one sample increase in index. This prevents features being excluded.

The green line in Figure 5.3 shows the optimal warping path calculated using the above method for our example. The warping path is then applied to the original sequences in order to align their features (see Figure 5.4).

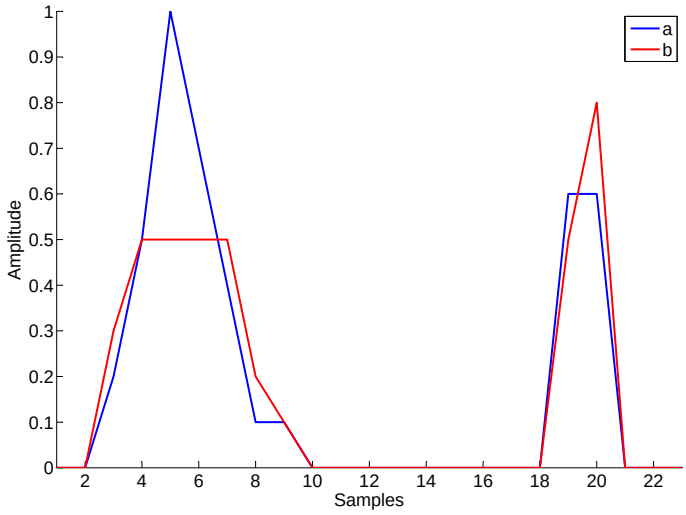


Figure 5.4: Warped sequences.

DTW has been described above in its simplest form. There are of course alterations and optimisations that can be made to the algorithm.

5.4 Application of DTW to RIR Interpolation

The mechanism of DTW has been explained in the previous section. In this section a novel technique will be described that allows for DTW to be used to spatially interpolate the early reflection component of measured room impulse responses. Consider the layout in Figure 5.5 with a loudspeaker at a given source position and two microphone receivers some distance apart.

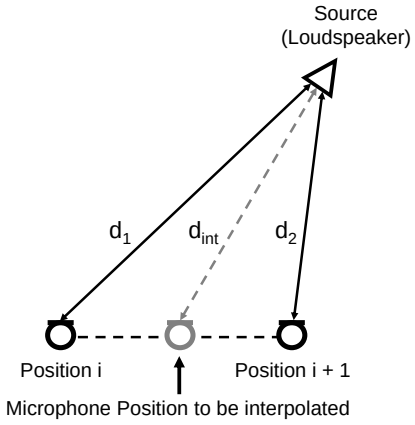


Figure 5.5: Interpolation between two microphones.

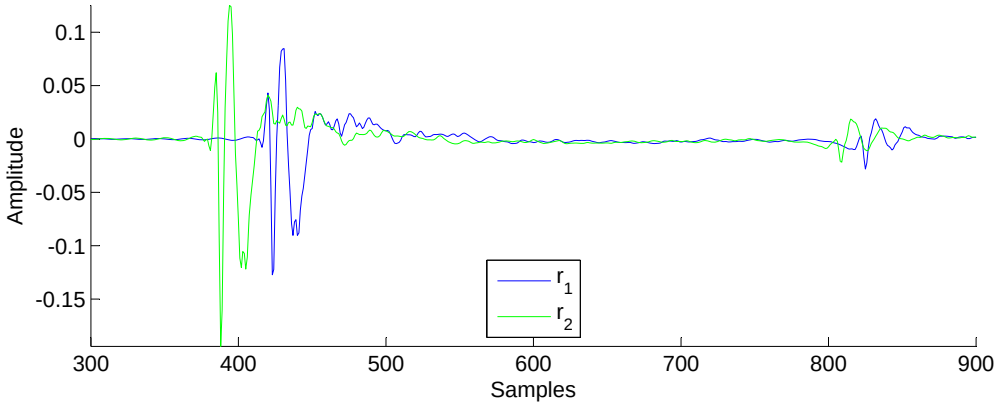


Figure 5.6: Original RIRs.

The objective is to interpolate to obtain a new RIR (direct sound and early reflections only) at some position on the line connecting the two existing microphone positions. Let r_1 and r_2 denote these known RIRs (see Figure 5.6) and r_{int} denote the interpolated RIR. In order to obtain r_{int} it is first necessary to apply DTW to r_1 and r_2 to align their main feature points. This produces two time warped vectors $r_1 w$ and $r_2 w$ and two warp vectors w_1 and w_2 which describe how r_1 and r_2 map onto their warped equivalents $r_1 w$ and $r_2 w$. As the feature points have been aligned it is now possible to safely linearly interpolated between $r_1 w$ and $r_2 w$ to obtain $r_{int} w$ without any blurring of the feature points using simple linear interpolation. The linear interpolation is weighted based a ratio of the inverse distances between the source and the microphones, as

sound pressure level is proportional to inverse distance.

$$\alpha = \frac{\frac{1}{d_2} - \frac{1}{d_{int}}}{\frac{1}{d_2} - \frac{1}{d_1}} \tag{5.6}$$

and hence,

$$r_{int}w = \alpha r_1 w + (1 - \alpha) r_2 w \tag{5.7}$$

Figure 5.7 demonstrates this with Figure 5.8 providing a more in depth view of the direct sound component to clearly demonstrate the effect of the warping.

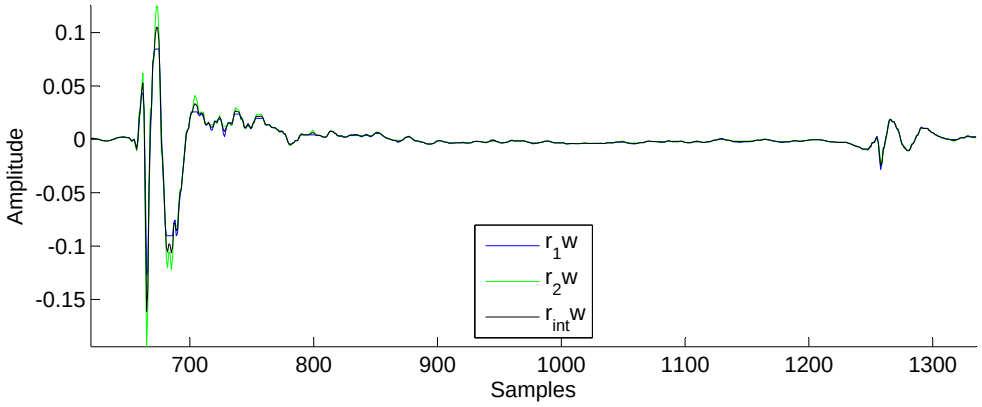


Figure 5.7: Two Warped RIRs and Interpolated RIR.

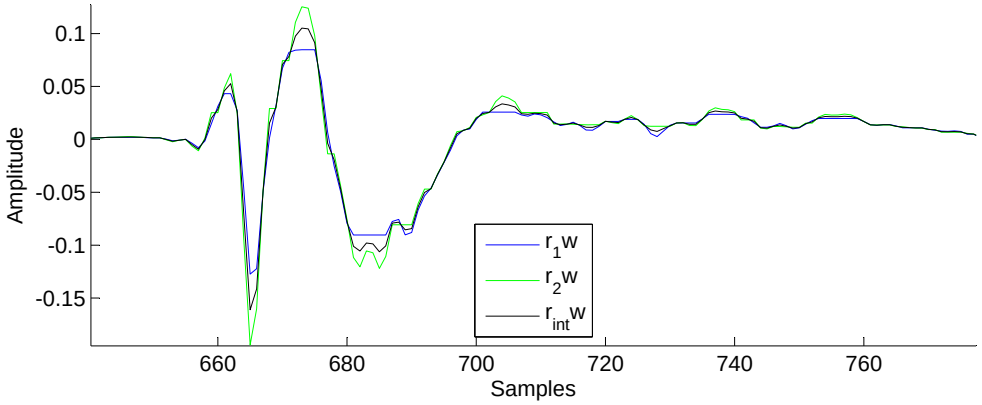


Figure 5.8: Magnified version of a key section of Figure 5.7.

The next step in the process is to interpolate the warp vectors w_1 and w_2 in order to obtain the correct warp vector w_{int} which will allow for $r_{int}w$ to be successfully mapped back to the unwarped time domain. Again simple linear interpolation is used to achieve this with the weights calculated based on a ratio of the distance from the source to the receiver locations. The main objective of the weighted interpolation is that the application of the newly interpolated warp vector will cause the direct sound impulse to be correctly positioned in time in the interpolated

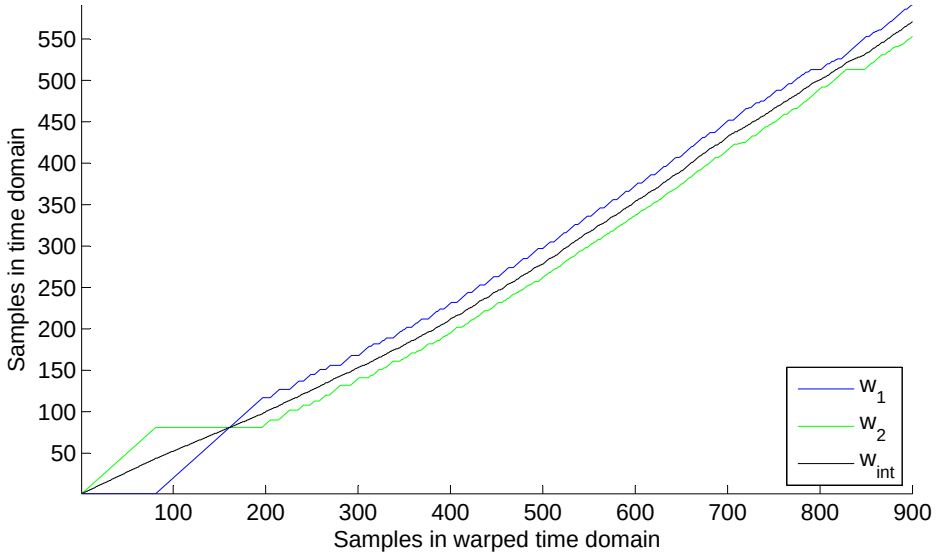


Figure 5.9: Interpolating warp vectors.

RIR. As the time for a sound wave to travel from its source to the receiver position is directly proportional to the distance travelled the weighting of the interpolation is based on a distance ratio described as follows

$$\beta = \frac{d_{int} - d_2}{d_1 - d_2} \tag{5.8}$$

and hence,

$$w_{int} = \beta w_1 + (1 - \beta) w_2 \tag{5.9}$$

Each value of w_{int} must be rounded to the nearest whole number as it is in essence a sample mapping vector. Figure 5.9 demonstrates the result of this interpolation. The interpolated warp vector is then applied to the interpolated warped RIR to reverse the warping process. This 'dewarping' process is described as follows

$r_{int}(1) = r_{int}w(1)$
$i = 2$
for $x = 2$ to length(w_{int}) **do**
  if $w_{int}(x) \neq w_{int}(x-1)$ **then**
    $i = i + 1$
    $r_{int}(i) = r_{int}w(x)$
  end if
end for

The result of the interpolation can be seen in Figure 5.10. It displays both the interpolated RIR and the actual measured RIR measured at that position along with the two measured RIRs, one on each side of the interpolated position, that were used for the interpolation operation. There is nearly complete agreement between the interpolated and measured RIRs for the given

position.

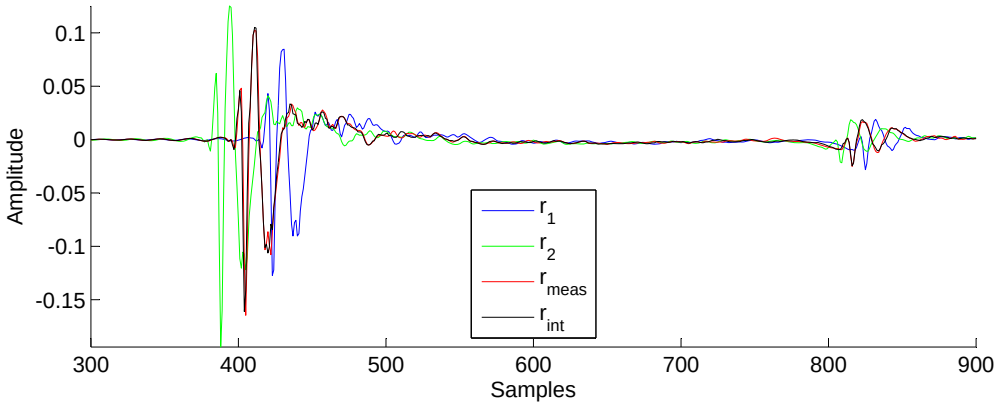


Figure 5.10: Unwarped time domain result.

The advantage of this technique over simple linear interpolation is now shown. A simple linear interpolation technique involves an initial shifting of one of the RIRs to align the direct sound features of both RIRs. Linear interpolation is then carried out on the responses based on the same weightings used in the warped case (see Equation 5.6). A scaled time shift ($-\beta$ * original time shift) is then applied to the interpolated RIR. This produces the result shown in Figure 5.11. Linear interpolation correctly interpolates the direct sound component but significantly distorts the first reflection. This distortion is more apparent in Figure 5.12 where the section containing the first reflection has been magnified. The DTW based approach avoids this pitfall.

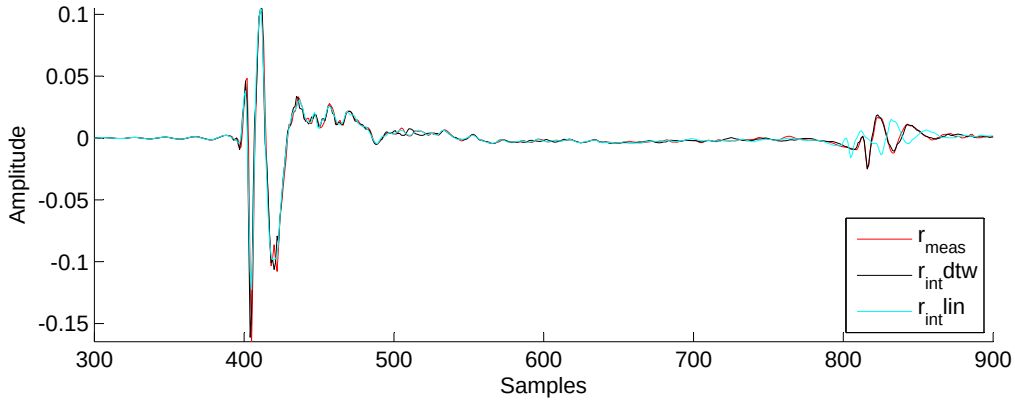


Figure 5.11: Compare linear interpolation to DTW method.

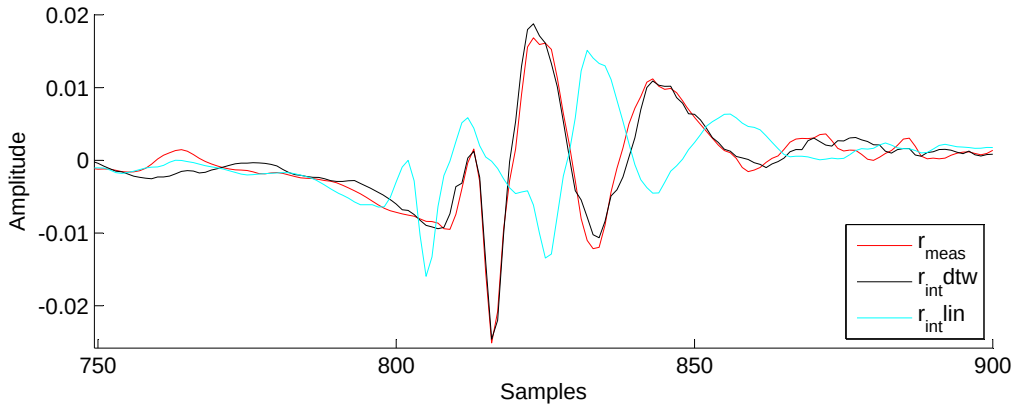


Figure 5.12: Magnified version of a key section of Figure 5.11.

5.5 Tail Synthesis

RIR modelling techniques, such as the image source method and other techniques introduced in Section 2.5.1, frequently model only the early reflection component due to the exponential nature of the growth in computational complexity when reflection order increases. As such there is a significant body of research on the synthesis of the natural sounding diffuse tails for RIR synthesis. Moorer [129] proposes the use of white noise with an exponentially decaying envelope as a reverberation tail and he describes the resultant sound when the RIR is convolved with source audio as 'natural sounding'. Jot [87] develops a real time approach for generating artificial reverberation using a feedback delay filter network with absorbent filters included. These absorbent filters are designed based on characteristics of a response measured in a real room. Lehmann and Johansson [108, 109] develop a method for modelling the diffuse tail as a decaying random process. This technique is used in conjunction with the image source method for early reflection synthesis. The authors develop a closed form expression for the energy decay based on a geometrical analysis of the image source principle which dictates the envelope to be applied to the random noise.

In the course of the current research study a significant number of RIRs were measured. The approach adopted here has been to synthesise a decorrelated tail from the existing measured tails in the dataset that retains their timbre. First the amount of decorrelation between all of the measured RIRs must be calculated. The cross correlation between each tail and every other tail in the dataset is calculated using the normalised cross correlation function [100] as follows:

$$R_{h_i^l h_j^l}(\tau) = \frac{\sum h_i^l[n] h_i^l[n+\tau]}{\sqrt{\sum h_i^l[n]^2 \sum h_j^l[n]^2}} \tag{5.10}$$

This allows for the correlation floor to be calculated by taking the average of the correlation values, excluding the autocorrelation of a tail with itself. The correlation of any synthesised tail with the tail it is based on will be compared to this value to ensure it is not overly decorrelated

form the original.

A 'best' tail is chosen from the available measured dataset. The criteria for this selection is based on the Euclidean distance between each tail and every other tail in the dataset. This 'best' tail is then separated into perceptual critical bands. A bank of 4096 tap equivalent rectangular bandwidth (ERB) FIR filters with a combined flat frequency response and piecewise linear phase are used to achieve this. The phase in each band of the ERB filterbank is then randomly delayed and the amount of forward or backward shift is dependent on the longest waveform period in the band in question, as well as the precedence effect as described in the work of Bouéri and Kyriakakis [23]. If the resultant synthesised signal is deemed too decorrelated from the original RIR (i.e. it is below the correlation floor) after comparison, then a small portion of the original tail can be added back in to compensate for this. The tail is then added to the interpolated early reflection component with crossfading at the transition. Figure 5.13 shows an example of a synthesised tail as well as the measured tail on which it is based.

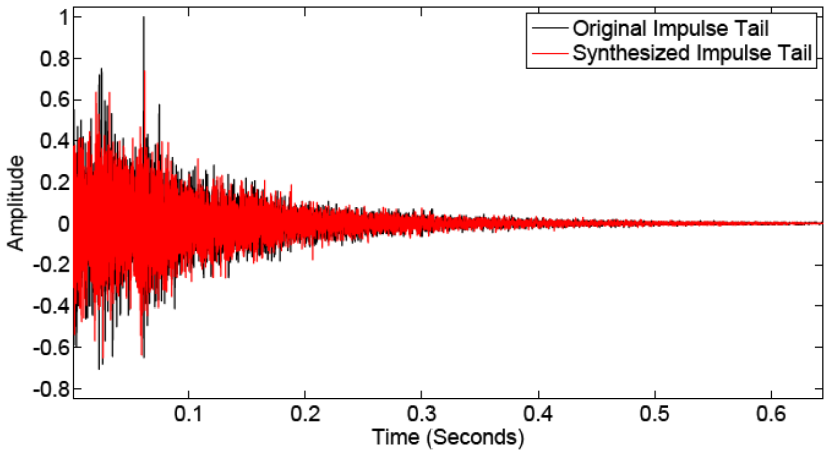


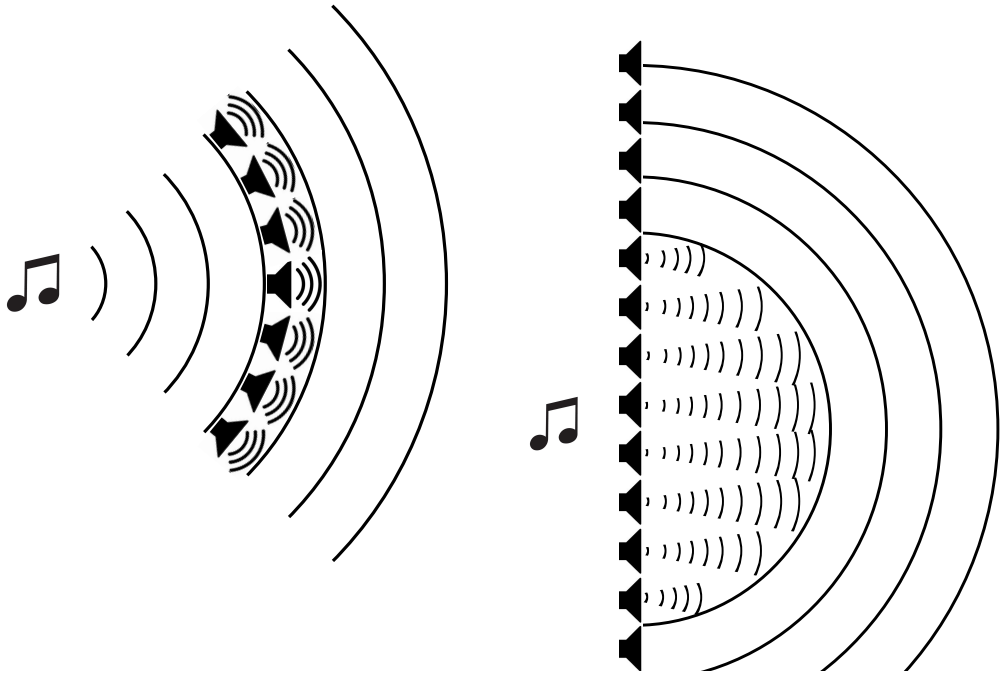
Figure 5.13: Comparison of original and synthesised tails.

5.6 Application of Interpolation to Wave Field Synthesis Reproduction

5.6.1 Introduction to Wave Field Synthesis

Wave Field Synthesis (WFS) is a sound reproduction technique which aims to reconstruct an acoustically correct sound field over an extended listening area. WFS offers the obvious benefit that it avoids the 'sweet spot' problem of traditional reproduction techniques and is therefore very suitable for distributed audiences. However, a convincing implementation requires large numbers of loudspeakers and significant processing power to drive them, which may be unfeasible from a practicality and monetary viewpoint for many applications. It was introduced by

Berkhout [17] in the late 80s and is based on Huygens Principle (see Figure 5.14a) which states that any wave front can be regarded as a superposition of elementary spherical waves. While Huygen's work was focussed on light waves it is also applicable to sound The mathematical



(a) Illustration of Huygen's Principle. (b) Wave Field Synthesis.

Figure 5.14

description of Huygens Principle is given by the Kirchhoff-Helmholtz integral. It states that the sound pressure is completely determined within a volume free of sources, if sound pressure and velocity are determined in all points on its surface, S.

$$P(\mathbf{r},\omega) = \frac{1}{4\pi}\iint_S P(\mathbf{r}_S,\omega)\frac{\partial}{\partial n}\left(\frac{e^{-jk|\mathbf{r}-\mathbf{r}_S|}}{|\mathbf{r}-\mathbf{r}_S|}\right) - \frac{\partial P(\mathbf{r}_S,\omega)}{\partial n}\frac{e^{-jk|\mathbf{r}-\mathbf{r}_S|}}{|\mathbf{r}-\mathbf{r}_S|}dS \tag{5.11}$$

where $P(\mathbf{r},\omega)$ is the sound pressure in the Fourier domain at an arbitrary point in the volume, k is the wave number, S is the surface of the volume, $\mathbf{r}$ is the coordinate vector of an observation point and $\mathbf{r}_S$ is the coordinate vector of the integrand functions on S. The first term in this expression represents the distribution of dipoles while the second represents a distribution of monopoles. Simplification can be achieved by considering a place surface instead of a closed surface. In this case the Kirchhoff-Helmholtz integral is replaced by one of the Raleigh integrals which describe how the wave field is generated by either secondary monopoles or dipoles.

In reality the reproduction is by a discrete line array of loudspeakers as opposed to the continuous, infinite array assumed by theory. A consequence of finite spatial reproduction by any real loudspeaker array is that there is a limited frequency range over which accurate wave

field reconstruction is possible, and above this range spatial aliasing effects occur [17]. The upper limit of this frequency range, f_a, is given by

$$f_a = \frac{c}{\Delta x \| sin(\theta^s) - sin(\theta^v) \|} \tag{5.12}$$

where Δx is the loudspeaker spacing, θ^v is the maximum source angle on the sampling side and θ^s is the maximum angle on the reproduction side as shown in Figure 5.15. c is the speed of sound. Above this frequency the reproduced wavefield suffers audible colouration and spectral

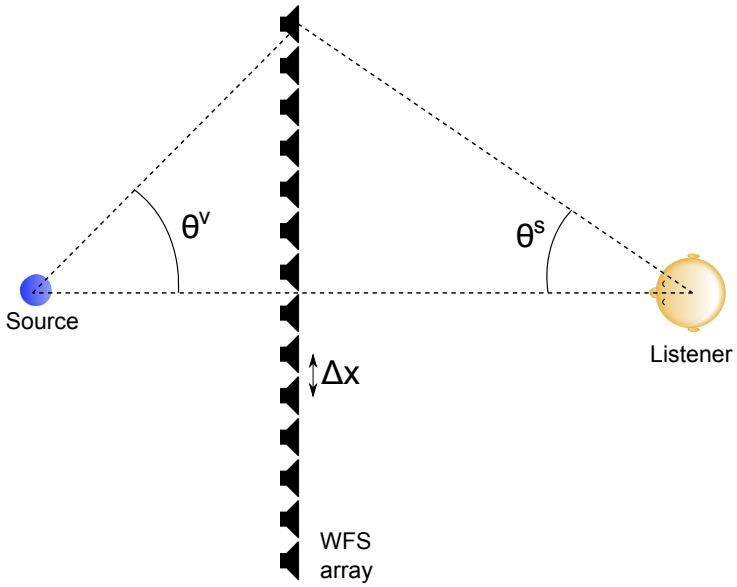


Figure 5.15: Illustration of elements of spatial aliasing formula (Equation 5.12).

and spatial distortion. In order for correct ITD cues to be produced it is desirable that this frequency limit be as high as possible. A 12.5cm spacing would allow accurate reproduction up to approximately 1.45kHz, which allows for maintenance of the ITD. Larger spacing would compromise the integrity of the ITD.

Several techniques have been devised in an attempt to overcome this aliasing problem [38, 166, 186]. In the current study of interpolation a technique called OPSI (Optimized Phantom Source Imaging in Wavefield Synthesis) is employed above the spatial aliasing frequency, as suggested by Wittek [186]. Stereophonic imaging is used to increase localisation focus and reduce colouration. Hence the wavefield is reproduced only below the spatial aliasing frequency.

5.6.2 RIR Capture

The dataset of 32 RIRs was captured in the Printing House Hall in Trinity College Dublin. The room dimensions are approximately 5.85m x 5.85m x 15m. RIR measurements were made at 0.12m spacings in a line across the breadth of the hall. RIRs from 3 different source positions, at -30°, 0° and 15° relative to the listening position shown in Figure 5.16, were captured. The

radial distance of the sources to the listening position was 2.5m. The impulse responses were

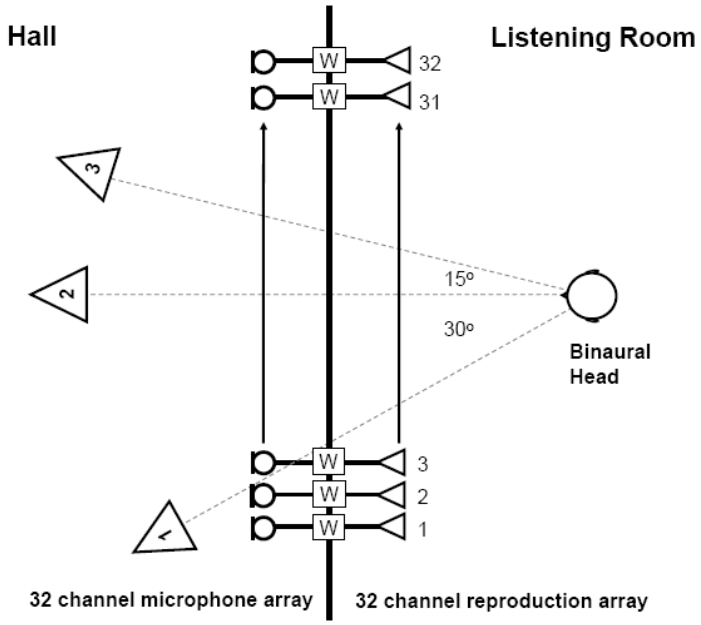


Figure 5.16: Setup for RIR capture and WFS reproduction.

obtained using Farina's logarithmic swept sine technique [50]. A Genelec 1029A loudspeaker was used as the source, and a Soundfield MK5 system was used to record the dataset. Forward facing pressure-gradient RIR responses were extracted from the B-Format recordings. The transition time of the RIRs was calculated using Equation 5.1 in Section 5.2 as 32mS and the measured RIRs were split into their constituent early and late sections. For the synthesis of early reflections, it is necessary to decide which impulse locations are important to use for the interpolation in each source case. For each speaker location the two ends of the microphone array must be used, i.e. positions 1 and 32, as well as the positions that are closest to the source. Table 5.1 shows the real measurements for each source position from which each 32-channel dataset was interpolated from.

Source	4 RIRs	8 RIRs	16 RIRs
1	1,6,19,32	1,6,10,14,16,23,28,32	1, 4:2:32
2	1,16,17,32	1,6,11,16,16,22,27,32	1:2:29, 32
3	1,11,22,32	1,6,11,17,22,26,29,32	1, 4:2:32

Table 5.1: Positions of RIRs used in interpolation.

Figures 5.17, 5.18 and 5.19 show the interpolated early reflections for the centre (source 2), left (source 1) and right (source 3) speaker cases respectively. For each of the three source positions the interpolation based on 16 measured RIRs produces excellent results with minimal

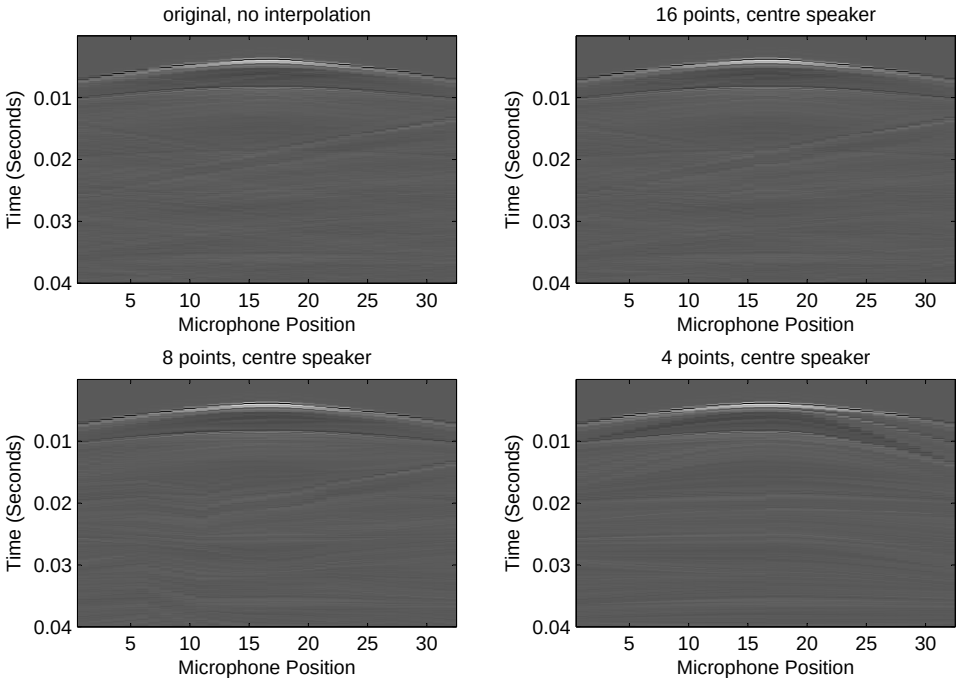


Figure 5.17: Centre source position.

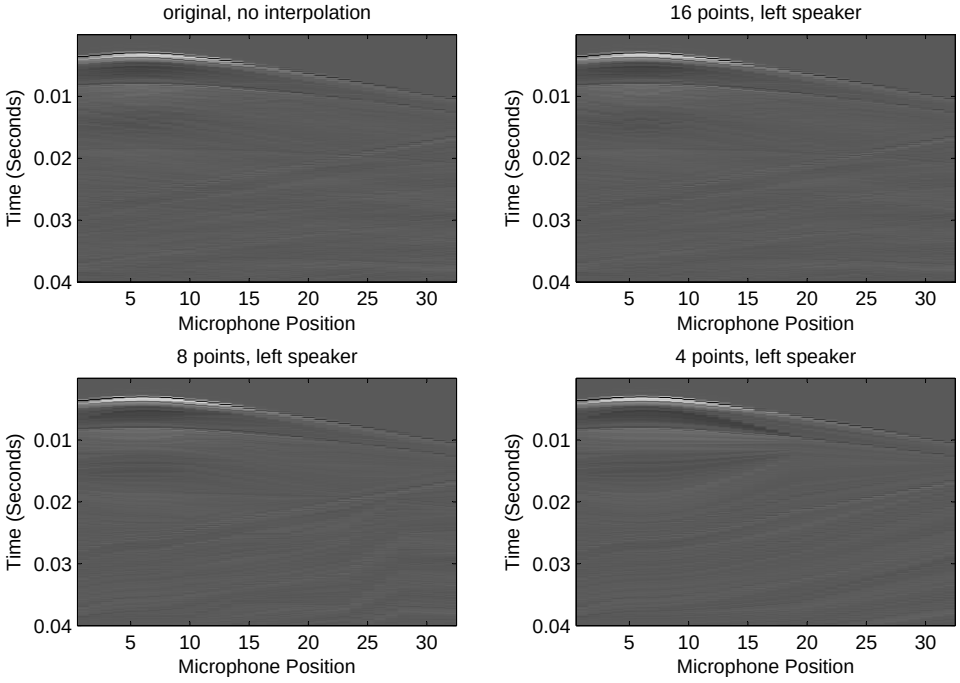


Figure 5.18: Left source position.

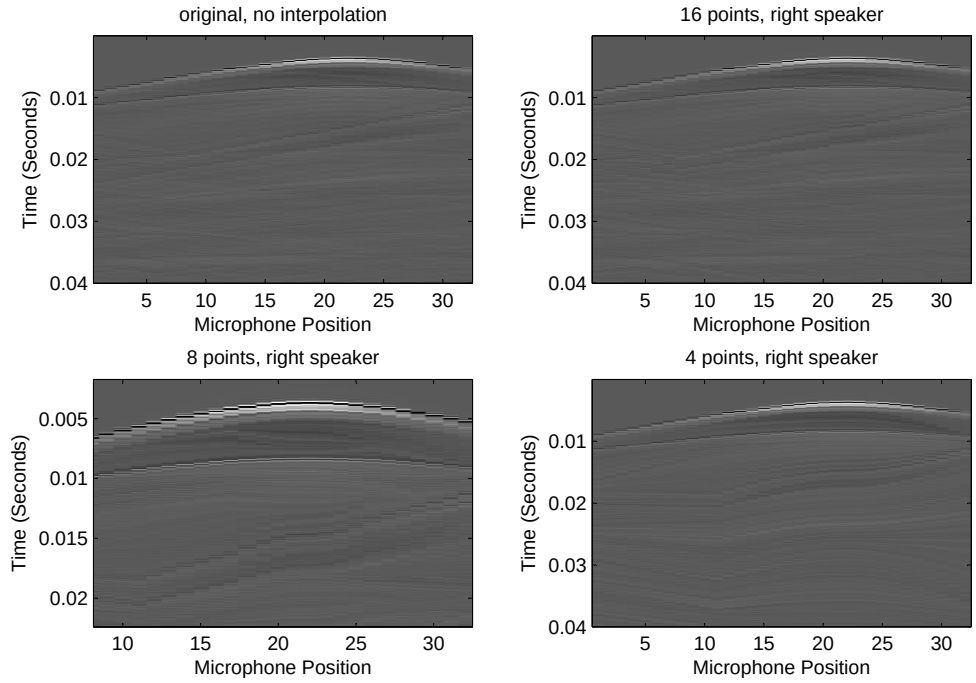


Figure 5.19: Right source position.

errors. The direct sound as well as the first and second order reflections of the floor and walls are well maintained. When the number of measured RIRs is reduced to 8 the direct sound and first order reflection from the floor are kept intact but distortion in the some of the first reflections of the walls becomes evident. A clear example of this can be seen in the bottom left panel of Figure 5.18. If there is not a measured RIR at the point where reflections cross each other then there will be distortion introduced by the interpolation procedure. This is especially evident in the 4 RIR case where there is significant distortion after the direct sound and first floor reflection. However considering the sparseness of the dataset the performance of the interpolation is reasonably good.

5.6.3 Implementation of Wave Field Synthesis

As discussed in Section 5.6.1 the frequency range in which there is accurate wave field reconstruction is limited due to spatial aliasing. For the array used in this work, the spatial aliasing frequency is at 1447Hz. Wittek's stereophonic OPSI technique (as introduced in Section 5.6.1) is used to reproduce the high frequency information.

Recording angles are formed from the perspective of the listener position using selected real and interpolated microphone responses. These angles are based on the localisation theory established by Williams [184] and Thiele and Wittek [187]. The recording angles in this case are primarily due to time differences at the microphones. Linked recording angles are then created

to accommodate the -30°, 0° and 15° sources using reproduction loudspeakers 6, 13, 20 and 27. Critical linking between recording angles can be achieved by employing time or intensity bias in the RIRs (above f_a) [185]. It should be noted, that in the case where B-Format microphones are used to capture the RIRs, virtual microphones can be steered to establish correct recording angles from the perspective of the listener position.

A 20dB per octave linear phase crossover at f_x =1.2kHz was used to separate the low and high frequency portions of the RIR dataset. Prior to reproduction over the array the WFS signals were also subject to the so-called $\sqrt{jk}$ filter, which applies a 3dB per octave boost [17]. The array signals were also tapered using a raised cosine filter, so as to avoid any dispersion effects at the ends of the array.

5.6.4 Objective Analysis

The new datasets of RIRs created by the interpolation and tail synthesis processes described previously were convolved with the swept sine excitation signals for reproduction over the loudspeaker array. A Neumann KU100 dummy head was used to capture the playback of the array at the listener position specified in Figure 5.16. The measured binaural signals were then compensated for the acoustic response of the playback environment by a pre-measured equalization FIR filter following the approach of Kirkeby [97]. Here both the magnitude and phase of the room at the binaural head position are taken into account and a regularization value is set so that large peaks do not occur in the magnitude response when the modulus of the transfer function denominator becomes too large. The binaural signals were then analysed in terms of the Interaural Cross Correlation Function (IACF). Refer back to Section 2.2.1 for an introduction to the IACF.

The point at which the function yields its maximum is known as the Interaural Cross Correlation Coefficient (IACC), and is commonly used as a measure of the acoustic quality in concert halls [140]. The $\mathrm{IACC_{E3}}$ function is employed here as defined by Beranek [16]. $\mathrm{IACC_{E3}}$ is an average measure of the IACC in the 500Hz, 1kHz and 2kHz centred octave bands within the first 80mS, as this frequency region contains wavelengths comparable to the width of the head. Hidaka et al. [77] have shown the relationship between the $\mathrm{IACC_{E3}}$ and apparent source width as

$$\mathrm{ASW} = |1 - \mathrm{IACC_{E3}}| \tag{5.13}$$

and have demonstrated its usefulness in measuring the acoustic quality of concert halls. They show typical ASW values of 0.5 to 0.7 for concert hall acoustics. ASW is also investigated through the width of the main IACC lobe, $\mathrm{W_{IACC}}$, which is defined as the interval of delay time 10% below the IACC peak. This parameter is useful in detecting frequency dependent changes in the ASW when the IACC remains constant.

As mentioned in Section 2.2.1, the time delay at which the IACC is maximum is representative of the position of the source as it gives a value for the ITD. This allows for the localisation

accuracy of the interpolated datasets to be investigated. It is also informative to look at the late IACC values (from 80mS to ∞) which give insight into the diffuseness and decorrelation of the reproduced wavefield.

The results of the binaural measurements are shown in Table 5.2. ITD is fully maintained in each of the interpolation cases suggesting localisation accuracy remains unaffected by interpolation. The early and late IACC measurements also show good degrees of correlation, indicating that the apparent source width remains similar between the full and reduced datasets. W_{IACC} also remains consistent, again indicating no significant changes in ASW. There are however small fluctuations in these values due to the distortion of the early reflection components caused by the interpolation procedure. The late IACC shows significantly lower correlation values than the $IACC_{E3}$ indicating that decorrelation is achieved as expected.

	RIRs	$IACC_{E3}$ (mS)	ITD (mS)	W_{IACC}	$IACC_L$
Pos 1	Full	0.585	0.313	0.194	0.208
	16ch	0.582	0.313	0.194	0.201
	8ch	0.573	0.313	0.195	0.208
	4ch	0.574	0.313	0.194	0.188
Pos 2	Full	0.612	0.104	0.195	0.308
	16ch	0.603	0.104	0.188	0.293
	8ch	0.587	0.104	0.188	0.209
	4ch	0.622	0.104	0.188	0.242
Pos 3	Full	0.597	-0.167	0.181	0.354
	16ch	0.609	-0.167	0.174	0.234
	8ch	0.578	-0.167	0.188	0.326
	4ch	0.552	-0.167	0.174	0.319

Table 5.2: Binaural parameters of reproduced playback.

5.6.5 Perceptual Analysis

A series of perceptual experiments was implemented to further investigate the reproduced wavefield. The primary aim of these tests was to investigate the localisation performance using the 32 channel WFS array for interpolated datasets constructed using 4, 8, 16 and 32 measured RIRs. For the tests, 3 types of sources were prepared: full bandwidth pink noise, and pink noise filtered below and above 1.2kHz. Each test source was a succession of 5 100ms noise bursts, separated by 100ms silences. The noise was convolved with each of the RIR datasets, for each of the 3 measured source positions. This gave a total of 36 localisation tests. Each participant was located at a distance of 1.5m from the centre of the array. 10 listeners took part, each

under 35 years of age and with good hearing. For each pink-noise presentation the participant was asked to identify the direction of localisation using a dedicated software pointer, projected onto an acoustically transparent screen in front of the array (see Figure 5.20). Listeners were

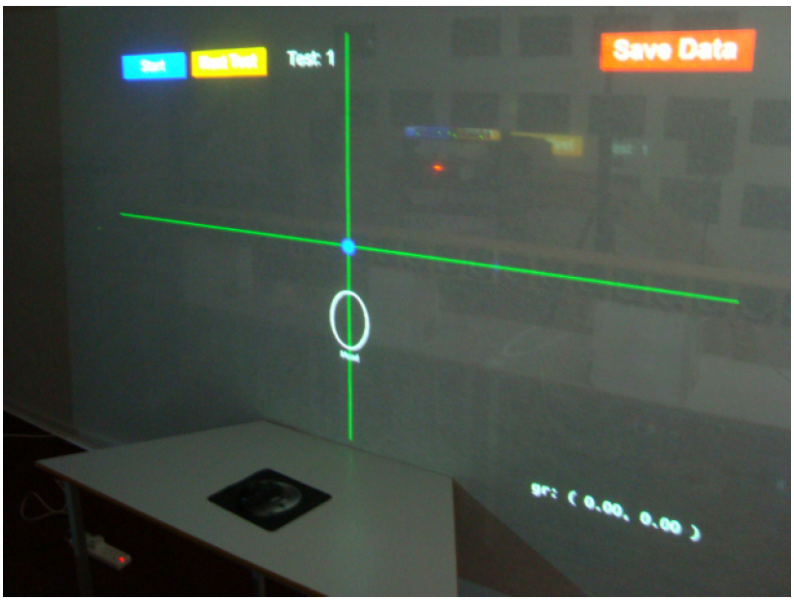
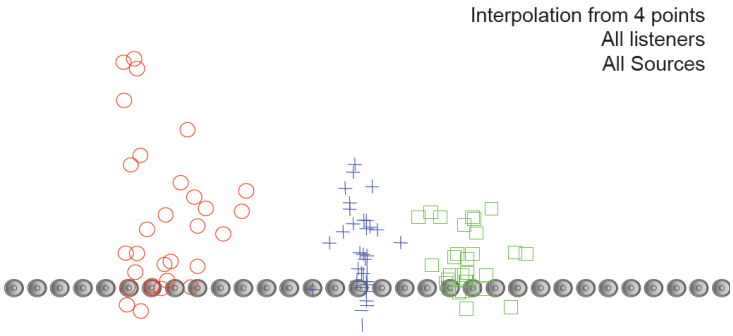


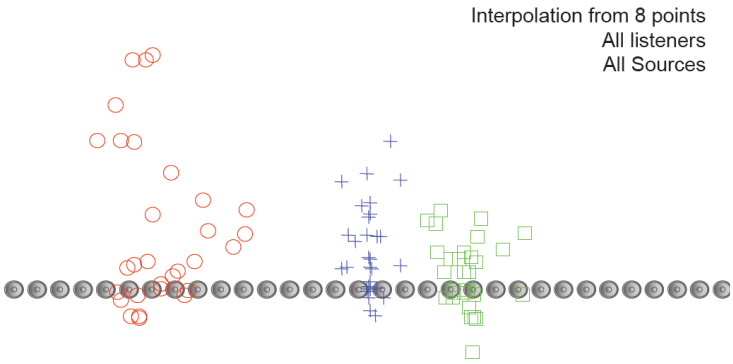
Figure 5.20: Photo of perceptual test setup.

allowed to rotate their heads, but were asked to not move it from the centre position. An on-screen head-sized circle was also projected to assist this. The data acquired using the pointing software was transformed to allow viewing of the total chosen source position for all listeners on a 2-D plot of the array. These scatter plots are shown in Figure 5.21. We see that over all sources a reasonable azimuthal localisation performance is achieved. Azimuthal deviations are greatest for the 15° source, which from the objective analysis holds the lowest $\mathrm{IACC}_{\mathrm{E3}}$ value (0.585 for no interpolation) as well as the largest $\mathrm{W}_{\mathrm{IACC}}$ (0.194 for no interpolation). The best localisation is achieved for the centre source, which holds the highest $\mathrm{IACC}_{\mathrm{E3}}$ value (0.612 for no interpolation). It is difficult however to gauge the true localisation performance from these plots. Instead, we can look at the mean and standard deviation of the localisation of each source over each type of presentation. This is plotted in Figure 5.22.

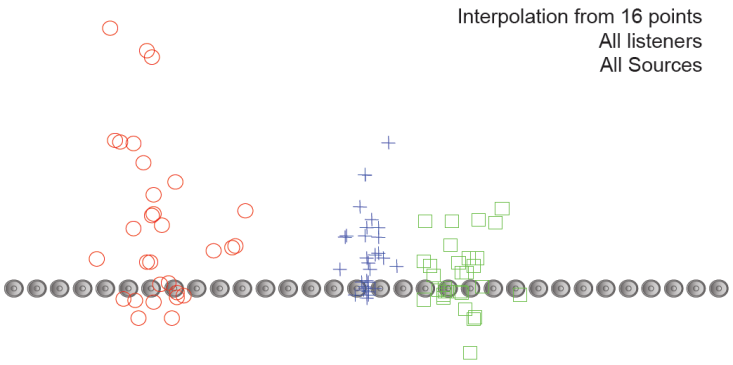
In the case of the pure wavefield presentations, the mean localisation fluctuations between datasets is marginal and equivalent localisation is achieved. Note that the largest deviations (approx 3°) from the full non-interpolated dataset occur with the dataset generated from 4 RIR measurements. This can be attributed to the distortions of the early reflections in this dataset. Localisation above f_a is again good and comparable across all datasets, with standard deviations similar to the low frequency case. This shows the viability of recording angle theory to the OPSI method. The same holds for the full bandwidth case, where localisation is again comparable.



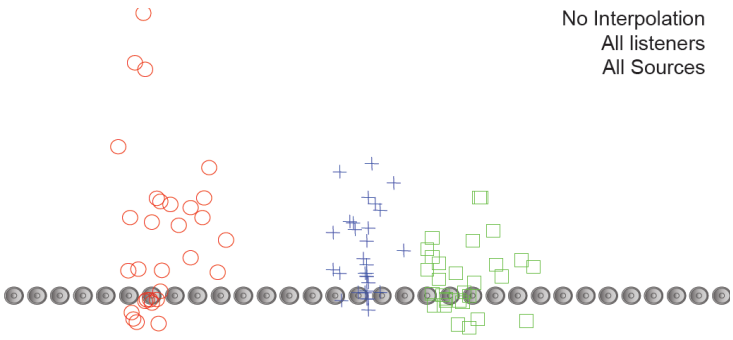
(a) 4 RIRs.



(b) 8 RIRs.

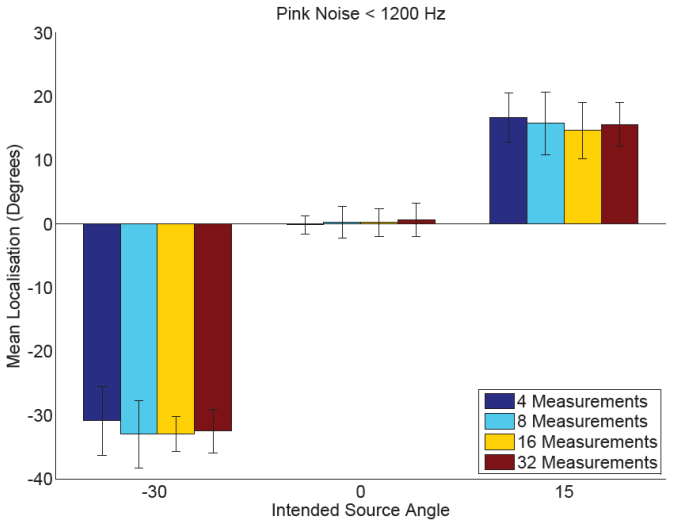


(c) 16 RIRs.

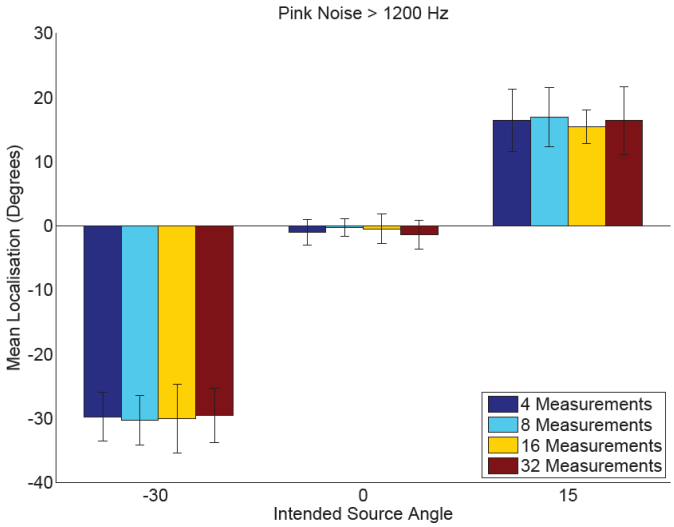


(d) 32 RIRs.

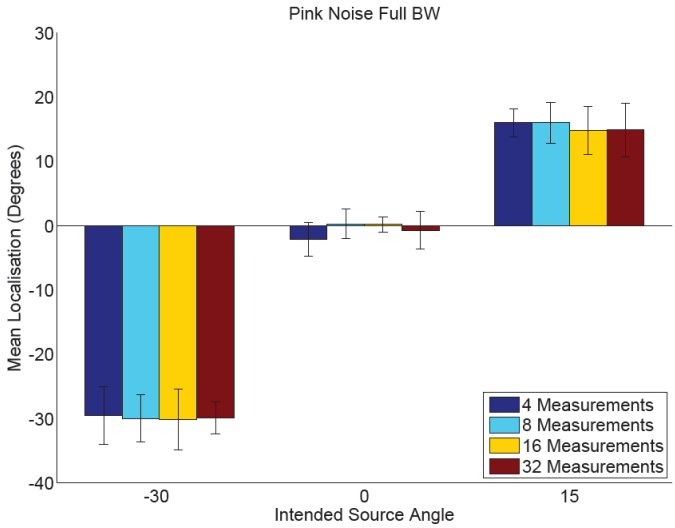
Figure 5.21: Listening Test Results.



(a) $<$ 1200Hz.



(b) $>$ 1200Hz.



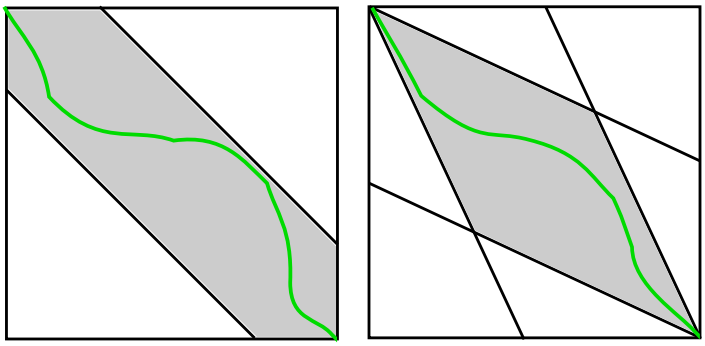
(c) Full bandwidth.

Figure 5.22: Statistical analysis of localisation accuracy.

5.7 Examination of Possible DTW Improvements

The DTW algorithm described in Section 5.3 is the concept in its most simple form. Refinements have been made to the algorithm by several authors in order to optimise it for certain criteria. In this section a brief review of these refined techniques will be undertaken and a comparison of the most relevant of these with the original DTW algorithm will be performed.

Sakoe and Chiba [159] propose applying an adjustment window condition and a slope constraint condition to the algorithm. The adjustment window condition requires the user to choose a window length to define the maximum deviation the warp path can take from the main diagonal of the accumulated distance matrix. This in effect reduces the available accumulated distance matrix to a band (see Figure 5.23a). Itakura [84] introduces a similar condition to this with a diamond shaped allowable area whose boundaries are determined by a user defined slope (see Figure 5.23b). The slope constraint described by Sakoe and Chiba [159] is simply a user defined limit on the slope of the warp path. If the slope becomes too severe (either too steep or too shallow) then the algorithm forces the next point in the path to be one which helps rectify the slope to a more normal value. These conditions are more suited however to the speech



(a) Sakoe band constraint. (b) Itakura parallelogram.

Figure 5.23: DTW constraints.

processing applications to which DTW has traditionally been applied. In these applications a steep gradient in the path through the accumulated distance matrix is unrealistic as it matches a very short part of one sequence to a long section of the other. In the application it is used for in this thesis however, steep gradients in the warp path are likely due to the nature of the RIRs and should be allowed. Hence these methods will not be pursued further in this context.

A more suitable approach may be to use derivative based methods to aid in aligning sharp local changes in the two sequences. Keogh and Pazzani [94] introduced derivative DTW (DDTW). They propose that local differences in feature magnitude between the two sequences can cause traditional DTW to fail somewhat and propose that an alternative difference measure is used instead of Euclidean distance. In the case of DDTW the distance measure is based on the

difference of the local derivatives of the two sequences. The distance matrix is formed as follows

$$D(i,j) = \left|\left| \frac{(a(i) - a(i-1)) + \frac{(a(i)-a(i-1))}{2}}{2} - \frac{(b(j) - b(j-1)) + \frac{(b(j)-b(j-1))}{2}}{2} \right|\right| \tag{5.14}$$

instead of as with Equation 5.2 in Section 5.3.

Xie and Wiltgen [190] propose feature based DTW (FBDTW). The authors state that the local difference metrics used in both DTW and DDTW are not sufficiently descriptive to assure correct warping. Instead they suggest using a mixture of global and local difference metrics to form the cost matrix.

$$D(i,j) = D_{local}(i,j) + D_{global}(i,j) \tag{5.15}$$

The local and global distance matrices, D_{local} and D_{global}, are formed from their relative feature vectors as shown below in Equations 5.16 and 5.17.

$$D_{local}(i,j) = |f_{local}(a(i))_1 - f_{local}(b(j))_1| + |f_{local}(a(i))_2 - f_{local}(b(j))_2| \tag{5.16}$$

$$D_{global}(i,j) = |f_{global}(a(i))_1 - f_{global}(b(j))_1| + |f_{global}(a(i))_2 - f_{global}(b(j))_2| \tag{5.17}$$

The global feature metric for a given sample of either of the two sequences a and b is a two element vector. The first element is the difference between the sample value and the mean of all the preceding elements while the second element is the difference between the sample value and the mean of all the proceeding samples. This is formulated in Equation 5.18 for sequence a. $f_{global}(b(i))$ is calculated in the same way.

$$f_{global}(a(i)) = \left[a(i) - \frac{\sum_{k=1}^{i-1} r(k)}{i-1},\ a(i) - \frac{\sum_{k=i+1}^{M} r(k)}{M-i} \right] \tag{5.18}$$

The local feature metric for sequence a is calculated as shown in Equation 5.19. $f_{local}(b(i))$ is calculated in the same way.

$$f_{local}(a(i)) = [a(i) - a(i-1),\ a(i) - a(i+1)] \tag{5.19}$$

A brief comparison of the usage of DDTW or FBDTW instead of DTW for RIR interpolation was completed. DDTW or FBDTW was used in place of DTW as the warping technique. The same weightings and dewarping technique were used in all cases. This was applied to the 32 RIR dataset for the centre source position described in Section 5.6.2. The MSE results for the usage of each of the three methods (DTW, DDTW and FBDTW) for interpolation when compared with the original full dataset are shown below in Table 5.3 for the 4, 8 and 16 RIR cases. The positions of the RIRs used in each case are as documented in Table 5.1. This comparison shows DTW outperforming DDTW by a small margin for each of three interpolations. FBDTW provides a miniscule MSE improvement over DTW in the 16 RIR interpolation case but is outperformed by DTW in the other two interpolation cases, especially for the 4 RIR case. It appears that the simplest technique is sufficient for this application.

No. RIRs	DTW	DDTW	FBDTW
16	2.88347e-006	3.02838e-006	2.87530e-006
8	2.33037e-005	2.37136e-005	2.35686e-005
4	4.85908e-005	4.96482e-005	6.12989e-005

Table 5.3: MSE comparison for 3 different RIR availability cases.

5.8 Conclusion

Large grids of measured RIRs distributed throughout the virtual listening environment for various source and receiver positions are necessary for convincing auralisation. In order to make this measurement task more efficient a robust spatial interpolation technique is required. A novel DTW based approach to early reflection interpolation has been presented in this chapter which, when coupled with the synthesis of the diffuse tail, allows for accurate spatial interpolation of RIRs. The application of the developed technique to measured RIRs shows good results even when only a small number of measured responses are used in the reconstruction. Both objective and perceptual tests on a wave field synthesis implementation indicate that there is no significant degradation of localisation quality brought about by interpolation. A brief study of more advanced DTW based techniques for use in the warping of early reflections yielded no significant improvements on the original DTW algorithm.

6

Virtual Auditory Environments

In previous chapters the contribution of head related filtering to spatial sound localisation and the effect of the environment on sound have been discussed. These elements are both integral to achieving convincing auralisation. This chapter first explores the concept of the Binaural Room Impulse Response (BRIR) in relation to the Room Impulse Response (RIR) and the Head Related Impulse Response (HRIR). Methods for reproduction of spatial audio over headphones using BRIRs are discussed, with particular attention being paid to the virtual loudspeaker approach. An introduction to Ambisonics encoding and decoding is given as a basis from which to explore the virtual loudspeaker approach. A real time implementation of the Ambisonics based virtual loudspeaker approach is then described which is responsive to head tracking information from the listener. The use of approximate HRIR factorisation and RIR interpolation are discussed in this context.

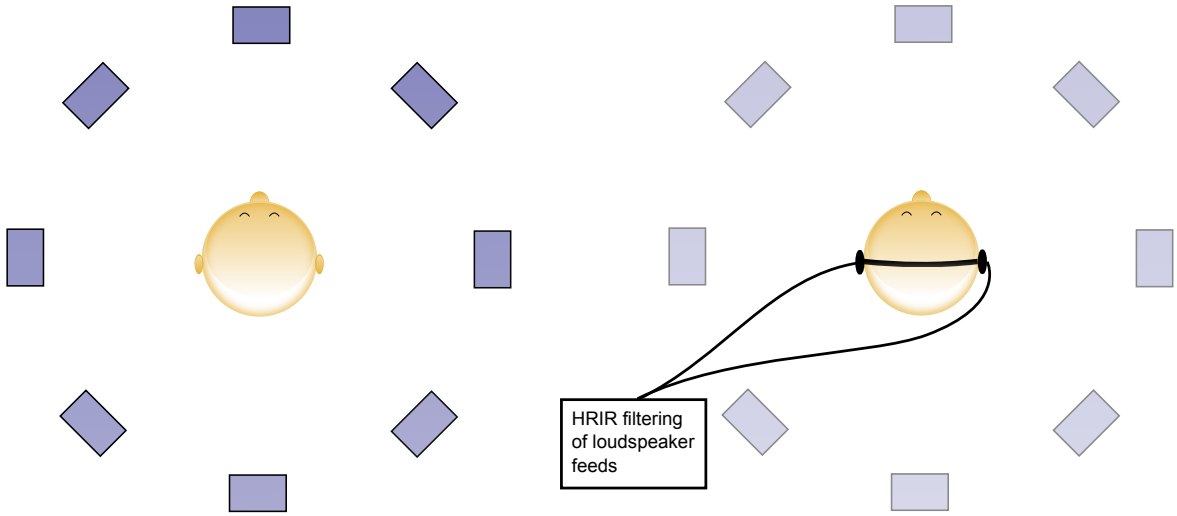
6.1 Binaural Room Impulse Response

HRIRs are measured in an anechoic environment and as such contain only reflections of the pinna and body of the subject. When a person hears a sound in a natural environment the signal received has been processed by both the transmission through the room and through the head response. This composite is the convolution of the RIR and the HRIR and is called the Binaural Room Impulse Response (BRIR). There is a lack of flexibility and practicality inherent in the measurement of such responses. The BRIR set are particular to the auditory environment in which they were measured. While the responses cater for the sound source moving to any

position on the specified measurement grid, if, in an interactive application, the listener wants to move to another position in the room, further BRIRs are required. So ideally, different listeners would require their own measured BRIRs for each environment and for a large number of positions in that environment to allow for maximum flexibility.

BRIRs can also be synthesised by combining measured or synthesised HRIRs and RIRs. There are two main techniques that can be used to achieve this combination. The first involves an analysis of the early reflection component of the RIRs which separates the individual reflections and determines their angle of incidence. The relevant HRIR for that direction is then convolved with the relevant reflection and the results of these convolution operations are summed together. Ahnert and Reiner implement such a system in [2]. They use a commercially available package called EASE to synthesise the RIR data. This package calculates the RIRs based on a CAD model of the room and given source and receiver positions. It also gives the angle at which specific reflections impinge on the receiver. This involves significant processing, especially if one considers that this may well be implemented in a real time, interactive system. More recently Menzer and Faller [118] have implemented a similar system using measured B-format RIRs. The B-format concept will be explored in Section 6.2. The extraction of the early reflections and their directions of arrival is done using a measurement based technique formulated by Merimaa and Pulkki [119] and hence no model of the environment is necessary. This still requires a dense set of HRIRs and significant processing.

An alternative approach based on virtual loudspeakers has been proposed by McKeag and McGrath [116] and further elaborated upon by Noisternig et al. [138]. In this case a set of HRIRs are measured at only a small number of positions. These positions correspond to the locations of loudspeakers in a virtual loudspeaker array which the listener is positioned at the centre or 'sweet spot' of, as illustrated in Figure 6.1. The loudspeaker feeds are convolved with the relevant HRIRs and the results for each ear are summed and played binaurally. The loudspeaker feeds are generated from B-format Ambisonic source data. McKeag and McGrath propose that this B format source data can be synthesised or measured using a sound field microphone. In Noisternig's paper [138] this audio is synthesised based on the image source method. First and second order image sources are calculated based on a simple user defined room geometry. The original source audio is delayed and attenuated depending on the location of these image sources. The reverberation tail is simulated separately by low pass filtering the output of a recursive reverberation network. The direct sound, early reflection and reverberation tail components are then encoded into B-format where each component (W, X, Y and Z for first order) is summed. Before further elaborating on the virtual loudspeaker approach to binaural sound reproduction, it is first necessary to briefly explore the background theory of Ambisonics.



(a) Real Loudspeaker Listening. (b) Virtual Loudspeaker Approach.

Figure 6.1: Illustration of Virtual Loudspeaker Approach.

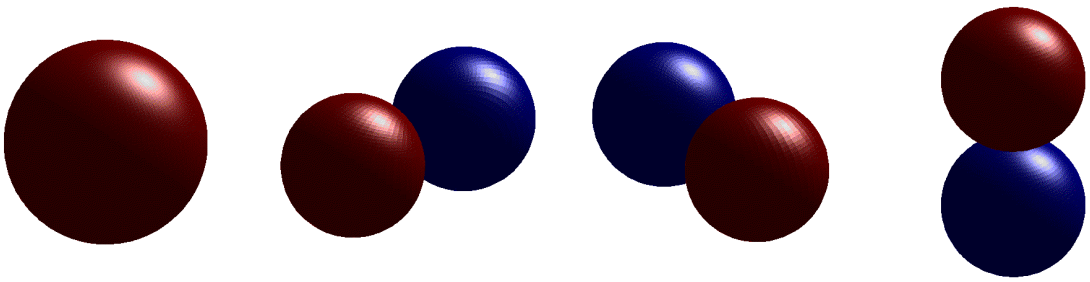
6.2 Ambisonics

Ambisonics is a flexible recording and reproduction process for multichannel loudspeaker playback. It was developed by Michael Gerzon in the 1970s in Oxford. The technique aims to record and reproduce the entire sound field at a given position and is based on the spherical harmonic decomposition of the sound field. It is different to normal recording and reproduction techniques in that recording and reproduction are disengaged from each other. Another key difference is that it aims to correctly recreate the wavefronts within the reproduction array rather than synthesising certain localisation cues at a given sweet spot as is the case with most other reproduction techniques. One significant advantage of Ambisonics is that all directions are given equal weight and therefore there is not the same concentration on frontal sources that is evident in other popular techniques.

The simplest and best known version is B-format first order Ambisonics. In this case the sound field is encoded into four channels: W, X, Y and Z (see Figure 6.2). W is the omnidirectional zero order component while X, Y and Z are figure of eight first order velocity components orientated along the x, y and z Cartesian axes. For horizontal only reproduction only W, X and Y are needed. Linear combinations of these channels can be decoded to form virtual microphones. Each component is weighted differently in the linear combination based on the angle of the virtual microphone and the resultant output is fed to the loudspeaker at that angle.

6.2.1 B-format Recording and Encoding

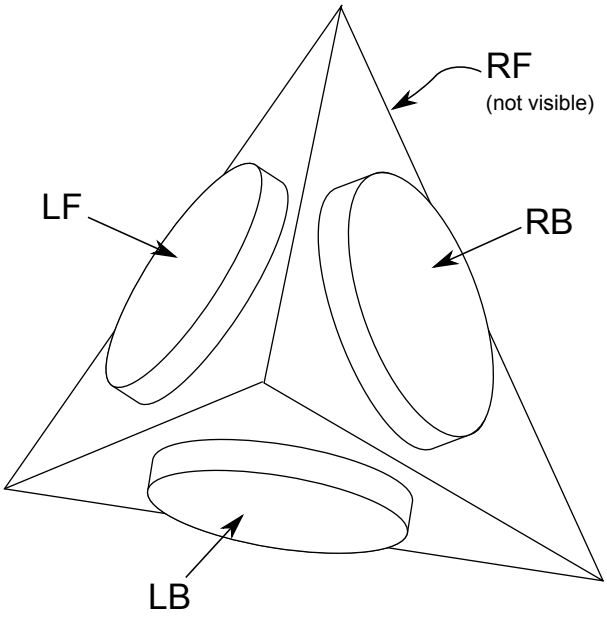
First order B-format channels can be measured using a soundfield microphone. A commercially available example is the Soundfield MK5 as shown in Figure 6.3a. Generally four cardioid



(a) W (b) X (c) Y (d) Z

Figure 6.2: First order B-format Spherical Harmonics.

(a) Soundfield MK5 Microphone.



(b) Tetrahedral Microphone Array.

Figure 6.3: Soundfield Microphone.

microphones arranged in a tetrahedral array are used to measure the soundfield. The set of four signals is referred to as A-format. This configuration was patented by Craven and Gerzon [40] in 1977 and is as shown in Figure 6.3b. LF is the left-front orientated cardioid, RB is the right-back orientated cardioid and so on. The B-format signal are derived from these using simple sum and difference equations as follows

$$W = LF + LB + RF + RB \tag{6.1}$$

$$X = LF - LB + RF - RB \tag{6.2}$$

$$Y = LF + LB - RF - RB \tag{6.3}$$

$$Z = LF - LB - RF + RB \tag{6.4}$$

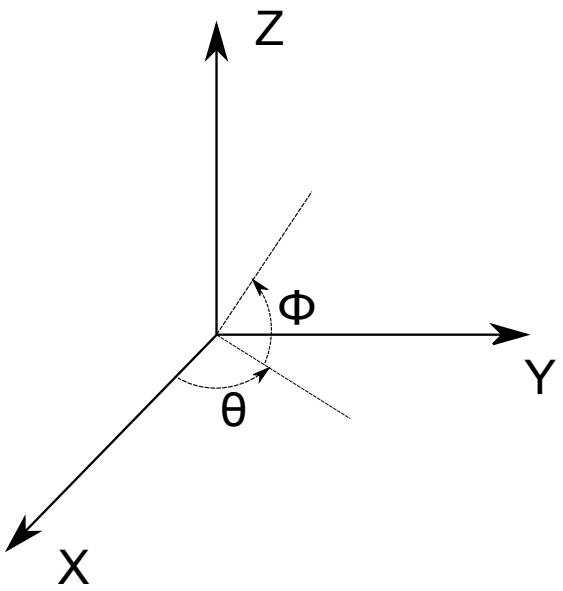


Figure 6.4: Frame of reference.

This conversion from A-format to B-format is generally implemented in hardware. At this point let us consider the frame of reference from which angular positions and axes are being referenced in order to maintain coherency throughout this exploration of B-format Ambisonics. Figure 6.4 demonstrates this.

Existing mono source recordings can be encoded into B-format based on an assigned position on the unit reproduction sphere using the following equations.

$$W = s.\frac{1}{\sqrt{2}} \tag{6.5}$$

$$X = s.\cos\theta\cos\phi \tag{6.6}$$

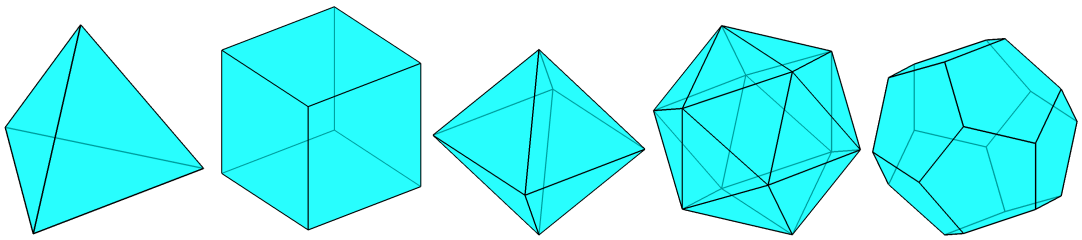
$$Y = s.\sin\theta\cos\phi \tag{6.7}$$

$$Z = s.\sin\phi \tag{6.8}$$

where s is the mono source, θ is the horizontal angle (anti-clockwise) and ϕ is the elevation angle. If multiple mono sources are present then each can be encoded based on their own specific positional information. The W, X, Y and Z components can then be simply summed and the resulting sum normalised by the number of sources.

6.2.2 B-Format Decoding

The Ambisonic encoded signals do not directly feed loudspeakers. Instead they can be decoded to practically any loudspeaker configuration. One fundamental limit on this is that the number of loudspeakers must be greater than the number of Ambisonic channels. In the case of first



(a) Tetrahedron. (b) Cube. (c) Octahedron. (d) Icosahedron. (e) Dodecahedron.

Figure 6.5: The Platonic solids.

order horizontal only B-format this means there must be a minimum of three loudspeakers, while the minimum for full sphere reproduction is four. While it is possible to decode for any loudspeaker configuration that satisfies this minimum channel number rule, it is advisable to make the loudspeaker array as regular as possible for optimum performance.

If a form of regularity is imposed on the loudspeaker array then the decoded loudspeaker feeds can be easily calculated by decoding through projection. The loudspeaker feed, L_i, at a given position described by the angles θ_i and ϕ_i, is as follows

$$L_i = \frac{1}{\sqrt{2}}W + (cos\theta_i cos\phi_i)X + (sin\theta_i cos\phi_i)Y + (sin\phi_i)Z \tag{6.9}$$

In the 3D case there is a very limited number of loudspeaker arrays that satisfy this regularity constraint on speaker spacing. This group of regular 3D objects is called the Platonic solids and consists of the tetrahedron, the cube, the octahedron, the icosahedron and the dodecahedron (shown in Figure 6.5). The octahedron and the icosahedron are the only two of these that offer a ring of loudspeakers in the horizontal plane. For further information refer to the following theses [41, 79].

Decoding can also be performed using the pseudo-inverse operator. B is defined as a column vector made up of the Ambisonics channels, p is the column vector of loudspeaker feeds and C is the re-encoding matrix i.e.

$$B = [WXYZ]^T \tag{6.10}$$

$$B = Cp \tag{6.11}$$

and hence

$$p = C^{-1}.B \tag{6.12}$$

C^{-1} is referred to as the decoding matrix. C will be square and invertible in the normal way if the number of loudspeakers is equal to the number of Ambisonics channels. Otherwise the pseudoinverse operator is used to solve for p which offers the least squares solution.

$$p = C^T(C.C^T)^{-1}.B \tag{6.13}$$

The condition number of C must be small in order for this technique to provide useful solutions [130] and is defined as follows using the matrix norm

$$cond(C) = \|C\|_2 \|C^+\|_2 \tag{6.14}$$

where the + superscript denotes the pseudoinverse operator.

$$C^+ = C^T (C.C^T)^{-1} \tag{6.15}$$

For perfectly regular loudspeaker layouts both decoding by projection and pseudoinverse are equivalent [41].

6.2.3 Decoder Optimisation

The decoding process should be optimised in order to maximise localisation accuracy. As discussed in Section 2.2 ITD is the dominant cue at low frequencies while ILD is more dominant at high frequencies. To this end Gerzon [61, 63] devised optimised velocity and energy vector decoding based on these psychoacoustic criterion. The aim is make both the velocity vector, r_e, and the energy vector, r_v, as close to 1 as possible.

For regular horizontal only reproduction this is achieved by using shelf filtering as shown in Figure 6.6. The frequency response above 700Hz is boosted for the W component and attenuated for the X and Y velocity components. The magnitude difference at high frequencies optimises the energy decode while the dominance of the low frequencies for the velocity components optimises the velocity decode. See [62, 88, 181] for further details.

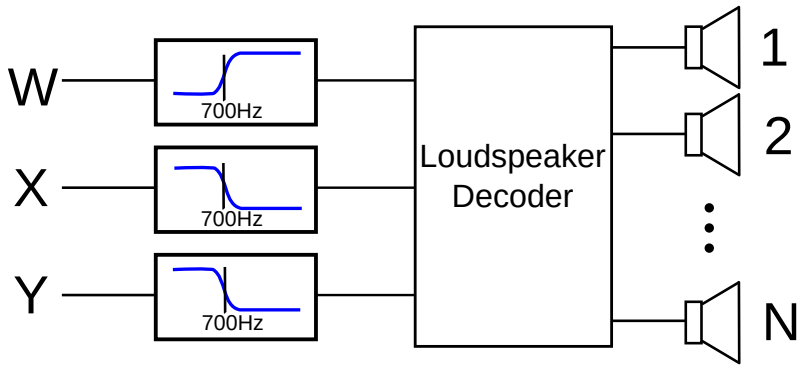


Figure 6.6: Shelf filtering.

6.2.4 Near Field Effect

Ambisonics assumes that reproduction loudspeakers and the virtual sources are in the far field. This means that in all but the largest of reproduction installations it is necessary to implement

near field compensation to compensate for the proximity effect which results in a low frequency boost. The proximity effect is caused by the non planar wavefronts emitted by near field sources. Daniel [42] proposes the introduction of such compensation at the encoding stage with special consideration to Higher Order Ambisonics (HOA). For the first order horizontal applications in this thesis it is sufficient to filter the velocity components (i.e. X, Y) with a high pass filter with a -3dB point at $c/(2\pi * \text{loudspeaker array radius in m})$ Hz where c is the speed of sound in air [107].

6.2.5 Rotating the Soundfield

A significant advantage that Ambisonic encoding offers, and one which is particularly advantageous for the virtual loudspeaker approach, is that the soundfield can be easily rotated in any direction using simple matrix operations. The W component will remain unaffected by these transformations as it is omnidirectional. In order to rotate the soundfield about the Z axis by an angle θ the following operations are applied

$$W' = W \tag{6.16}$$
$$X' = X\cos\theta + Y\sin\theta \tag{6.17}$$
$$Y' = -X\sin\theta + Y\cos\theta \tag{6.18}$$
$$Z' = Z \tag{6.19}$$

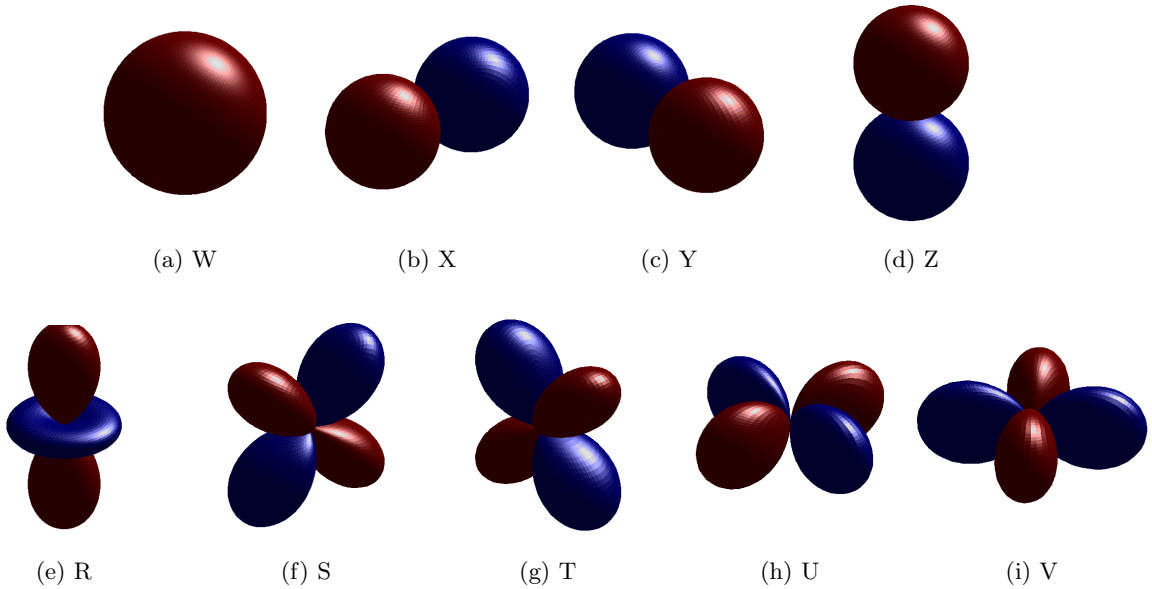
In order to tilt the soundfield (i.e. rotate about the X axis) the following operations are applied

$$W' = W \tag{6.20}$$
$$X' = X \tag{6.21}$$
$$Y' = Y\cos\phi - Z\sin\phi \tag{6.22}$$
$$Z' = Y\sin\phi + Z\cos\phi \tag{6.23}$$

In order to tumble the soundfield (i.e. rotate about the Y axis) the following operations are applied

$$W' = W \tag{6.24}$$
$$X' = X\cos\phi - Z\sin\phi \tag{6.25}$$
$$Y' = Y \tag{6.26}$$
$$Z' = X\sin\phi + Z\cos\phi \tag{6.27}$$

To allow for angular rotation of the full soundfield it is necessary to combine two of these as shown below. The matrix shown in Equation 6.29 shows the combined implementation of both



(a) W (b) X (c) Y (d) Z

(e) R (f) S (g) T (h) U (i) V

Figure 6.7: Second order B-format Spherical Harmonics.

the rotate and tilt operations.

$$\begin{pmatrix} X' \\ Y' \\ Z' \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \tag{6.28}$$

$$= \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta\cos\phi & \cos\theta\cos\phi & -\sin\phi \\ \sin\theta\sin\phi & \cos\theta\sin\phi & \cos\phi \end{pmatrix} \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \tag{6.29}$$

6.2.6 Higher Order Ambisonics

The previous sections discuss first order B-format Ambisonics. It is possible to extend this to Higher Order Ambisonics (HOA). This offers benefits such as an increase in the size of the sweet spot at which the soundfield is correctly reproduced and better localisation. Figure 6.7 shows the spherical harmonics required for second order Ambisonic reproduction. The first rows shows the zeroth and first order components shown earlier while the second row shoes the second order components. While the soundfield is more accurately encoded as the Ambisonic order increases, recording of the soundfield becomes more difficult. Recall that the first order soundfield can be captured using a tetrahedral array of 4 cardioid microphones. The second order soundfield requires a dodecahedral, or similarly complex, microphone array [39]. The complexity of design increases with order. More recently Moreau et al. [130] have proposed a design for a fourth order microphone made up of 32 sensors. The minimum number of loudspeakers needed to reproduce the soundfield is also increased as the number of loudspeakers must be greater than

or equal to the number of Ambisonics channels. For 3D reproduction the number of channels is given by $(M+1)^2$ where M is the order of the system. For 2D reproduction this falls to $2M+1$. Recall from Section 6.2.2 that there are only 5 perfectly regular 3D shapes and of these the dodecahedron offers the highest number of vertices at 20. This means that for high orders there must be some deviation from the Platonic solids (see [41, 79] for further information on this). The increase in the number of channels also makes rotation matrices etc considerably more complex.

6.3 Virtual Loudspeaker Approach

The application of Ambisonics to the virtual loudspeaker approach was briefly introduced in Section 6.1 of this chapter. The convolution of mono source audio with first order B-format Ambisonic RIRs is the preferred method in this thesis for the production of the loudspeaker feeds. The choice of a regularly spaced loudspeaker array allows for decoding by projection of the Ambisonics channels to obtain the loudspeaker feeds. As alluded to in Section 6.2.5 the soundfield can be easily rotated by applying simple rotation matrices to the B-format Ambisonic channels. Hence there is only a need to measure a small static set of HRIRs, one for each ear for each of the discrete loudspeaker positions, as rotation of the soundfield can be used to compensate for head movement and switching HRIRs is not necessary. This reduces auditory glitches in the playback system as well as offering significant time and memory savings.

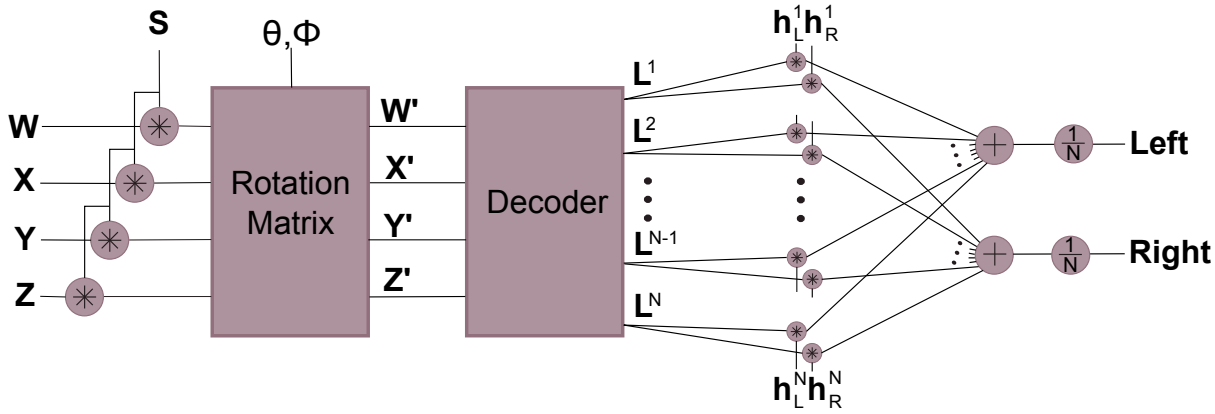


Figure 6.8: Virtual loudspeaker approach.

The block diagram shown in Figure 6.8 outlines the processing steps of the virtual loudspeaker approach. S is the mono source audio, θ and ϕ describe the rotation of the head in azimuth and elevation respectively, L^i is the virtual loudspeaker feed for position i and h_L^i and h_R^i are the left and right ear HRIRs for that position.

The rotation block is governed by the equations introduced in Section 6.2.5. W, X, Y, Z are converted to W', X', Y', Z' based on the angular rotation of the head described by the

angles θ and ϕ. These angles are obtained using a head tracking apparatus, several of which were investigated in the course of this work. These included optical based systems involving infrared emitters and cameras. However due to the small size and fast speed of natural head movements and due to the fact that it was desirable to afford complete rotational freedom to the user, optical techniques were found to be mostly insufficient. For high accuracy and freedom of movement a precise accelerometer is desirable such as the InertiaCube headtracker [82].

The process of decoding the B-format channels to a set of virtual loudspeaker channels, L^i is described in Section 6.2.2. Regularity will be imposed on the practical implementations in this thesis and hence, as described in the previous section, each of the loudspeaker feed can be calculated as follows

$$L_i = \frac{1}{\sqrt{2}}W + (cos\theta_i cos\phi_i)X + (sin\theta_i cos\phi_i)Y + sin\phi_i Z \tag{6.30}$$

The left and right ear headphone feeds are then calculated, as shown graphically in Figure 6.8, by the following equations

$$\text{Left} = \sum_{i=1}^{N} L^i * h_L^i \tag{6.31}$$

$$\text{Right} = \sum_{i=1}^{N} L^i * h_R^i \tag{6.32}$$

where each loudspeaker feed, L^i is convolved with the relevant HRIR for each each ear, h_L^i or h_R^i, and the resulting signals summed. N is the number of loudspeakers in the reproduction array. The $\frac{1}{N}$ component prevents clipping of the audio signal when all the loudspeaker have been summed together.

6.3.1 Application of HRIR Factorisation to VLA

As the convolution of the loudspeaker feeds with the relevant HRIRs must be done in real-time, it would be desirable to shorten the filter lengths. Novel HRIR factorisation techniques have been introduced in Chapter 4 and their application to this problem yields shorter filters for real-time convolution. The factorisation techniques allow for a direction independent component to be extracted from a HRIR data set. As a result each HRIR can be considered as the convolution of this common component and a direction dependent component. If this technique is applied to the set of HRIRs (both left and right ear) corresponding to the virtual loudspeaker positions, then each HRIR can be rewritten as

$$h_L^i = f * g_L^i \tag{6.33}$$

$$h_R^i = f * g_R^i \tag{6.34}$$

where f denotes the direction independent component and g_L^i and g_R^i denote the direction dependent components. The mono source audio, S, can be pre-convolved with the direction

independent component offline and the direction dependent component can replace the longer complete HRIRs at the outputs of the loudspeaker decoder. Hence the left and right ear headphone feeds are calculated as follows

$$\text{Left} = \frac{1}{N}\sum_{i=1}^{N} L^i * g_L^i \tag{6.35}$$

$$\text{Right} = \frac{1}{N}\sum_{i=1}^{N} L^i * g_R^i \tag{6.36}$$

A block diagram of this improved system is shown in Figure 6.9.

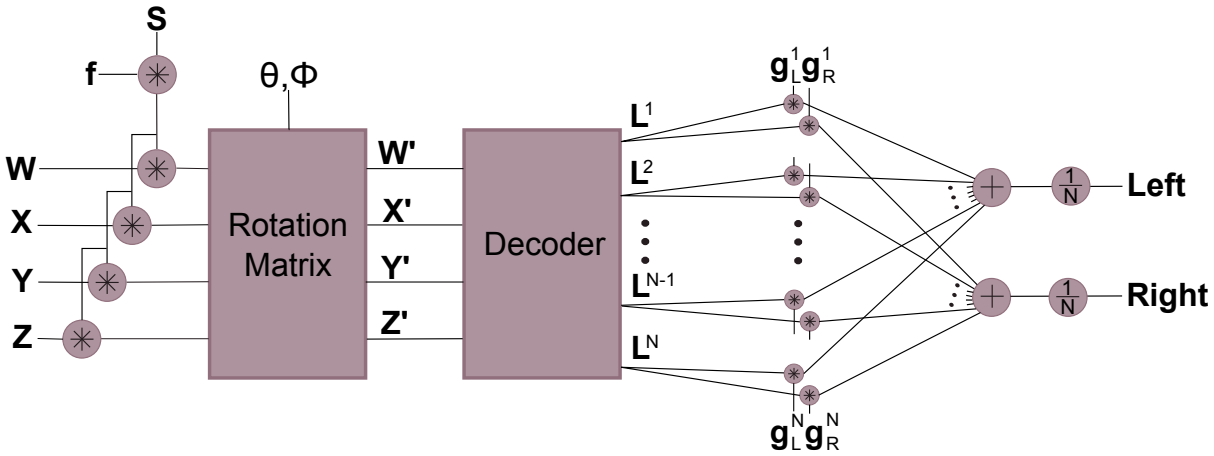


Figure 6.9: Virtual loudspeaker approach incorporating HRIR factorisation.

6.3.2 Application of RIR Interpolation to VLA

In the previous sections auralisation is considered with the limitation that the listener is positioned at one position in the virtual listening environment. While rotation of the head is allowed, no translational movements are facilitated. In order to allow for this large datasets of B-format RIRs are required for various source/receiver positions. This involves much tedious measurement and significant storage capabilities. In Chapter 5 a technique for RIR interpolation was developed which is applicable to this problem.

In Chapter 5 the early reflection interpolation technique was applied to mono RIRs. The technique can also applied separately to the four first order B-format channels. A linear array of 32 measurements was taken in the Printing House Hall in Trinity College at a height of 1.65m with a 1m spacing between the source and the array as illustrated in Figure 6.10. The spacing between measurements was 0.125m. Figure 6.11 demonstrates the effectiveness of the B-format interpolation. The second column shows the RIRs for the case where the RIRs at even positions have been discarded and reformed by interpolation. Only positions 1 to 31 are shown

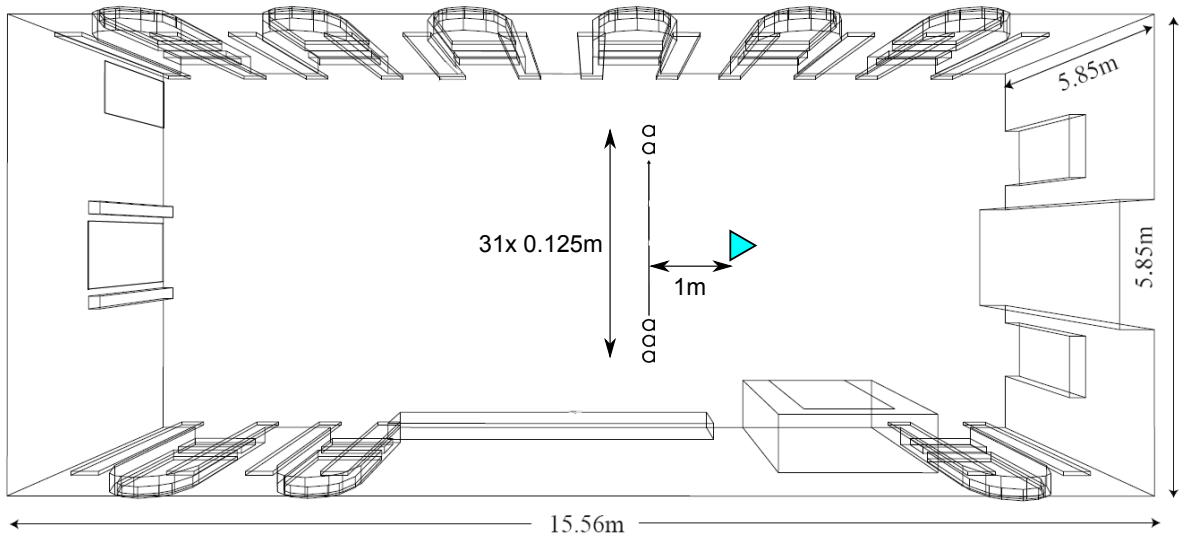


Figure 6.10: Illustration of microphone and loudspeaker positions in the Printing House Hall.

as extrapolation for position 32 is not possible. The first column shows the original measured responses. As can be seen the interpolation technique is very effective when applied in this manner.

6.4 Real-time Implementation

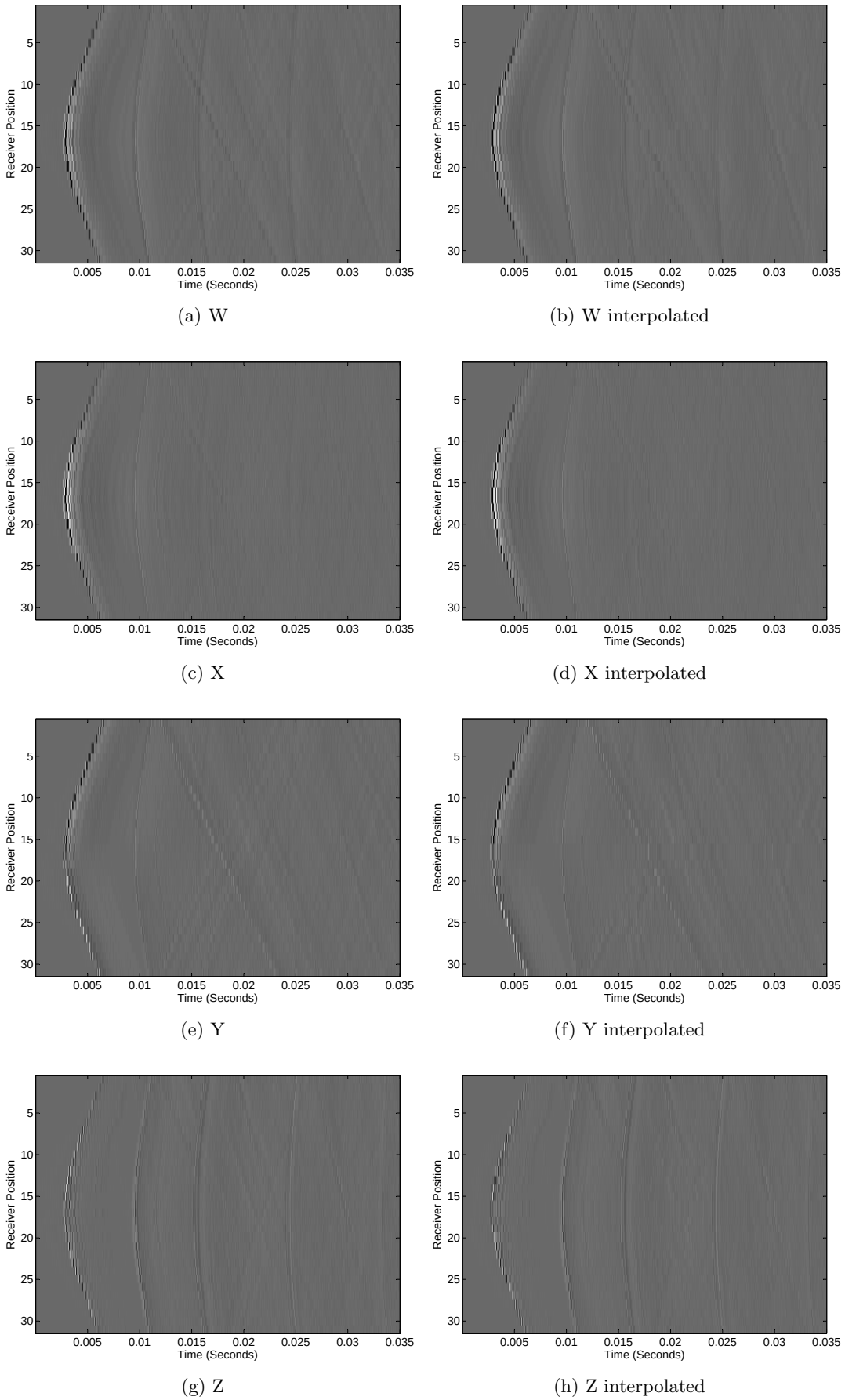
The block diagrams in the previous section (Figures 6.8 and 6.9) show theoretically how the virtual loudspeaker approach is arranged. An actual implementation of the system in real time involves reconfiguring the layout slightly to allow for efficiency and ease of coding. The real time systems in this thesis implement horizontal only reproduction of the sound field. This simplifies the calculation of the loudspeaker feeds from the first order B-format components. By letting $\phi = 0$, Equation 6.30 becomes

$$L_i = \frac{1}{\sqrt{2}}W + (cos\theta_i)X + (sin\theta_i)Y \tag{6.37}$$

Eight regularly spaced virtual loudspeakers are used, positioned as shown in Figure 6.12

The system as defined in Figure 6.8 can hence be written as follows (only left ear feed considered for brevity):

$$\begin{aligned}\text{Left} = S * [&(\frac{1}{\sqrt{2}}W + \cos\theta_1 X' + \sin\theta_1 Y') * h_l(\theta_1) \\ &+ (\frac{1}{\sqrt{2}}W + \cos\theta_2 X' + \sin\theta_2 Y') * h_l(\theta_2) \\ &+ \dots \\ &+ (\frac{1}{\sqrt{2}}W + \cos\theta_8 X' + \sin\theta_8 Y') * h_l(\theta_8)]\end{aligned}$$



(a) W

(b) W interpolated

(c) X

(d) X interpolated

(e) Y

(f) Y interpolated

(g) Z

(h) Z interpolated

Figure 6.11: Comparison of original and interpolated B-format RIRs.

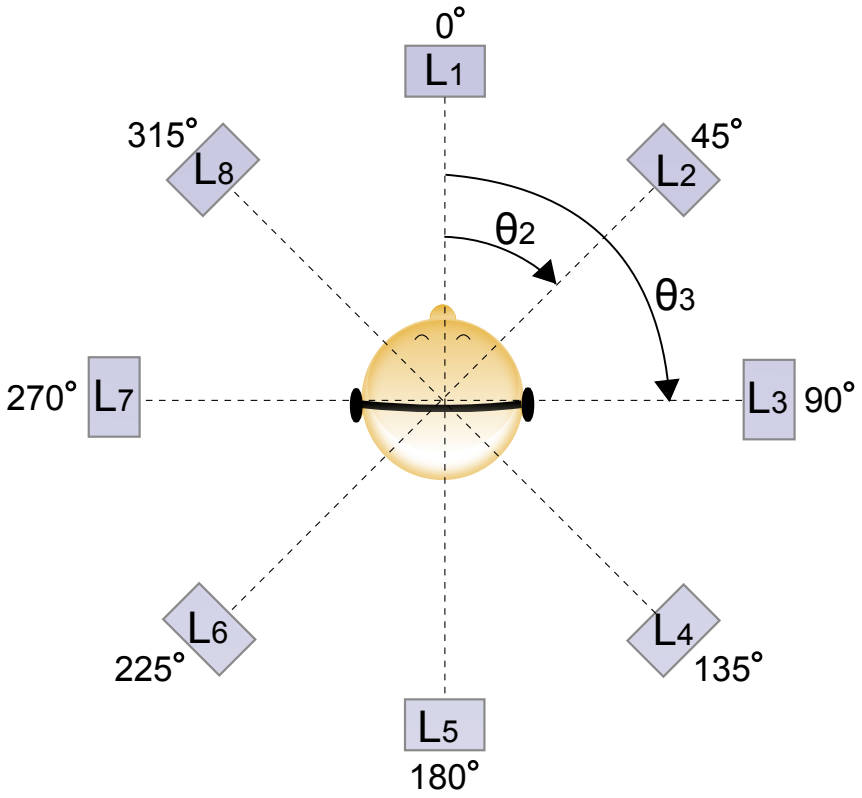


Figure 6.12: Virtual loudspeaker array.

Due to the distributive property of convolution this can rearranged to give

$$\begin{aligned}\text{Left} = S * [W * (\frac{1}{\sqrt{2}}h_l(\theta_1) + \frac{1}{\sqrt{2}}h_l(\theta_2) + \cdots + \frac{1}{\sqrt{2}}h_l(\theta_8)) \\ + X' * (\cos\theta_1 h_l(\theta_1) + \cos\theta_2 h_l(\theta_2) + \cdots + \cos\theta_8 h_l(\theta_8)) \\ + Y' * (\sin\theta_1 h_l(\theta_1) + \sin\theta_2 h_l(\theta_2) + \cdots + \sin\theta_8 h_l(\theta_8))]\end{aligned}$$

This can be reformulated as

$$\text{Left} = S * (W * W_{HRIR_L}) + (X' * X_{HRIR_L}) + (Y' * Y_{HRIR_L}) \tag{6.38}$$

where the HRIRs are essentially encoded into spherical harmonic components as follows:

$$W_{HRIR_L} = \frac{1}{\sqrt{2}} \sum_{i=1}^{8} h_l(\theta_i) \tag{6.39}$$

$$X_{HRIR_L} = \sum_{i=1}^{8} \cos\theta_i h_l(\theta_i) \tag{6.40}$$

$$Y_{HRIR_L} = \sum_{i=1}^{8} \sin\theta_i h_l(\theta_i) \tag{6.41}$$

Thus, any number of HRIRs can be represented by a set of Ambisonic channels of a given order. If the factorisation technique is applied then Equations 6.39 to 6.41 become

$$W_{HRIR_L} = \frac{1}{\sqrt{2}} \sum_{i=1}^{8} g_l(\theta_i) \tag{6.42}$$

$$X_{HRIR_L} = \sum_{i=1}^{8} \cos\theta_i g_l(\theta_i) \tag{6.43}$$

$$Y_{HRIR_L} = \sum_{i=1}^{8} \sin\theta_i g_l(\theta_i) \tag{6.44}$$

where $g_l(\theta_i)$ is the direction dependent component at a position θ_i resulting from the factorisation process. Again the right ear components are calculated in the same manner. The direction independent component must also be incorporated and the overall system output for the left channel becomes

$$\text{Left} = S * f * [(W * W_{HRIR_L}) + (X' * X_{HRIR_L}) + (Y' * Y_{HRIR_L})] \tag{6.45}$$

The next consideration is the rotation transformations to be applied to the velocity channels. For the horizontal only case with rotation about the z-axis these transforms are only applicable to the velocity components, X and Y, and from Equation 6.29 it follows

$$\begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} \tag{6.46}$$

where θ is the clockwise angular deviation of the soundfield from straight ahead. The rotation angle will be calculated as a result of an input from the InertiaCube head tracker unit. If the listener head rotates by an angle γ then it is necessary for the soundfield to rotate by $-\gamma$ in order for the sound source to be maintained at the same spatial location. The left channel can be rewritten as

$$\text{Left} = S * f * [(W * W_{HRIR_L}) \tag{6.47}$$

$$+ ([X\cos\theta - Y\sin\theta] * X_{HRIR_L}) \tag{6.48}$$

$$+ ([X\sin\theta + Y\cos\theta] * Y_{HRIR_L})] \tag{6.49}$$

which, due to the distributivity property, is equivalent to

$$\text{Left} = ([S * f * W] * W_{HRIR_L}) \tag{6.50}$$

$$+ ([[S * f * X]\cos\theta - [S * f * Y]\sin\theta] * X_{HRIR_L}) \tag{6.51}$$

$$+ ([[S * f * X]\sin\theta + [S * f * Y]\cos\theta] * Y_{HRIR_L}) \tag{6.52}$$

As can be seen the source audio and f component are combined with the W, X and Y B-format components offline to form the 3 base inputs to the system. The X and Y channels are

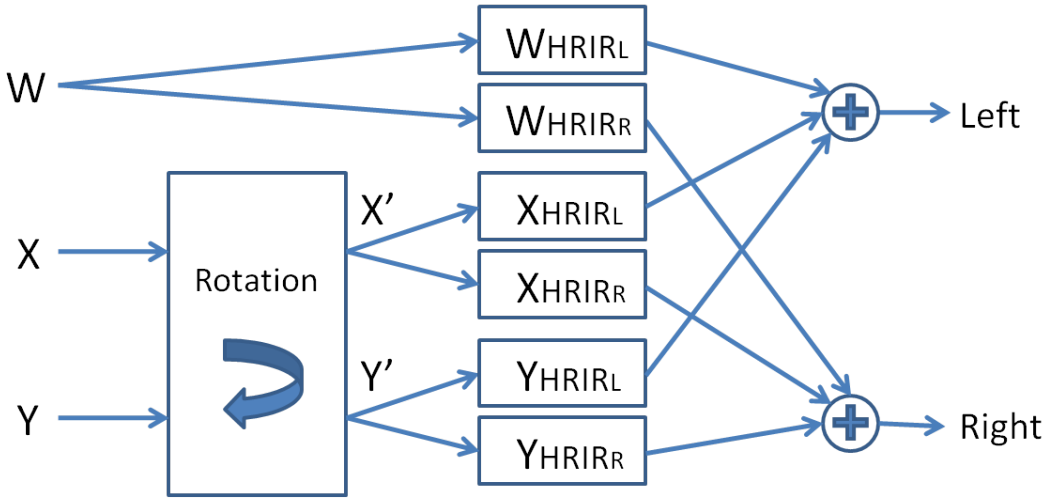


Figure 6.13: Horizontal only VLA implementation.

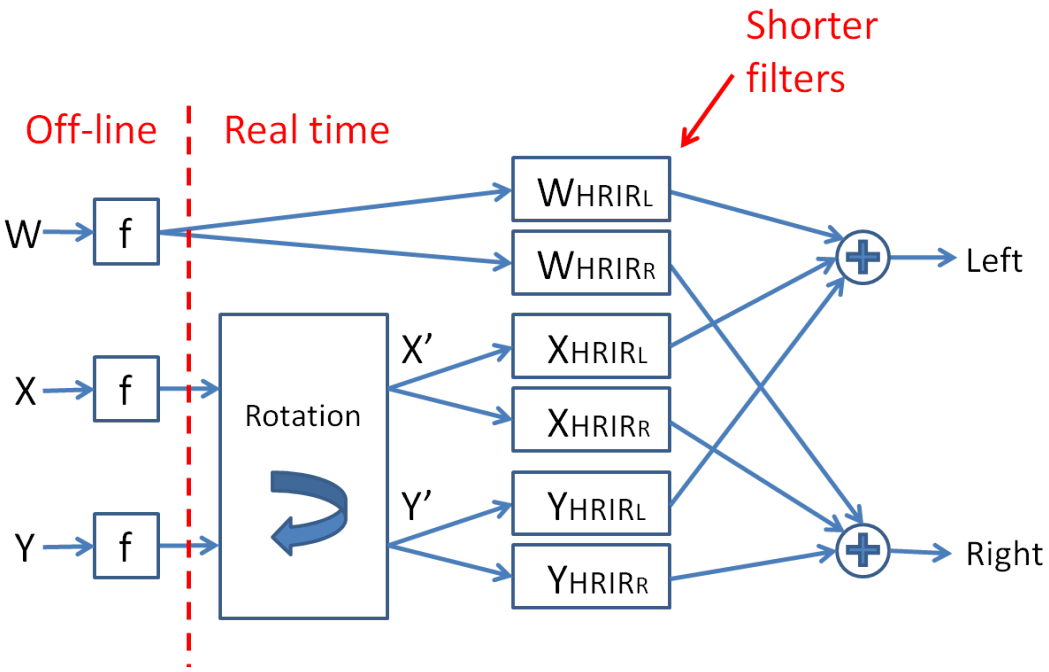


Figure 6.14: Incorporation of factorisation into VLA implementation.

linearly combined with weightings based on user dependent positional information before being convolved with the HRIR spherical harmonic components. Figures 6.13 and 6.14 outline the system in block diagram form with and without factorisation.

The above is applied to real HRIR data from the IRCAM database [83]. This database consists of HRIRs measured at 187 positions on a measurement sphere of radius 1.95 m for 51 subjects. It includes both raw measured HRIRs and their diffuse field equalised (recall

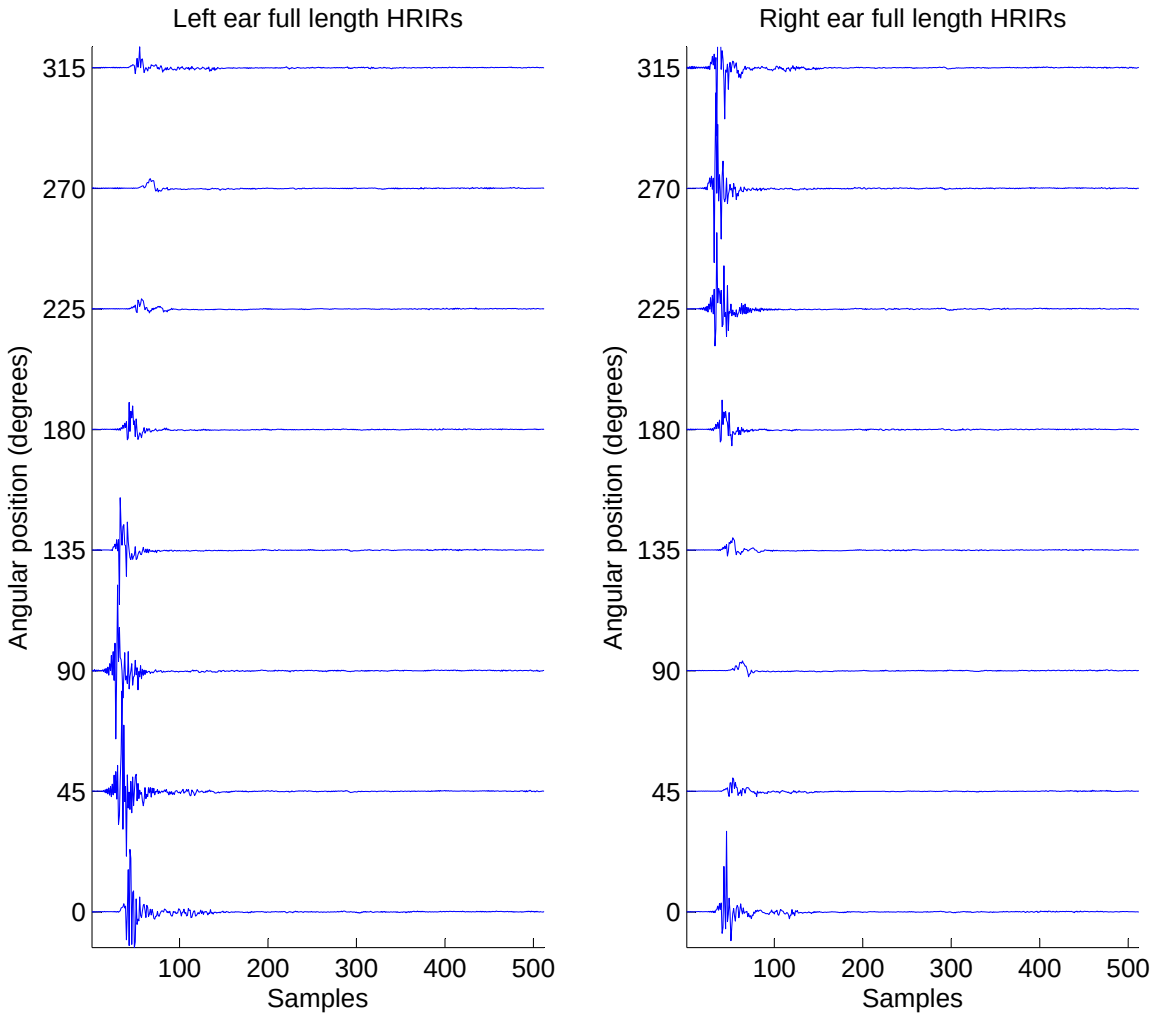


Figure 6.15: HRIRs for each ear for 8 loudspeaker positions.

Section 3.2.2) counterparts. Each HRIR is 512 samples long at 44.1kHz. In this implementation the diffuse field equalised data for Subject 1021 is used. Figure 6.15 shows the left and right ear HRIRs used. The HRIRs are in the horizontal plane and spaced at 45° intervals i.e. $[0°, 45°, 90°, \ldots 315°]$. The 16 HRIRs (8 for each ear) are factorised in order to extract a direction independent component, f. The factorisation process uses regularisation on the g^{ϕ} components and a 450 sample long f component is extracted. Figure 6.16 shows the 63 tap long direction dependent components for each ear while Figure 6.17 shows the direction independent component. Figure 6.18 shows the absolute difference between the original 16 HRIRs and the HRIRs reformed by convolving together the factorised components. The difference is minimal between the two and it is clear that the factorisation has been successful.

The next stage is the encoding of the direction dependent components into spherical harmonic components as dictated by Equations 6.42-6.44. These components can be seen in Figure 6.19.

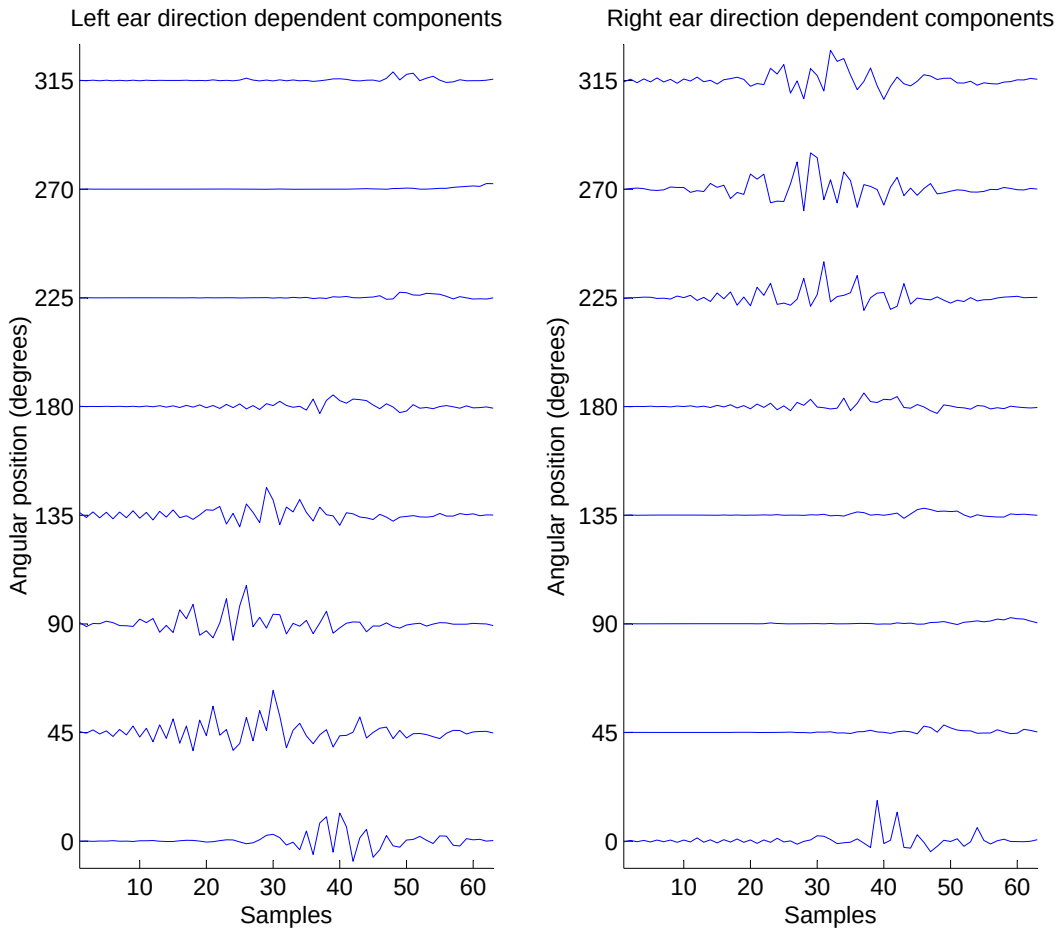


Figure 6.16: Direction dependent components for each ear for 8 loudspeaker positions.

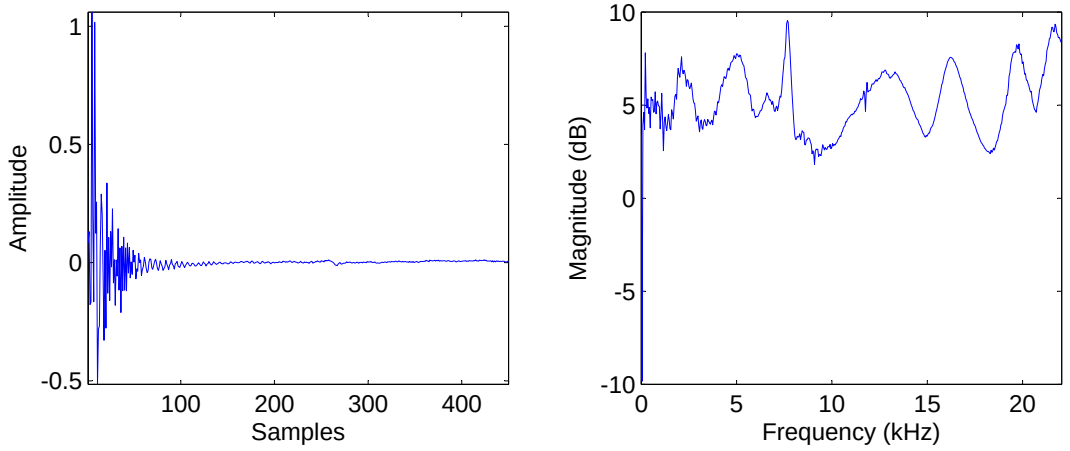


Figure 6.17: Direction independent component in the time and frequency domain.

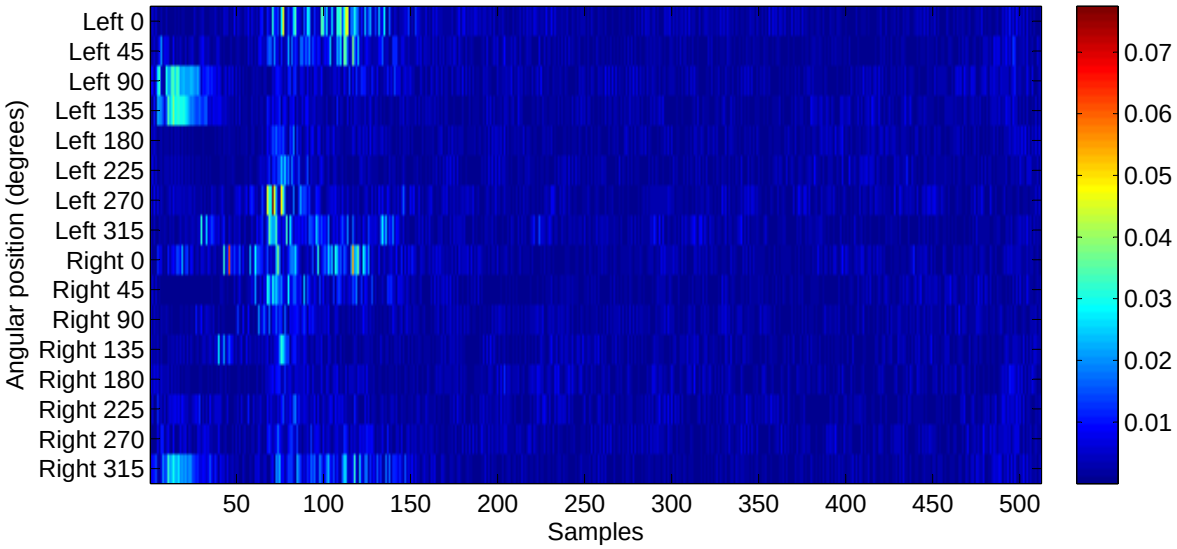


Figure 6.18: Difference between original HRIRs and HRIRs reformed after factorisation.

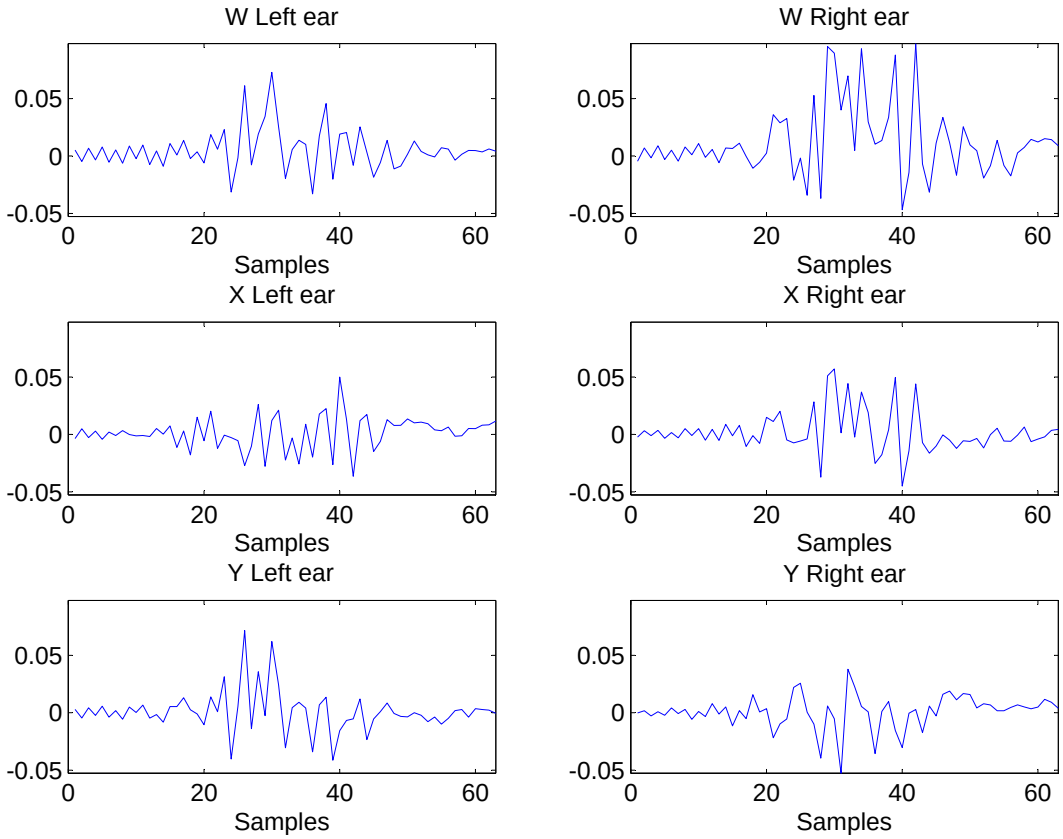


Figure 6.19: W_{HRIR}, X_{HRIR} and Y_{HRIR} for each ear.

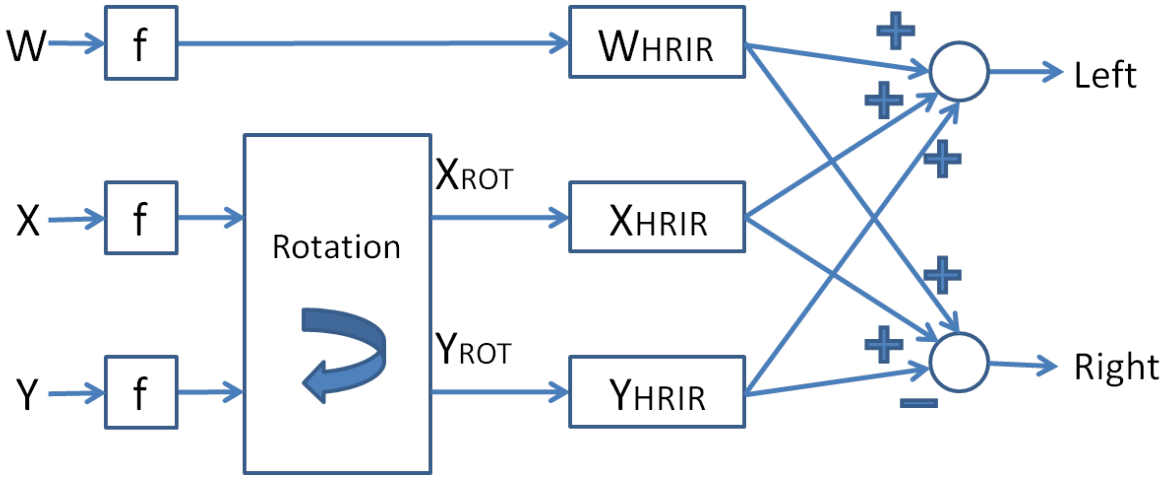


Figure 6.20: Virtual loudspeaker approach asumming left/right symmetry as defined by McKeag [116].

This results in 6 convolutions (3 for each ear). If left/right symmetry is assumed in the system as in McKeag's paper [116] then only 3 convolutions are required as shown in Figure 6.20. If the left ear HRIR data is used, McKeag's approach requires negating the Y channel before the summation for the right ear feed. By including both the left and right ear data instead of using this assumption it is hoped that the auralisation will be more convincing, especially if the subjects own HRIRs are being used.

6.4.1 Fmod

The Fmod Ex API is an audio engine made by Firelight Technologies for the creation and manipulation of real-time interactive audio. The creators allow for the library to be freely downloaded by users who will not profit commercially through its use. Fmod products are widely used in the games industry and support a large number of platforms (eg. Windows, Mac, Linux iPhone, Xbox, Playstation 3, Nintendo Wii) and audio formats.

The C++ API is used in this study. A standard 2.2GHz dual core PC with 3.5GB of RAM running Windows XP is used for the implementation. The system is initialised as follows:

```
FMOD::System          *system;
FMOD::Channel         *channel[2]; channel[0] = 0; channel[1] = 0;
FMOD_RESULT            result;
unsigned int           version;
int                    channelsplaying = 0;
FMOD_CREATESOUNDEXINFO exinfo;

// Create a System object and initialize.
result = FMOD::System_Create(&system);
ERRCHECK(result);
```

```
result = system->getVersion(&version);
ERRCHECK(result);

if (version < FMOD_VERSION){
    printf("Error! You are using an old version of FMOD %08x. This program
        requires %08x\n", version, FMOD_VERSION);
    exit(0);
}
result =system->setSoftwareFormat(44100, FMOD_SOUND_FORMAT_PCM16, 0, 0,
    FMOD_DSP_RESAMPLER_LINEAR);
ERRCHECK(result);
result = system->init(32, FMOD_INIT_NORMAL, 0); ERRCHECK(result);
```

Fmod allows for user defined custom DSP units to be created (as shown in the code fragment below) thus supporting the implementation of the sequence of filters described in Section 6.4.

```
FMOD::DSP  *mydsp
{
      FMOD_DSP_DESCRIPTION  dspdesc;
      memset(&dspdesc, 0, sizeof(FMOD_DSP_DESCRIPTION));
      strcpy(dspdesc.name, "L");
      dspdesc.channels = 1;
      dspdesc.read = myDSPCallback;
      dspdesc.userdata = (void *)Luserdata;
      result = system->createDSP(&dspdesc, &mydsp);
      ERRCHECK(result);
  }
```

Such custom DSP units are accessed via a callback from the main program. The mono source audio which has been pre-convolved with the direction independent component is passed in through the input buffer of the callback to the custom DSP unit. The callback is called twice in each cycle of the program, once for each ear. A number of other pieces of information are passed into the callback through the userdata pointer. These include a flag to indicate if the left or right ear feed is being created, an angle used to control the rotation to be applied to the soundfield and a pointer to an overlap buffer and are shown in the code fragment below.

```
float        *Luserdata[5];
Luserdata[0] = new float[1]; Luserdata[1] = new float[1];
Luserdata[2] = new float[1]; Luserdata[3] = new float[1];
Luserdata[4] = new float[1];
Luserdata[0] = &AngleValue_az;// azimuth angle
Luserdata[1] = &AngleValue_el;// elevation angle (not needed for horz only case)
Luserdata[2] = overlapbufferL;// overlap buffer for left channel convolution
Luserdata[3] = &lflag;        // leftflag
Luserdata[4] = &pageidentifier;
```

The overlap buffer stores the overlap of the convolution at each iteration so it can added on at the beginning of the next cycle to prevent discontinuities in the audio feed. The code for the

C++ callback is shown in Appendix A for the horizontal only case discussed previously. This allows for head rotation for the listener.

In order to allow the listener full freedom in the virtual environment it would be necessary to switch the W, X and Y components containing the RIR information as the listener moves though the room. A discrete grid of measurements in the rooms to be synthesised would be taken and the responses would be switched between depending on user proximity to particular grid points. Care would need to be taken to avoid any distortion or artifacts in the playback resulting from such switching.

6.5 Conclusion

The modular composition of BRIRs and their applicability to real time binaural auralisation has been explored in this chapter. A combination of preprocessing of HRIR data based on regularised factorisation and real-time filtering is used to implement an efficient system which allows for head movement of the listener. The application of RIR interpolation to this problem is also explored to enable a listener to move around the artificial listening environment. A real time implementation using the Fmod Ex API is detailed which includes rotation of the soundfield in response to head movements by the listener. The factorisation process described in Chapter 4 is applied to the HRIR data and this results in shorter filters for use in the real time process. This implementation is motivated by the need of immersive and convincing spatial audio in e-learning and gaming scenarios. Externalisation of the virtual sound source from the listener's head will reduce the dissociation between the visual and audio cues that are presented.

7

Application to E-learning

7.1 Application to E-learning

Consider the situation where educational information is being presented to a subject through visual and audio media in a classroom environment, the visual information being presented on a monitor in front of the subject and the audio information being presented over headphones. It is desirable that a stable audio image be created at the location of the visual source in order to establish a truly immersive learning experience. However there can be a significant disassociation between visual and audio cues due to the head movements of the listener. The use of external stationary loudspeakers as the spatial audio reproduction medium would remove this issue but this is of course not applicable to a classroom type situation as it would cause disruption to other participants. The purpose of this chapter is to discuss the benefits of a spatial audio in e-learning applications while also providing a method for the removal of this dissociative distraction from headphone listening.

7.2 Applicability of Spatial Audio to E-learning

The majority of current e-learning systems place huge emphasis on the visual while largely neglecting and underestimating the importance of audio. However visual and audio information when presented in the real world are complimentary and intrinsically linked media and should be presented as such in virtual environments. The role of spatial audio in interactive environments such as e-learning and desktop conferencing has been explored by several authors. Baldis [9]

completed a comprehensive examination of the effect of spatial audio on participant memory, comprehension and focal assurance for desktop conferences. Whilst the spatial audio apparatus in this experiment was very basic, essentially 4 loudspeakers placed directly above a monitor with 4 images of speakers displayed, this paper proposes that memory and comprehension of the conference material is improved by the use of spatial audio when compared with non spatial audio. Ericson and McKinley [48] show findings indicating that spatial audio increases speech intelligibility in noisy environments. These findings would certainly be advantageous in an e-learning application, especially if one considers the participant using the application in a crowded classroom environment with a propensity for significant background noise.

While at this initial development stage of this application there is only one speaker/sound source, in further, more complex, e-learning scenarios it is conceivable that there may be multiple sources. Several authors have found that spatialisation of such individual sound sources to different positions increases their intelligibility to the user (e.g. Drullman and Bronkhorst [45]). Another interesting finding in Drullman and Bronkhorst's paper is that they find no performance difference when comparing the use of individual and non individual HRIRs in this situation. Ericson et al. [47] use a intelligibility measure called the coordinate response measure to determine the usefulness of spatial audio, their focus being on its applications in aviation. They find an increase of 25 to 35% in performance when competing speakers are spatially separated. They highlight the fact that this spatial separation improves the intelligibility of all the speakers as opposed to the use of other techniques which only improve the intelligibility of one of the speakers. It is hoped in this application that a spatial audio environment that responses interactively to a users movement and offers good externalisation of the virtual sound source will aid in cementing the participants concentration to the screen.

7.3 ReciTell

The particular e-learning situation being dealt with here is a primary school level reading book which has been ported to an interactive, flash or java based, player named the ReciTell Player (see Figure 7.1). Darren Kavanagh[1] has developed an innovative system for segmenting and synchronising the spoken words with their corresponding visual text counterpart. As such, the player offers an efficient means of creating visualised/highlighted text that is displayed in synchronism with a natural reading of teaching material. The use of synthetic speech would trivialise the synchronisation problem. However, the use of real speech is lauded by Kavanagh as key to optimising the engagement level of the user. He claims that the natural accent, intonation and prosody of a real narration cannot be imparted using synthetic speech. An overview of the basic operation of the system is shown in Figure 7.2. While this product has been initially piloted for primary schools it also has significant applicability to the 'English as a

[1]Darren Kavanagh is a PhD student preparing his thesis on speech segmentation with applications to e-learning and has a patent pending on his speech/video synchronisation technique.

Second Language' market.

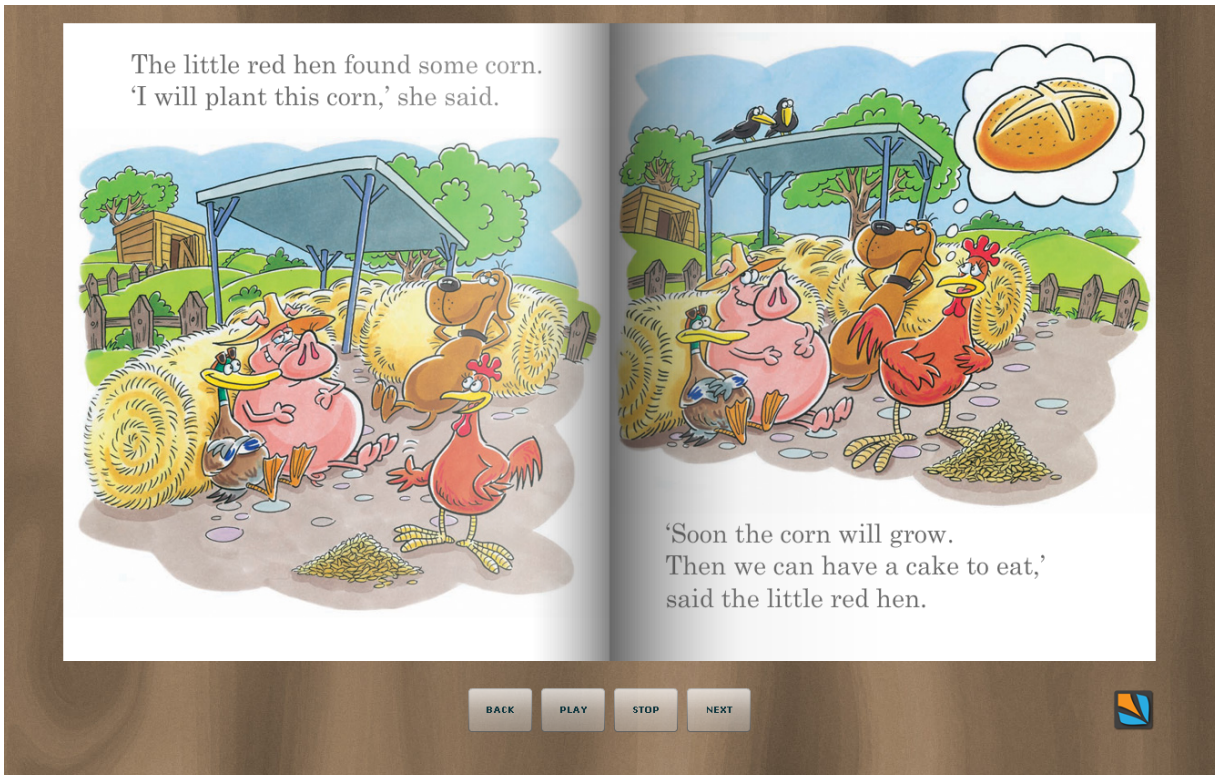


Figure 7.1: Example of ReciTell in operation.

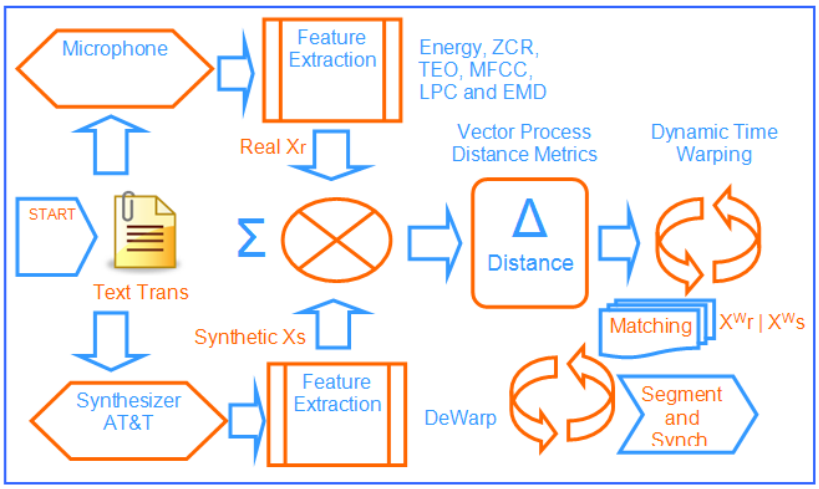


Figure 7.2: Overview of operation of ReciTell system.

7.4 Implementation

The use of a head tracking apparatus and the binaural virtual loudspeaker approach as outlined in the previous chapter offers a practical solution to the problem of adjusting the virtual source position with user head movement. Fmod offers the necessary real-time filtering capabilities while the inertia based head tracker provides accurate tracking and free movement to the listener. As previously discussed the factorisation of HRIRs is directly applicable to this form of Ambisonic based implementation and would allow for increased efficiency. The inclusion of realistic B-format RIRs and the dynamically rotating soundfield based on head position should help provide externalisation of the sound image and increase the connection between the visual and audio information.

A suitable B-format RIR is needed to impose a realistic reverberance on the reproduced signal. Hence it is necessary to measure a B-format RIR in a classroom sized room with a desk between the loudspeaker and soundfield microphone. The measurement setup is shown below in Figure 7.3. The measurement took place in a medium sized classroom (approximately 7m x 5.4m x 2.4m) in Trinity College Dublin. Similarly to the RIR capture discussed in Chapter 5, a Genelec 1029A loudspeaker was used to play the swept sine excitation signal and the Soundfield MK5 system was used to capture the B-format response. Each B-format component of the RIR is shown in Figure 7.4.

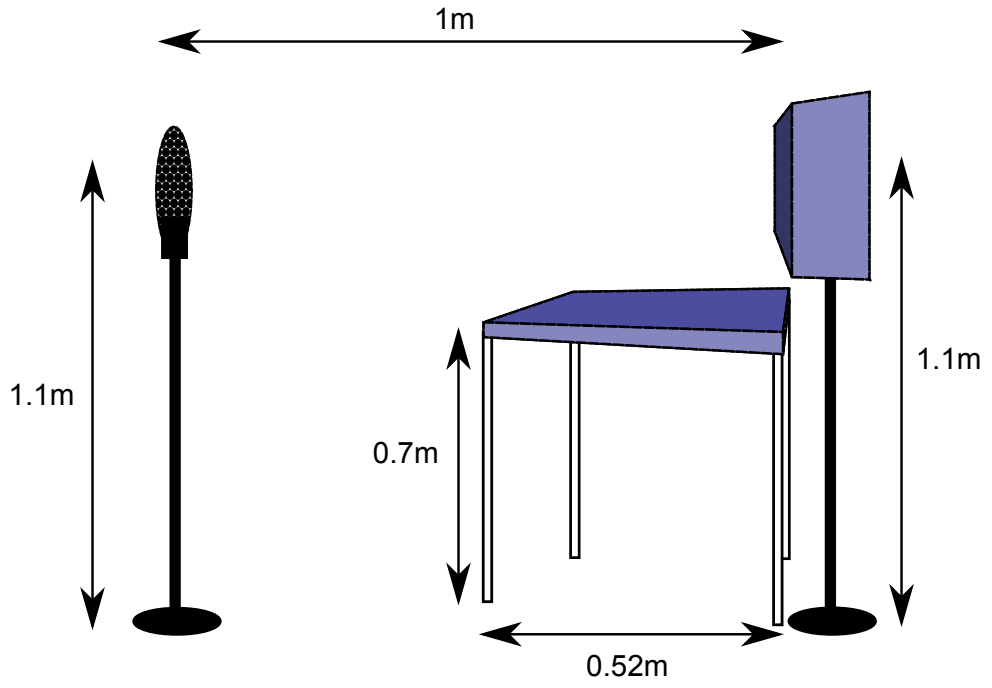


Figure 7.3: Measurement Apparatus.

The HRIRs used to position the eight virtual loudspeakers in this implementation are from the CIPIC database. This database offers the smallest radius measurement sphere (at 1m) of

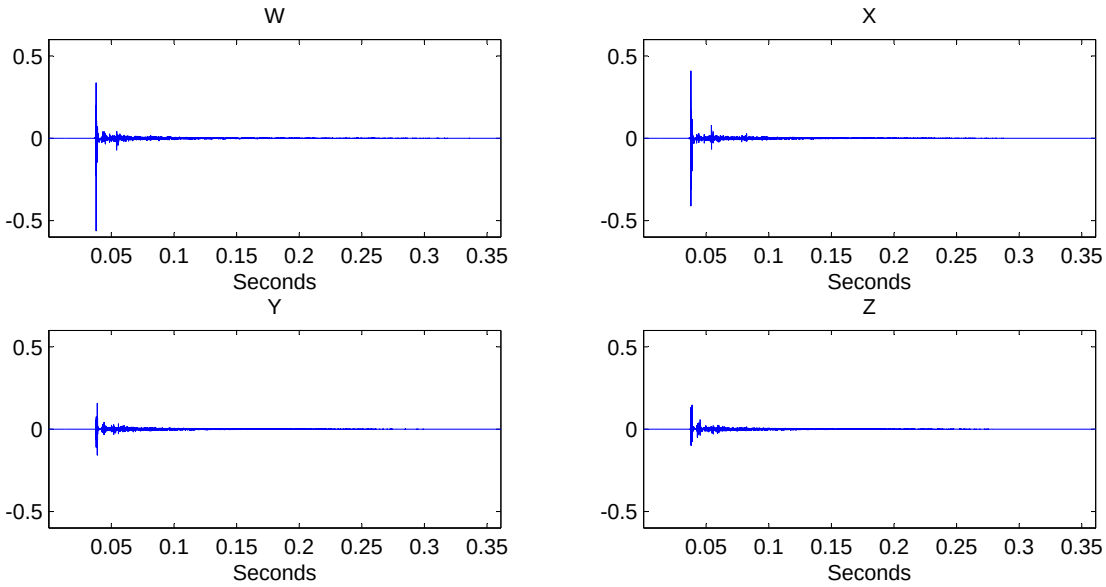


Figure 7.4: B-format Classroom RIR.

all the freely available databases which makes it the most suitable for this application as the proposed listener is unlikely to be sitting any further than this from the screen. It also has a large dataset of subjects along with anthropometric data which would allow for future work in the assignment of a set of 'best fit' HRIRs based on user inputted measurements or images of the pinna. The use of non individual HRIR data has been shown to reduce localisation accuracy and externalisation by some authors (refer to Section 3.2.1). However in this case the audio data being presented is speech and as such is largely limited to being lower than 3.5kHz in frequency. The highly individual characteristics in HRIRs caused by the fine structure of the pinna are generally present at frequencies above 3.5kHz (refer to Section 3.1 for further detail on this) and as such, speech frequencies remain mostly unaffected. The bandlimited nature of the signal, coupled with the head tracking controlled soundfield rotation and added room reverberant information, should allow for externalised, localisable audio. Factorisation is applied to the 16 HRIRs to extract a common component and leave significantly shorter filters for use in real-time convolution. These filters are then encoded into B-format as described in the previous chapter and the common component is convolved with the source audio. The real-time convolution and rotation of the soundfield is done in the Fmod API.

7.5 Conclusion

The applicability of the virtual loudspeaker approach introduced in Chapter 6 to a particular e-learning implementation has been explored in this chapter. The Recitell e-learning system allows for the presentation of synchronised audio and visualised text. The inclusion of spatial audio

based on the virtual loudspeaker approach, dynamically reacting in real time to head tracking data, is motivated by a need for an externalised sound source and is aimed at improving the immersivity of the environment.

There is a considerable breath of work that could follow on from this point. It would be desirable to measure sets of HRIR data with the sound source at varying distances from the head and for several subjects to allow more flexibility in reproduction. Whilst the CIPIC database offers a 1m distance from the measurement position to the head it is feasible that some people would sit closer to the monitor than this and would require alternative HRIR data. Recall also that for a 3D layout the loudspeakers should ideally be placed at the vertices of Platonic solids. These vertices, in most instances, do not occur on the regular (often 5° spaced) grids that the available HRIR datasets offer. A study of the effect of audio spatialised in this manner in comparison with monophonic audio on concentration levels and comprehension could yield interesting results. A gaze tracker is a possible quantifiable technique that could be used to measure the attentiveness of the user to the material. Alternatively questionnaires could be used to examine the user's comprehension, memory and preferences.

8 Conclusion

In the Introduction the aim of this thesis was stated to be

> "to make the auralisation of audio more efficient, both in its real time computational and memory requirements and in the preparation time required to acquire necessary data and measurements."

The following novel features have been developed over the course of this thesis to achieve these aims:

- A least squares based algorithm for the approximate factorisation and order reduction of large HRIR datasets.
- A refinement of this algorithm to include two regularisation techniques, one applicable to minimum phase data, the other to full, ITD inclusive HRIRs.
- A technique for the spatial interpolation of the early reflection component of room impulse responses using Dynamic Time Warping (DTW).
- The development of a real time implementation of the Virtual Loudspeaker Approach with head tracking using the Fmod API and the application of this to a particular e-learning scenario.

The thesis has been structured in such a way as to provide the reader with a brief introduction to the necessary fundamentals before delving into the more novel attributes of the work. Chapter 2 served to introduce the reader to the general area by providing the necessary psychoacoustic

background information pertaining to how humans process sound. The nature of localisation cues, the neurological processing involved and the spatial resolution they offer were examined. The importance of the listener head movements was recognised following an examination of existing literature and this motivated the use of head tracking data in a real time implementation in Chapter 6. The importance of the influence of the environment on the sound that reaches the human ear, along with measurement and modelling techniques to capture this filtering effect, were also discussed as a necessary precursor to Chapter 5.

The HRIR, as introduced in Chapter 3, has been a central concept in this thesis. Chapter 3 outlined the physiological reasons for the composition of the HRIR as well as techniques for its capture and synthesis. Topics that frequent the existing literature on the subject such as the minimum phase assumption and commonly used equalisation strategies were explained, the minimum phase explanation being of particular relevance for Chapter 4. Various existing HRIR reduction techniques were discussed.

An approximate factorisation algorithm, implemented as a deconvolution operation, was introduced in Chapter 4 which addressed improving the real time implementational efficiency of the use of HRIRs in real time interactive systems. The capability to extract a significantly lengthy direction independent component from a large dataset of HRIRs while still maintaining much of the fine detail of each HRIR has considerable benefits from a memory conservation and real time processing viewpoint. The extension of this algorithm to offer two regularised factorisation techniques allows the user choice between the maintenance of the non minimum phase properties of the HRIR versus minimal direction dependent filter length. There is an important trade off here and it is necessary to refer back to the minimum phase discussion in Section 3.3 to appreciate this. While the minimum phase assumption is widely used in literature there is some evidence that HRIRs often exhibit non minimum phase behaviour. This coupled with the uncertainty on how ITD is detected in the brain presents an important argument against the automatic use of the assumption. The use of the minimum phase assumption has been shown to allow for a significantly longer common component to be factorised from the dataset but requires the separate storage of the initial delay. The benefit of the two separate variations of the factorisation algorithm, allowing for either the inclusion or exclusion of the minimum phase assumption, is clear in this light.

Head related filtering is only one component of the system required to create a convincing virtual auditory environment. As discussed in Chapter 5 the influence of the room must also be considered and included for maximal immersivity. For an interactive walk-through situation a grid of RIR measurements spatially distributed across the room is needed. The RIR interpolation technique described in Chapter 5 reduces the number of measurements that must be taken. DTW based early reflection interpolation was used in conjunction with synthesis of the diffuse tail. The technique showed good results, even when only a small number of measured responses are used and the interpolation is performed over a significant distance. Another sound reproduction technique that this method is useful for is wave field synthesis. Both objective

and perceptual tests were carried out on a wave field synthesis rig comparing implementations where no interpolation and different levels of interpolation of RIRs were used. No significant degradation in localisation quality was found to be brought about by interpolation.

Chapter 6 described the implementation of a real time virtual auditory system which is responsive to head tracking input. The real-time convolution and rotation of the soundfield was executed in the Fmod API on a standard PC (2.2GHz Dual Core, 3.5GB RAM, Win XP). The application of the factorisation technique resulted in significantly shorter filters for use in the system. A 450 sample long direction independent component was extracted from the set of sixteen 512 sample long HRIRs used leaving 63 sample long filters for real time convolution, with the direction independent component being convolved with the source audio offline. The Ambisonics based virtual loudspeaker implementation allowed for simple rotation of the soundfield in response to head movement, negating the need for constant HRIR switching and avoiding the noise artifacts that may result from this. The implementation described in Chapter 6 is only in the horizontal plane but it is possible to extend this to three dimensions. This will be explored in the next section on future work.

Chapter 7 described the application of the implementation detailed in Chapter 6 to the Recitell e-learning system. A review of literature in the area found that the use of spatial audio in a teleconferencing environment improved comprehension and memory of the material. This motivated its use in ReciTell system.

8.1 Future Work

The Fmod implementation of the virtual loudspeaker approach described in Chapter 6 is only in the horizontal plane. A full 3D implementation of the virtual loudspeaker approach would be desirable to further improve the immersivity of the system. As Ambisonics decoding requires regularity from the reproduction loudspeaker array, while also imposing a lower bound on the number of loudspeakers to be used, there is a very limited number of layouts that could be used. Theoretically only the five platonic solids satisfy this regularity constraint. The octahedron and the icosahedron are the only two of these that offer a ring of loudspeakers in the horizontal plane which is preferable from a psychoacoustic viewpoint. While first order Ambisonics was used in this thesis, one should also consider the use of higher order Ambisonics to improve localisation quality. The use of higher order Ambisonics places further restrictions on the speaker layout. This necessitates the examination of other semi regular layouts as the platonic solids are no longer adequate. The method of geodesic spheres is one technique which allows for the creation semi-regular 3D shapes with higher numbers of vertices. It involves creating new vertices on existing platonic polyhedra by subdividing each face and projecting the new vertices out onto the polyhedrons circumscribing sphere. Further information on this can be found in the work of Daniel [41] and Hollerweger [79].

Larger HRTF sets with spatial measurements available at the vertices of these geodesic

semi-regular solids would be required to implement this system. Existing freely available HRTF databases are generally only measured over a fairly coarse spatial grid so synthesis or spatial interpolation of responses would be necessary. Another issue with these databases is that the measurement sphere radius is often quite large (1.5m to 2m). As it is unlikely that a user would be sitting more than 1m from the monitor in an e-learning application the HRTF measurement sphere radius should reflect this. There is also an absence of HRTF data for children, whom are the main target audience for the Recitell product. Obtaining such data by measurement would be problematic due to the understandable reluctance of parents to allow devices to be placed into their childrens' ears as well as the difficultly inherent in keeping children still for the sustained periods of time required for such measurements. It would be preferable to examine modelling techniques such as those investigated in Section 3.2 as an alternative.

In order to examine the benefits offered by implementing spatialised audio using the virtual loudspeaker approach in this context, a study should be undertaken investigating comprehension and concentration levels, using simple monophonic reproduction as a control. One possible strategy for measuring attentiveness in a quantifiable way is through the use of gaze tracking technology. Alternatively, or additionally, more subjective means of assessment such as questionnaires could be used to assess the comprehension and preferences of the user.

While in this thesis the presentation system is applied to a particular e-learning application there is significant opportunity in other areas such as personal music consumption and gaming. There is also potential to explore liaising with headphone manufacturers with a view to incorporating a head tracking device and possibly even a DSP device to headphone sets. Recent commercial products such as Sennheiser's 7.1 surround sound PC 333D headphones, Beyerdynamic's Headzone Professional 5.1 Monitoring Headphone System and Smyth Research's Realiser A8 [165] indicate that the importance and potential of spatial audio reproduction over headphones is finally being realised in industry. Beyerdynamic claim their system to be the first headphone system that to deliver virtual 5.1 surround sound and use a ultrasonic headtracking system. While HRTF data is used to position the virtual loudspeakers, it is unclear what processing is used for real time adaptation to head movement. There is no mention of an Ambisonics based approach. However, the nature of the ultrasonic headtracking may not be suitable in a multi-user environment such as a classroom. Also the list price is in the region of €2000 which would preclude the average person/educational centre from purchasing it. This motivates a low cost alternative, which the work presented in this thesis is directed towards providing.

The availability of cheap head tracking apparatus is another major driver of this work. In this work an inertial unit is used as it was easily available to the author. However, there are numerous other options depending on the particular application and its monetary restrictions. Camera based systems are in existence where face tracking technology is used to determine head position. With the availability of cheap webcam type devices such as the Playstation Eye with frame rates up to 120Hz this is an attractive option. There is also a number of infrared systems available including commercial products and DIY systems. One such homemade system

is based on the use of a wiimote and cheap IR LEDs. While these systems require line of sight in order to function correctly they may be suitable for applications where there is reasonably constrained head movement. A full examination and comparison of these techniques based on their price, flexibility, reliability of performance and appropriateness for various applications should be undertaken.

8.2 Final Remarks

Technology for the reproduction of visual information has been advancing rapidly in recent decades as high definition, 3D video has become a commercial reality. The quality of the spatial auralisation of the accompanying audio reproductions has not advanced at a similar pace. Typical headphone presentations result in disjointed audio-visual presentations as the audio fails to 'externalise' outside the head. This thesis has examined techniques of implementing interactive, high quality, immersive spatial sound with particular emphasis on binaural reproduction.

A
Appendix

```
FMOD_RESULT F_CALLBACK myDSPCallback(FMOD_DSP_STATE *dsp_state, float *inbuffer,
    float *outbuffer, unsigned int length, int inchannels, int outchannels){

char name[256];
FMOD::DSP *thisdsp = (FMOD::DSP *)dsp_state->instance;
thisdsp->getInfo(name, 0, 0, 0, 0);
void *number;
thisdsp->getUserData(&number);   //get the hrir and the overlap data from userdata

float** ptr2ptrs = static_cast<float**>(number);
float* angle_azm=ptr2ptrs[0];
float* angle_el=ptr2ptrs[1];
float* overlapbuffer=ptr2ptrs[2];
float* leftflag=ptr2ptrs[3];
float* identifier=ptr2ptrs[4];

//////////////////   ROTATE SOUNDFIELD   ///////////////////////
//negate angle as rotation of head by x degrees requires soundfield rotation
//in the oppositie direction to maintain a stable source i.e. -x degrees.
float cos_angle_azm=cos(-.017453*(*angle_azm));
float sin_angle_azm=sin(-.017453*(*angle_azm));

//Rotate X,Y
float Xrot[buffsize],Yrot[buffsize],Zrot[buffsize];
for (int i=0; i<buffsize; i++){
```

```
    Xrot[i] = (buffXf[placeholder+i]*cos_angle_azm) - (buffYf[placeholder+i]*
        sin_angle_azm);
    Yrot[i] = (buffXf[placeholder+i]*sin_angle_azm) + (buffYf[placeholder+i]*
        cos_angle_azm);
}

////////////////////////  LEFT EAR  ////////////////////////////
if(*leftflag==1.0){
  //W component convolution
  for (unsigned int n=0; n<(length+hrir_leng-1); n++){
    float sum=0;
    for (int m=0; m< hrir_leng; m++){
      if (((n-m)>=0) && ((n-m)<length)){
        sum=sum + (Whrirl[m]*buffWf[placeholder+n-m]);  //main convolution formula
      }
    }
    W[n] = sum;
  }

  //X component convolution
  for (unsigned int n=0; n<(length+hrir_leng-1); n++){
    float sum=0;
    for (int m=0; m< hrir_leng; m++){
      if (((n-m)>=0) && ((n-m)<length)){
        sum=sum + (Xhrirl[m]*Xrot[n-m]);
      }
    }
    X[n] = sum;
  }

  //Y component convolution
  for (unsigned int n=0; n<(length+hrir_leng-1); n++){
    float sum=0;
    for (int m=0; m< hrir_leng; m++){
      if (((n-m)>=0) && ((n-m)<length)){
        sum=sum + (Yhrirl[m]*Yrot[n-m]);
      }
    }
    Y[n] = sum;
  }

  //PUT it all together and order it in the outbuffer
  for (unsigned int n=0; n<(length+hrir_leng-1); n++){
    float sum=0;
    if(n<length){  //Get outbuffer values
        if (n<lengoverlap){
          sum = W[n] + X[n] + Y[n] + overlapbuffer[n];
          outbuffer[n] = sum ;
```

```
        }else{
          sum = W[n] + X[n] + Y[n] ;
          outbuffer[n] = sum ;
        }

    }else{ //Put overlap in overlapbuffer for next DSP callback
      overlapbuffer[n-length] = W[n] + X[n] + Y[n] ;
    }
  }

////////////////////////  RIGHT EAR  ////////////////////////////
}else if(*leftflag==-1.0){
  //W component convolution
  for (unsigned int n=0; n<(length+hrir_leng-1); n++){
    float sum=0;
    for (int m=0; m< hrir_leng; m++){
      if (((n-m)>=0) && ((n-m)<length)){
        sum=sum + (Whrirr[m]*buffWf[placeholder+n-m]);
      }
    }
   W[n] = sum;
  }

  //X component convolution
  for (unsigned int n=0; n<(length+hrir_leng-1); n++){
    float sum=0;
    for (int m=0; m< hrir_leng; m++){
      if (((n-m)>=0) && ((n-m)<length)){
        sum=sum + (Xhrirr[m]*Xrot[n-m]);
      }
    }
    X[n] = sum;
  }

  //Y component convolution
  for (unsigned int n=0; n<(length+hrir_leng-1); n++){
    float sum=0;
    for (int m=0; m< hrir_leng; m++){
      if (((n-m)>=0) && ((n-m)<length)){
        sum=sum+ (Yhrirr[m]*Yrot[n-m]);
      }
    }
    Y[n] = sum;
  }

  //PUT it all together and order it in the outbuffer
  for (unsigned int n=0; n<(length+hrir_leng-1); n++){
    float sum=0;
```

```
    if(n<length){  //Get outbuffer values
      if (n<lengoverlap){
        sum = W[n] + X[n] + Y[n]  +overlapbuffer[n];
        outbuffer[n] = sum ;
      }else{
        sum = W[n] + X[n] + Y[n] ;
        outbuffer[n] = sum ;
      }
    }else{ //Put overlap in overlapbuffer for next DSP callback
      overlapbuffer[n-length] = W[n] + X[n] + Y[n];
    }
  }
  placeholder=placeholder + buffsize;

}else{
  cout<<"Neither left or right! Somethings wrong! \n";
}

return FMOD_OK;
}
```

Bibliography

[1] N. Adams and G. Wakefield. State-space synthesis of virtual auditory space. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(5):881–890, July 2008.

[2] W. Ahnert and R. Feistel. Ears auralization software. In *Audio Engineering Society Convention 93*, October 1992.

[3] T. Ajdler. *The plenacoustic function and its applications.* PhD thesis, École Polytechnique Fédérale de Lausanne, 2006.

[4] T. Ajdler, L. Sbaiz, and M. Vetterli. Plenacoustic function on the circle with application to HRTF interpolation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[5] V. Algazi, C. Avendano, and R. Duda. Elevation localization and head-related transfer function analysis at low frequencies. *Journal of the Acoustical Society of America*, 109(3):1110–1122, 2001.

[6] V. Algazi, R. Duda, D. Thompson, and C. Avendano. The CIPIC HRTF database. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 99–102, 2001.

[7] J. Allen and D. Berkley. Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

[8] C. Avendano, R. Duda, and V. Algazi. Modeling the contralateral HRTF. In *Proceedings of the 16th International Audio Engineering Society Conference: Spatial Sound Reproduction*, March 1999.

[9] J. Baldis. Effects of spatial audio on memory, comprehension, and preference during desktop conferences. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, pages 166–173, 2001.

[10] M. Barron and A. Marshall. Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure. *Journal of Sound and Vibration*, 77(2):211 – 232, 1981.

[11] D. Begault. *3-D Sound for Virtual Reality and Multimedia*. Academic Press, USA, 1994.

[12] D. Begault, E. Wenzel, and M. Anderson. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *Journal of the Audio Engineering Society*, 49(10):904–916, 2001.

[13] B. Beliczynski, I. Kale, and G. Cain. Approximation of FIR by IIR digital filters: An algorithm based on balanced model reduction. *IEEE Transactions on Signal Processing*, 40(3):532–542, 1992.

[14] R. Bellman and R. Kalaba. On adaptive control processes. *IRE Transactions on Automatic Control*, 4(2):1–9, 1959.

[15] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines. A blind source separation technique using second-order statistics. *IEEE Transactions on Signal Processing*, 45(2):434–444, 1997.

[16] L. Beranek. *Concert and Opera Halls: How They Sound*. Acoustical Society of America, New York, 1996.

[17] A. Berkhout. A holographic approach to acoustic control. *Journal of the Audio Engineering Society*, 36(12):977–995, 1988.

[18] P. Bloom. Creating source elevation illusions by spectral manipulation. *Journal of the Audio Engineering Society*, 25(9):560–565, 1977.

[19] A. Blumlein. Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems. British Patent Specification 394,325, 1931.

[20] M. Boone and E. Verheijen. Sound reproduction applications with wave-field synthesis. In *Audio Engineering Society Convention 104*, May 1998.

[21] J. Borish. Extension of the image model to arbitrary polyhedra. *Journal of the Acoustical Society of America*, 75(6):1827–1836, 1984.

[22] I. Bork. A comparison of room simulation software - The 2nd round robin on room acoustical computer simulation. *Acustica*, 86(6):943–956, 2000.

[23] M. Bouéri and C. Kyriakakis. Audio signal decorrelation based on a critical band approach. In *Audio Engineering Society Convention 117*, October 2004.

[24] A. Brand, O. Behrend, T. Marquardt, D. McAlpine, and B. Grothe. Precise inhibition is essential for microsecond interaural time difference coding. *Nature*, 417(6888):543–547, 2002.

[25] S. Busson. *Individualisation d'indices acoustiques pour la synthèse binaurale*. PhD thesis, Université de la Méditerranée Aix-Marseille II, 2006.

[26] R. Campbell and A. King. Auditory neuroscience: A time for coincidence? *Current Biology*, 14(20):R886–R888, 2004.

[27] S. Carlile, C. Jin, and V. van Raad. Continuous virtual auditory space using HRTF interpolation: acoustic and psychophysical errors. In *Proceedings of the IEEE International Symposium on Multimedia Information Processing*, 2000.

[28] C. Carr and M. Konishi. Axonal delay lines for time measurement in the owl's brainstem. *Proceedings of the National Academy of Sciences*, 85(21):8311–8315, 1988.

[29] C. Carr and M. Konishi. A circuit for detection of interaural time differences in the brain stem of the barn owl. *Journal of Neuroscience*, 10(10):3227–3246, 1990.

[30] B. Carty and V. Lazzarini. Frequency-Domain Interpolation of Empirical HRTF Data. In *Audio Engineering Society Convention 126*, May 2009.

[31] W. Cavanaugh, G. Tocci, and J. Wilkes. *Architectural Acoustics: Principles and Practice*. Wiley & Sons, 2009.

[32] S. Cerd, A. Gimnez, J. Romero, R. Cibrin, and J. Miralles. Room acoustical parameters: A factor analysis approach. *Applied Acoustics*, 70(1):97–109, 2009.

[33] D. Chandler and D. Grantham. Minimum audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth, and velocity. *Journal of the Acoustical Society of America*, 91(3):1624–1636, 1992.

[34] J. Chen, B. Van Veen, and K. Hecox. A spatial feature extraction and regularization model for the head-related transfer function. *Journal of the Acoustical Society of America*, 97(1):439–452, 1995.

[35] P. Chin, R. Corless, and G. Corliss. Optimization strategies for the approximate GCD problem. In *Proceedings of the International Symposium on Symbolic and Algebraic Computation*, 1998.

[36] N. Collins. *Introduction to Computer Music*. John Wiley & Sons Inc, 2010.

[37] R. Corless, M. Giesbrecht, and D. Jeffrey. Approximate polynomial decomposition. In *Proceedings of the International Symposium on Symbolic and Algebraic Computation*, 1999.

[38] E. Corteel, K. NGuyen, O. Warusfel, T. Caulkins, and R. Pellegrini. Objective and subjective comparison of electrodynamic and map loudspeakers for wave field synthesis. In *Proceedings of the 30th International Audio Engineering Society Conference: Intelligent Audio Environments*, March 2007.

[39] P. Cotterell. *On the Theory of the Second-Order Soundfield Microphone.* PhD thesis, University of Reading, 2002.

[40] P. Craven and M. Gerzon. Coincident microphone simulation covering three dimensional space and yielding various directional outputs. US Patent Specification 4042779, 1977.

[41] J. Daniel. *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia.* PhD thesis, Université Pierre et Marie Curie (Paris VI): Paris, 2001.

[42] J. Daniel. Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new ambisonic format. In *Proceedings of the 23rd International Audio Engineering Society Conference: Signal Processing in Audio Recording and Reproduction*, May 2003.

[43] M. F. Davis. History of spatial coding. *Journal of the Audio Engineering Society*, 51(6):554–569, 2003.

[44] N. DoCoMo. Press release on NTT DoCoMo homepage. `http://www.nttdocomo.com/pr/2009/001438.html`. Accessed September 13, 2010.

[45] R. Drullman and A. Bronkhorst. Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *Journal of the Acoustical Society of America*, 107(4):2224–2235, 2000.

[46] R. Duraiswami, D. Zotkin, and N. Gumerov. Interpolation and range extrapolation of HRTFs. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 45–48, 2004.

[47] M. Ericson, D. Brungart, and B. Simpson. Factors that influence intelligibility in multitalker speech displays. *International Journal of Aviation Psychology*, 14(3):313–334, 2004.

[48] M. Ericson and R. McKinley. The intelligibility of multiple talkers separated spatially in noise. In *Binaural and Spatial Hearing in Real and Virtual Environments*, pages 701–724. Lawrence Erlbaum Associates, Mahwah NJ, 1997.

[49] M. Evans, J. Angus, and A. Tew. Analyzing head-related transfer function measurements using surface spherical harmonics. *Journal of the Acoustical Society of America*, 104(4):2400–2411, 1998.

[50] A. Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Audio Engineering Society Convention 108*, February 2000.

[51] J. Flanagan. *Speech Analysis and Perception.* Springer-Verlag, Berlin, 1965.

[52] B. Fox. Hundred years of stereo: Fifty of hi-fi. *New Scientist*, 92:908–11, 1981.

[53] F. Freeland, L. Biscainho, and P. Diniz. Efficient HRTF interpolation in 3D moving sound. In *Proceedings of the 22nd International Audio Engineering Society Conference: Virtual, Synthetic and Entertainment Audio*, 2002.

[54] M. Gardner. Distance Estimation of 0° or Apparent 0°-Oriented Speech Signals in Anechoic Space. *Journal of the Acoustical Society of America*, 45(1):47–53, 1969.

[55] M. Gardner. Some monaural and binaural facets of median plane localization. *Journal of the Acoustical Society of America*, 54(6):1489–1495, 1973.

[56] M. Gardner and R. Gardner. Problem of localization in the median plane: effect of pinnae cavity occlusion. *Journal of the Acoustical Society of America*, 53(2):400–408, 1973.

[57] W. Gardner. *3-D audio using loudspeakers*. Kluwer Academic Publishers, 1998.

[58] W. Gardner and K. Martin. HRTF measurements of a KEMAR. *Journal of the Acoustical Society of America*, 97(6):3907–3908, 1995.

[59] P. Georgiou and C. Kyriakakis. Modeling of head related transfer functions for immersive audio using a state-space approach. In *Conference Record of the Thirty-Third Asilomar Conference on Signals, Systems, and Computers*, 1999.

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

[61] M. Gerzon. Practical periphony: The reproduction of full-sphere sound. In *Audio Engineering Society Convention 65*, February 1980.

[62] M. Gerzon. Ambisonics in multichannel broadcasting and video. *Journal of the Audio Engineering Society*, 33(11):859–871, 1985.

[63] M. Gerzon. General metatheory of auditory localisation. In *Audio Engineering Society Convention 92*, March 1992.

[64] D. Grantham, B. Hornsby, and E. Erpenbeck. Auditory spatial resolution in horizontal, vertical, and diagonal planes. *Journal of the Acoustical Society of America*, 114(2):1009–1022, 2003.

[65] D. Grantham, J. Willhite, K. Frampton, and D. Ashmead. Reduced order modeling of head related impulse responses for virtual acoustic displays. *Journal of the Acoustical Society of America*, 117(5):3116–3125, 2005.

[66] H. Gray. *Anatomy of the human body*. Lea & Febiger, 1918.

[67] G. Grindlay and M. Vasilescu. A multilinear (tensor) framework for HRTF analysis and synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 161–164, 2007.

[68] B. Grothe. New roles for synaptic inhibition in sound localization. *Nature Reviews Neuroscience*, 4(7):540–550, 2003.

[69] B. Grothe, C. Carr, J. Casseday, B. Fritzsch, and C. Koppl. The evolution of central pathways and their neural processing patterns. In *Evolution of the vertebrate auditory system*, pages 289–359. Springer Verlag, 2004.

[70] P. Guillon, T. Guignard, and R. Nicol. Head-related transfer function customization by frequency scaling and rotation shift based on a new morphological matching method. In *Audio Engineering Society Convention 125*, October 2008.

[71] H. Hacihabiboglu, B. Gunel, and F. Murtagh. Wavelet-based spectral smoothing for head-related transfer function filter design. In *Proceedings of the 22nd International Audio Engineering Society Conference: Virtual, Synthetic and Entertainment Audio*, pages 131–136, 2002.

[72] K. Hancock and B. Delgutte. A physiologically based model of interaural time difference discrimination. *Journal of Neuroscience*, 24(32):7110–7117, 2004.

[73] Y. Haneda, Y. Kaneda, and N. Kitawaki. Common-acoustical-pole and residue model and its application to spatial interpolation and extrapolation of a room transfer function. *IEEE Transactions on Speech and Audio Processing*, 7(6):709–717, 1999.

[74] Y. Haneda, S. Makino, Y. Kaneda, and N. Kitawaki. Common-acoustical-pole and zero modeling of head-related transfer functions. *IEEE Transactions on Speech and Audio Processing*, 7(2):188–196, 1999.

[75] J. Hebrank and D. Wright. Spectral cues used in the localization of sound sources on the median plane. *Journal of the Acoustical Society of America*, 56(6):1829–1834, 1974.

[76] G. Henning. Detectability of interaural delay in high-frequency complex waveforms. *Journal of the Acoustical Society of America*, 55(1):84–90, 1974.

[77] T. Hidaka, L. Beranek, and T. Okano. Interaural cross-correlation, lateral fraction, and low-and high-frequency sound levels as measures of acoustical quality in concert halls. *Journal of the Acoustical Society of America*, 98(2):988–1007, 1995.

[78] T. Hidaka, Y. Yamada, and T. Nakagawa. A new definition of boundary point between early reflections and late reverberation in room impulse responses. *Journal of the Acoustical Society of America*, 122(1):326–332, 2007.

[79] F. Hollerweger. Periphonic sound spatialization in multi-user virtual environments. Master's thesis, Austrian Institute of Electronic Music and Acoustics (IEM), 2006.

[80] Q. Huang and K. Liu. A reduced order model of head-related impulse responses based on independent spatial feature extraction. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 281–284, 2009.

[81] C. Huszty, B. Nemeth, P. Baranyi, and F. Augusztinovicz. Measurement-based fuzzy interpolation of room impulse responses. *Journal of the Acoustical Society of America*, 123(5):3771, 2008.

[82] InertiaCube Sensor. Intersense homepage. `http://www.intersense.com/InertiaCube_Sensors.aspx`. Accessed October 5, 2010.

[83] IRCAM. Listen HRTF database. `http://recherche.ircam.fr/equipes/salles/listen/index.html`, 2003. Accessed September 13, 2010.

[84] F. Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1:67–72, 1975.

[85] Y. Iwaya. Individualization of head-related transfer functions with tournament-style listening test: Listening with others ears. *Acoustical Science and Technology*, 27(6):340–343, 2006.

[86] L. Jeffress. A place theory of sound localization. *Journal of comparative and physiological psychology*, 41(1):35–39, 1948.

[87] J.-M. Jot. An analysis/synthesis approach to real-time artificial reverberation. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 221–224, 1992.

[88] J.-M. Jot, V. Larcher, and J.-M. Pernaux. A comparative study of 3-d audio encoding and rendering techniques. In *Proceedings of the 16th International Audio Engineering Society Conference: Spatial Sound Reproduction*, March 1999.

[89] J.-M. Jot, V. Larcher, and O. Warusfel. Digital signal processing issues in the context of binaural and transaural stereophony. In *Audio Engineering Society Convention 98*, February 1995.

[90] Y. Kahana. *Numerical Modelling of the Head-Related Transfer Function.* PhD thesis, University of Southampton, 2000.

[91] Y. Kahana and P. A. Nelson. Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric models. *Journal of Sound and Vibration*, 300(3-5):552 – 579, 2007.

[92] B. Kapralos, N. Mekuz, A. Kopinska, and S. Khattak. Dimensionality reduced HRTFs: a comparative study. In *Proceedings of the International Conference on Advances in Computer Entertainment Technology*, pages 59–62, 2008.

[93] B. Katz. Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation. *Journal of the Acoustical Society of America*, 110(5):2440–2448, 2001.

[94] E. Keogh and M. Pazzani. Derivative dynamic time warping. In *First SIAM International Conference on Data Mining*, 2001.

[95] F. Keyrouz and K. Diepold. A rational HRTF interpolation approach for fast synthesis of moving sound. In *Proceedings of the 4th IEEE Signal Processing Education Workshop*, pages 222–226, 2006.

[96] S.-M. Kim and W. Choi. On the externalization of virtual sound images in headphone reproduction: A wiener filter approach. *Journal of the Acoustical Society of America*, 117(6):3657–3665, 2005.

[97] O. Kirkeby, P. Nelson, H. Hamada, and F. Orduna-Bustamante. Fast deconvolution of multichannel systems using regularization. *IEEE Transactions on Speech and Audio Processing*, 6(2):189–194, 1998.

[98] D. Kistler and F. Wightman. A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction. *Journal of the Acoustical Society of America*, 91(3):1637–1647, March 1992.

[99] M. Kleiner, B.-I. Dalenbäck, and P. Svensson. Auralization-an overview. *Journal of the Audio Engineering Society*, 41(11):861–875, 1993.

[100] C. Knapp and G. Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(4):320–327, 1976.

[101] S. Kopuz and N. Lalor. Analysis of interior acoustic fields using the finite element method and the boundary element method. *Applied Acoustics*, 45(3):193–210, 1995.

[102] G. Kuhn and R. Guernsey. Sound pressure distribution about the human head and torso. *Journal of the Acoustical Society of America*, 73(1):95–105, 1983.

[103] A. Kulkarni, S. Isabelle, and H. Colburn. On the minimum-phase approximation of head-related transfer functions. In *Proceedings of the IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 84–87, 1995.

[104] H. Kuttruff. *Room acoustics*. Taylor & Francis, 2000.

[105] V. Larcher and J.-M. Jot. Techniques d'interpolation de filtres audio-numériques: Application à la reproduction spatiale des sons sur écouteurs. In *Proceedings of the Congres Français d'Acoustique (CFA)*, 1997.

[106] V. Larcher, J.-M. Jot, and G. Vandernoot. Equalization methods in binaural technology. In *Audio Engineering Society Convention 105*, September 1998.

[107] R. Lee. Shelf filters for ambisonic decoders. `http://www.ambisonia.com/Members/ricardo/shelfs.zip/download`. Accessed October 18, 2010.

[108] E. Lehmann and A. Johansson. Prediction of energy decay in room impulse responses simulated with an image-source model. *Journal of the Acoustical Society of America*, 124(1):269–277, 2008.

[109] E. Lehmann and A. Johansson. Diffuse reverberation model for efficient image-source simulation of room impulse responses. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1429–1439, 8 2010.

[110] J. Leung and S. Carlile. PCA Compression of HRTFs and localization performance. In *Proceedings of the International Workshop on the Principles and Applications of Spatial Hearing*, 2009.

[111] T. Liitola. Headphone Sound Externalization. Master's thesis, Helsinki University of Technology, 2006.

[112] J. Mackenzie, J. Huopaniemi, V. Valimaki, and I. Kale. Low-order modeling of head-related transfer functions using balanced model truncation. *IEEE Signal Processing Letters*, 4(2):39, 1997.

[113] K. Maki and S. Furukawa. Reducing individual differences in the external-ear transfer functions of the Mongolian gerbil. *Journal of the Acoustical Society of America*, 118(4):2392–2404, 2005.

[114] M. Matsumoto, S. Yamanaka, M. Toyama, and H. Nomura. Effect of Arrival Time Correction on the Accuracy of Binaural Impulse Response Interpolation-Interpolation Methods of Binaural Response. *Journal of the Audio Engineering Society*, 52(1/2):56–61, 2004.

[115] D. McAlpine. Creating a sense of auditory space. *The Journal of Physiology*, 566(1):21, 2005.

[116] A. McKeag and D. McGrath. Sound field format to binaural decoder with head tracking. In *Audio Engineering Society Convention 6r*, August 1996.

[117] S. Mehrgardt and V. Mellert. Transformation characteristics of the external human ear. *Journal of the Acoustical Society of America*, 61(6):1567–1576, June 1977.

[118] F. Menzer and C. Faller. Obtaining binaural room impulse responses from b-format impulse responses. In *Audio Engineering Society Convention 125*, October 2008.

[119] J. Merimaa and V. Pulkki. Spatial impulse response rendering I: Analysis and synthesis. *Journal of the Audio Engineering Society*, 53(12):1115–1127, 2005.

[120] J. Middlebrooks, E. Macpherson, and Z. Onsan. Psychophysical customization of directional transfer functions for virtual sound localization. *Journal of the Acoustical Society of America*, 108(6):3088–3091, 2000.

[121] A. Mills. On the minimum audible angle. *Journal of the Acoustical Society of America*, 30(4):237–246, 1958.

[122] P. Minnaar, J. Plogsties, S. Olesen, F. Christensen, and H. Möller. The interaural time difference in binaural synthesis. In *Audio Engineering Society Convention 108*, February 2000.

[123] A. Møller. *Hearing: Anatomy, physiology, and disorders of the auditory system*. Academic Press, 2006.

[124] H. Møller. Fundamentals of binaural technology. *Applied Acoustics*, 36(3 & 4):171–218, 1992.

[125] H. Møller, M. Sørensen, C. Jensen, and D. Hammershøi. Binaural technique: Do we need individual recordings? *Journal of the Audio Engineering Society*, 44(6):451–469, 1996.

[126] B. Moore. *Hearing*. Academic Press, 1995.

[127] B. Moore and B. Glasberg. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, 74(3):750–753, 1983.

[128] B. Moore, S. Oldfield, and G. Dooley. Detection and discrimination of spectral peaks and notches at 1 and 8 kHz. *Journal of the Acoustical Society of America*, 85(2):820–836, 1989.

[129] J. A. Moorer. About this reverberation business. *Computer Music Journal*, 3(2):13–28, 1979.

[130] S. Moreau, J. Daniel, and S. Bertet. 3D Sound Field Recording with Higher Order Ambisonics–Objective Measurements and Validation of a 4th Order Spherical Microphone. In *Audio Engineering Society Convention 120*, May 2006.

[131] D. Murphy, A. Kelloniemi, J. Mullen, and S. Shelley. Acoustic modeling using the digital waveguide mesh. *IEEE Signal Processing Magazine*, 24(2):55–66, 2007.

[132] A. Musicant and R. Butler. The influence of pinnae-based spectral cues on sound localization. *Journal of the Acoustical Society of America*, 75(4):1195–1200, 1984.

[133] MyEars. Myears homepage. `http://www.myears.net.au/`. Accessed September 13, 2010.

[134] D. Nagasaka, N. Kaneda, K. Itoh, M. Otani, M. Shimizu, M. Sugimoto, M. Hashimoto, and M. Kayama. A real-time network board game system using tactile and auditory senses for the visually impaired. In *Computers Helping People with Special Needs*, volume 6179 of *Lecture Notes in Computer Science*, pages 255–262. Springer Berlin / Heidelberg, 2010.

[135] J. Nam, M. Kolar, and J. Abel. On the minimum-phase nature of head-related transfer functions. In *Audio Engineering Society Convention 125*, October 2008.

[136] G. Naylor and J. Rindel. Predicting Room Acoustical Behaviour with the ODEON Computer Model. In *124th Meeting of Acoustical Society of America*, 1992.

[137] T. Nishino, S. Mase, S. Kajita, K. Takeda, and F. Itakura. Interpolating HRTF for auditory virtual reality. In *Proceedings of the 3rd Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan*, 1996.

[138] M. Noisternig, A. Sontacchi, T. Musil, and R. Höldrich. A 3D ambisonic based binaural sound reproduction system. In *Proceedings of the 24th International Audio Engineering Society Conference: Multichannel Audio*, 2003.

[139] T. Okamoto, B. Katz, M. Noisternig, Y. Iwaya, and Y. Suzuki. Implementation of real-time room auralization using a surrounding 157 loudspeaker array. In *Proceedings of the International Workshop on the Principles and Applications of Spatial Hearing (IWPASH)*, 2009.

[140] T. Okano, L. Beranek, and T. Hidaka. Relations among interaural cross-correlation coefficient (IACC), lateral fraction (LF), and apparent source width (ASW) in concert halls. *Journal of the Acoustical Society of America*, 104(1):255–265, 1998.

[141] A. Oppenheim, R. Schafer, and J. Buck. *Discrete-time signal processing*. Prentice hall Englewood Cliffs, NJ, 1989.

[142] A. Palmer. Reassessing mechanisms of low-frequency sound localisation. *Current Opinion in Neurobiology*, 14(4):457–460, 2004.

[143] M. Pecka, A. Brand, O. Behrend, and B. Grothe. Interaural time difference processing in the mammalian medial superior olive: the role of glycinergic inhibition. *Journal of Neuroscience*, 28(27):6914–6925, 2008.

[144] D. Perrott and A. Musicant. Minimum auditory movement angle: Binaural localization of moving sound sources. *Journal of the Acoustical Society of America*, 62(6):1463–1466, 1977.

[145] D. Perrott and K. Saberi. Minimum audible angle thresholds for sources varying in both elevation and azimuth. *Journal of the Acoustical Society of America*, 87(4):1728–1731, 1990.

[146] D. Perrott and J. Tucker. Minimum audible movement angle as a function of signal frequency and the velocity of the source. *Journal of the Acoustical Society of America*, 83(4):1522–1527, 1988.

[147] J. Plogsties, S. Olesen, P. Minnaar, F. Christensen, and H. Møller. Audibility of all-pass components in head-related transfer functions. In *Audio Engineering Society Convention 108*, Febuary 2000.

[148] G. Pollak. Model hearing. *Nature*, 417(6888):502–503, 2002.

[149] V. Raykar, R. Duraiswami, and B. Yegnanarayana. Extracting the frequencies of the pinna spectral notches in measured head related impulse responses. *Journal of the Acoustical Society of America*, 118(1):364–374, 2005.

[150] L. Rayleigh. On our perception of sound direction. *Philosophical Magazine*, 13(74):214–232, 1907.

[151] W. Reichardt, O. Abdel Alim, and W. Schmidt. Abhngigkeit der grenzen zwischen brauchbarer und unbrauchbarer durchsichtigkeit von der art des musikmotives, der nachhallzeit und der nachhalleinsatzzeit. *Applied Acoustics*, 7(4):243–264, 1974.

[152] S. Robeson. Spherical methods for spatial interpolation: Review and evaluation. *Cartography and Geographic Information Science*, 24(1):3–20, 1997.

[153] S. Roffler and R. Butler. Factors that influence the localization of sound in the vertical plane. *Journal of the Acoustical Society of America*, 43(6):1255–1259, 1968.

[154] S. Rosiles. *Auditory localization under spatial disorientation.* PhD thesis, Texas Tech University, 1997.

[155] M. Rothbucher, H. Shen, and K. Diepold. Dimensionality reduction in HRTF by using multiway array analysis. *Human Centered Robot Systems*, pages 103–110, 2009.

[156] P. Runkle, M. Blommer, and G. Wakefield. A comparison of head related transfer function interpolation methods. In *Proceedings of the IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, 1995.*, pages 88–91, 1995.

[157] K. Saberi, L. Dostal, T. Sadralodabai, and D. Perrott. Minimum audible angles for horizontal, vertical, and oblique orientations: Lateral and dorsal planes. *Acustica*, 75(1):57–61, 1991.

[158] K. Saberi and D. Perrott. Minimum audible movement angles as a function of sound source trajectory. *Journal of the Acoustical Society of America*, 88(6):2639–2644, 1990.

[159] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49, 1978.

[160] L. Savioja, J. Huopaniemi, T. Lokki, and R. Vaananen. Creating interactive virtual acoustic environments. *Journal of the Audio Engineering Society*, 47(9):675–705, 1999.

[161] J. Schnupp and C. Carr. On hearing with more than one ear: Lessons from evolution. *Nature Neuroscience*, 12(6):692–697, 2009.

[162] M. Schroeder. Integrated-impulse method measuring sound decay without using impulses. *Journal of the Acoustical Society of America*, 66(2):497–500, 1979.

[163] E. Shaw. Acoustic response of external ear replica at various angles of incidence. *86th Annual Meeting of the Acoustical Society of America*, 1973.

[164] E. Shaw and R. Teranishi. Sound pressure generated in an external-ear replica and real human ears by a nearby point source. *Journal of the Acoustical Society of America*, 44(1):240–249, 1968.

[165] S. Smyth, M. Smyth, and S. Cheung. Smyth SVS headphone surround monitoring for studios. In *Proceedings of the 23rd UK Audio Engineering Society Conference: Music Everywhere*, 2008.

[166] E. Start. *Direct sound enhancement by wave field synthesis*. PhD thesis, Technische Universiteit Delft, 1997.

[167] J. Steinberg and W. Snow. Physical factors. *Bell System Technical Journal*, 13:245–258, 1934.

[168] R. Stewart and M. Sandler. Statistical measures of early reflections of room impulse responses. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx)*, pages 59–62, 2007.

[169] U. Svensson, R. Fred, and J. Vanderkooy. An analytic secondary source model of edge diffraction impulse responses. *Journal of the Acoustical Society of America*, 106(5):2331–2344, 1999.

[170] C. Tan and W. Gan. User-defined spectral manipulation of HRTF for improved localisation in 3D sound systems. *Electronics Letters*, 34(25):2387–2389, 1998.

[171] J. Torres, M. Petraglia, and R. Tenenbaum. Low-order modeling of head-related transfer functions using wavelet transforms. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, volume 3, pages 513–516, 2004.

[172] B. Treeby, J. Pan, and R. Paurobally. The effect of hair on auditory localization cues. *Journal of the Acoustical Society of America*, 122(6):3586–3597, 2007.

[173] M. Van Wanrooij and A. Van Opstal. Relearning sound localization with a new ear. *Journal of Neuroscience*, 25(22):5413–5424, 2005.

[174] M. Vorländer. *Auralization: Fundamentals of acoustics, modelling, simulation, algorithms and acoustic virtual reality*. Springer Verlag, 2008.

[175] vSpace. vspace homepage. `http://www.vspace.net/`. Accessed September 13, 2010.

[176] G. Wahba. *Spline models for observational data*. Society for Industrial Mathematics, 1990.

[177] H. Wallach. The role of head movements and vestibular and visual cues in sound localization. *Journal of Experimental Psychology*, 27(4):339–368, 1940.

[178] L. Wang, F. Yin, and Z. Chen. Head-related transfer function interpolation through multivariate polynomial fitting of principal component weights. *Acoustical Science and Technology*, 30(6):395–403, 2009.

[179] E. Wenzel, M. Arruda, D. Kistler, and F. Wightman. Localization using nonindividualized head-related transfer functions. *Journal of the Acoustical Society of America*, 94(1):111–123, 1993.

[180] G. White, G. Fitzpatrick, and G. McAllister. Toward accessible 3D virtual environments for the blind and visually impaired. In *Proceedings of the 3rd International Conference on Digital Interactive Media in Entertainment and Arts (DIMEA)*, 2008.

[181] B. Wiggins. *An investigation into the real-time manipulation and control of three-dimensional sound fields*. PhD thesis, University of Derby, 2004.

[182] F. L. Wightman and D. J. Kistler. Monaural sound localization revisited. *Journal of the Acoustical Society of America*, 101(2):1050–1063, 1997.

[183] F. L. Wightman and D. J. Kistler. Resolution of front–back ambiguity in spatial hearing by listener and source movement. *Journal of the Acoustical Society of America*, 105(5):2841–2853, 1999.

[184] M. Williams. Unified theory of microphone systems for stereophonic sound recording. In *Audio Engineering Society Convention 82*, March 1987.

[185] M. Williams and G. Le Dû. The quick reference guide to multichannel microphone arrays part 1: Using cardioid microphones. In *Audio Engineering Society Convention 110*, May 2001.

[186] H. Wittek. *Perceptual differences between wavefield synthesis and stereophony.* PhD thesis, University of Surrey, 2007.

[187] H. Wittek and G. Theile. The recording angle-based on localisation curves. In *Audio Engineering Society Convention 112*, April 2002.

[188] R. Woodworth and H. Schlosberg. *Experimental Psychology.* Holt, Rinehart and Winston, New York, 1962.

[189] Z. Wu, F. Chan, F. Lam, and J. Chan. A time domain binaural model based on spatial feature extraction for the head-related transfer function. *Journal of the Acoustical Society of America*, 102(4):2211–2218, 1997.

[190] Y. Xie and B. Wiltgen. Adaptive Feature Based Dynamic Time Warping. *International Journal of Computer Science and Network Security*, 10(1):264–273, 2010.

[191] P. Young. The role of head movements in auditory localization. *Journal of Experimental Psychology*, 14(2):95–124, 1931.

[192] P. Zahorik. Assessing auditory distance perception using virtual acoustics. *Journal of the Acoustical Society of America*, 111(4):1832–1846, 2002.

[193] Z. Zeng and B. Dayton. The approximate GCD of inexact polynomials. In *Proceedings of the International Symposium on Symbolic and Algebraic Computation*, pages 320–327, 2004.

[194] W. Zhang, T. Abhayapala, and R. Kennedy. Horizontal plane HRTF reproduction using continuous Fourier-Bessel functions. In *Proceedings of the 31st International Audio Engineering Society Conference: New Directions in High Resolution Audio*, June 2007.

[195] D. Zotkin, R. Duraiswami, and N. Gumerov. Regularized HRTF fitting using spherical harmonics. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (ICASSP)*, pages 257–260, 2009.

[196] D. Zotkin, J. Hwang, R. Duraiswami, and L. Davis. HRTF personalization using anthropometric measurements. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (ICASSP)*, pages 157–160, 2003.

[197] E. Zwicker. Subdivision of the audible frequency range into critical bands (frequenzgruppen). *Journal of the Acoustical Society of America*, 33(2):248–248, 1961.