Auditory Scene Synthesis using Virtual Acoustic Recording and Reproduction

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

Gavin Kearney
March 2010

Department of Electronic and Electrical Engineering
Trinity College Dublin



For Margaret, Al, Karen and Roz

Contents

Contents **i**

1 Introduction **1**
1.1 Introduction . . . 1
1.2 Analysis and Synthesis of Auditory Scenes . . . 2
1.2.1 Auralization of Room Responses . . . 3
1.2.2 Interactive Virtual Auditory Environment Generation . . . 6
1.3 Focus of the Thesis: Virtual Acoustic Recording and Reproduction . . . 7
1.3.1 Structure of the Thesis . . . 8
1.3.2 Contributions to the Field . . . 9
1.3.3 List of Publications . . . 9

2 Spatial Hearing **13**
2.1 Introduction . . . 13
2.2 Sound Localization in a 'Free Field' . . . 14
2.2.1 Source Propagation in a Free-Field . . . 15
2.2.2 The Nature of the Source . . . 17
2.2.3 Resolution of the Auditory System . . . 18
2.2.4 Localization Cues . . . 20
2.2.4.1 Interaural Level Difference . . . 21
2.2.4.2 Interaural Time Difference . . . 23
2.2.4.3 Monaural Spectral Cues . . . 26
2.2.5 Localization Blur . . . 27
2.2.6 Distance Perception in a Free Field . . . 28
2.2.7 Free-Field Localization Summary . . . 30
2.3 Sound Localization in a Reverberant Field . . . 31
2.3.1 Source Propagation in a Reverberant Field . . . 31
2.3.2 The Precedence Effect . . . 34
2.3.3 Distance Perception in Reverberant Environments . . . 35
2.3.4 Apparent Source Width in Reverberant Rooms . . . 37

2.3.5 Properties of the IACC . . . 38
2.3.6 A Head Model for Broadband ITD to Angle Mapping . . . 43
2.4 An Improved Model for ITD Estimation in Reverberant Environments . . . 45
2.4.1 Generalized Cross-Correlation and the IACF . . . 46
2.4.2 Objective Evaluation of IACC-PHAT in a Reverberant Environment . . . 48
2.4.3 Localization Experiment I: Evaluation of IACC-PHAT Through Perceived Localisation in a Reverberant Environment . . . 51
2.5 Conclusions . . . 54

3 Spatial Audio Systems for VAE presentations 55
3.1 Introduction . . . 55
3.2 Spatial Reproduction Classes . . . 56
3.2.1 Binaural Reproduction . . . 56
3.2.2 Multi-Loudspeaker Reproduction . . . 62
3.3 Stereophony . . . 67
3.3.1 Stability of Stereophonic Images . . . 70
3.3.2 Recording for Intensity Stereo . . . 72
3.3.3 VBAP: Stereophony for Arbitrary Numbers of Loudspeakers . . . 74
3.3.4 Stereophony Summary . . . 77
3.4 Ambisonic Soundfields . . . 77
3.4.1 Recording for Ambisonics . . . 80
3.4.2 Decoding Ambisonics . . . 81
3.4.2.1 Decoding Through Projection . . . 81
3.4.2.2 Decoding Through Pseudoinverse . . . 82
3.4.3 Psychoacoustic Optimization of Ambisonic Decoders . . . 83
3.4.4 Ambisonic Decoder Types . . . 85
3.4.5 Higher Order Ambisonics . . . 90
3.4.5.1 Recording for Higher Order Ambisonics . . . 95
3.4.5.2 Decoding for Higher Order Ambisonics . . . 96
3.4.6 Ambisonics Summary . . . 97
3.5 Wavefield Synthesis . . . 98
3.5.1 Reproduction using Discrete Linear Arrays . . . 102
3.5.2 Recording For Wavefield Synthesis . . . 109
3.5.3 Wave Field Synthesis Summary . . . 110
3.6 Wave Field Synthesis and Higher Order Ambisonics . . . 111
3.7 Conclusions . . . 111

4 A Comparative Study of Spatialization Techniques Under Real Listening Conditions 115

4.1 Introduction . . . 115
4.2 VBAP and 1st Order Ambisonics in Real Listening Conditions . . . 116
4.2.1 Localization Study Objectives . . . 118
4.2.2 Objective Analysis for Single and Multiple Listeners . . . 119
4.2.2.1 Sweet-Spot Localization . . . 120
4.2.2.2 Off-Centre Localization Estimators . . . 125
4.2.2.3 Off-Centre Localization Analysis . . . 126
4.2.2.4 Decoder Optimization . . . 131
4.2.3 Comments on Objective Analysis . . . 134
4.2.4 Subjective Listening Experiment II.A: Localization of VBAP and 1st Order Ambisonics Under Real Listening Conditions . . . 135
4.2.4.1 Experimental Procedure . . . 135
4.2.4.2 Results . . . 138
4.2.4.3 Room Investigations . . . 141
4.2.5 Subjective Listening Experiment II.B: Coloration of Moving Phantom Sources using VBAP and 1st Order Ambisonics under Real Listening Conditions . . . 146
4.2.5.1 Experimental Procedure . . . 146
4.2.5.2 Results . . . 147
4.2.6 Conclusions from VBAP and 1st Order Ambisonics Analysis . . . 149
4.3 Higher Order Ambisonic Decoders for Virtual Auditory Environments . . . 150
4.3.1 Objective Analysis of Higher Order Ambisonic Localization . . . 151
4.3.1.1 Sweet Spot Analysis . . . 151
4.3.1.2 Off-Centre Localization . . . 156
4.3.1.3 Comments on Objective Analysis of HOA decoders . . . 160
4.3.2 Subjective Listening Experiment III: Analysis of Higher Order Ambisonic Localization . . . 161
4.3.2.1 Results: Subjective Localization of HOA at the Sweet Spot . . . 164
4.3.2.2 Results: Subjective Localization of HOA at Off-Centre Listening Positions . . . 167
4.4 Conclusions . . . 168

5 Virtual Acoustic Recording 171
5.1 Introduction . . . 171
5.2 Natural Acoustic Recording . . . 172
5.2.1 Direct Field Recording . . . 173
5.2.2 Applicability of Direct-Field Stereophonic Recording Techniques . . . 173
5.2.3 Multichannel Recording in the Reverberant Field . . . 175
5.2.4 Natural Acoustic Recording using Ambisonic Techniques . . . 177

5.3 Virtual Acoustic Recording . . . 178
5.3.1 Convolution and Deconvolution in Multichannel Audio . . . 179
5.3.2 Spatial Impulse Responses . . . 180
5.3.3 Equalization of the Convolution Chain . . . 182
5.3.3.1 Measurement of Convolution Coloration . . . 183
5.3.3.2 Perceptual Effects of Coloration in the Convolution Chain . . . 187
5.4 Evaluation of Virtual Acoustic Recording . . . 191
5.4.1 Database Acquisition . . . 191
5.4.2 Binaural Comparison of Reproduced Recordings . . . 192
5.4.3 Subjective Experiment IV.A: Subjective Attributes of Actual Vs. Virtual Convolution Based Recordings . . . 199
5.4.3.1 Results . . . 201
5.4.4 Subjective Experiment IV.B: Subjective Preference of Actual Vs. Virtual Convolution Based Recordings . . . 203
5.4.4.1 Conclusions on Actual Vs. Virtual Recording Assessment . . . 203
5.5 Directivity Enhancement for Virtual Acoustic Recordings . . . 206
5.5.1 Source Simulation Radiators . . . 206
5.5.2 Measurement of Source Radiation . . . 208
5.6 Conclusions . . . 213

6 Spatial Impulse Response Interpolation 215
6.1 Introduction . . . 215
6.2 Spatial Resolution of SRIR Datasets . . . 216
6.2.1 Subjective Experiment V: Spatial Resolution in Virtual Environments with Motion Parallax . . . 220
6.2.1.1 Wavefield Measurement and Reproduction . . . 221
6.2.1.2 Test Implementation . . . 224
6.2.1.3 Results . . . 225
6.2.2 Derivation of Perceptual-Based SRIR Topology . . . 226
6.3 Interpolation of Spatial Impulse Responses . . . 227
6.3.1 Transition Region in Room Impulse Responses . . . 229
6.3.2 Interpolation of Early Reflections using Dynamic Time Warping . . . 229
6.4 Synthesis of Diffuse Decay . . . 233
6.5 Analysis of DTW Interpolation Algorithm: Ambisonic Room Simulations . . . 234
6.6 Analysis of DTW Interpolation Algorithm: Measured Soundfields . . . 237
6.6.1 Wavefield Capture and Spatial Downsampling . . . 238
6.6.2 Objective Analysis of the Reproduced Wavefields . . . 240
6.6.3 Perceptual Localization . . . 242
6.7 SRIR Topology Based on DTW Interpolation Algorithm . . . 244

6.8 Conclusions 246

7 Virtual Acoustic Reproduction 249
7.1 Introduction 249
7.2 Headphone Reproduction: Virtual Loudspeaker Optimization 250
7.2.1 Algorithm 251
7.2.2 Factorization Analysis using the CIPIC Dataset 253
7.3 Headphone Equalization 257
7.4 Higher Order Ambisonic Reproduction over Virtual and Real Loudspeakers . . . 261
7.4.1 Virtual Ambisonic Reproduction 261
7.4.2 Higher Order Ambisonic Synthesis 262
7.4.3 Examples of HOA Re-encoding using Directional Analysis 264
7.5 Conclusions 266

8 Conclusions 269
8.1 Summary and Conclusions 269
8.2 Further Work 272
8.2.1 Integration of Signal Processing Strategies to VAE Production Frameworks 273
8.2.2 Development of Virtual Acoustic Recording and Reproduction Strategies for Computational Based Auralization 275
8.2.3 Further Optimization Procedures for Binaural Synthesis 275
8.3 Final Remarks 275

Bibliography 277

A Test Environment 1 295
A.1 Test environment 1 measurement and simulation 296
A.1.1 Test Environment 1:Acoustic measurements 298
A.1.2 Test Environment 1: EASE Simulations 306

B Loudspeaker Layouts 315

C VBAP Localization Vectors 317

D 1^{st} Order Ambisonic Energy Vectors 325

E Higher Order Ambisonic Energy Vectors 335

F Speech Data 345
F.1 TIMIT Speech Corpus Phrases 346

G ISM and Real Measurement Comparison 349

G.0.1 Test Environment 1: SRIR Measurements at Violin Positions 350
G.0.2 Test Environment 1: SRIR Measurements at Female Vocal Positions . . . 352
G.0.3 Test Environment 1: ISM Simulation at Performance Distances 354

H Measured Wavefields **357**

I DVD Contents **361**

Abstract

This thesis is concerned with recording of real acoustic events for spatial audio presentations in interactive Virtual Auditory Environments (VAEs). A VAE is an environment where auditory perception does not correspond to the physical environment of the listener, but rather to a virtual one. The current state of the art of spatial audio reproduction in VAEs focuses on real-time walk-through audio presentations using computational-based auralization, where the transfer functions describing the acoustic interaction between a sound source and a listener's ears in a reverberant room are computationally modeled. The purpose of such systems is that the end-listener will no longer be limited to single-perspective recordings of acoustic events such as musical performances, but can instead move around a VAE as if they were at an original performance. Whilst such methods can yield a reasonable spatial impression of an acoustic event, auralization using real acoustic measurements (i.e. data-based auralization) offers a level of authenticity not yet fully achieved with computational-based methods. On the other hand, significant challenges exist in the use of data-based auralization, since the incorporation of listener movement in the virtual space requires a dataset of room impulse responses to be measured. Furthermore, the plausibility of the reproduction of acoustic events in a VAE is also dictated by the spatial audio reproduction method. This thesis hypothesizes that practical data-based auralization can be achieved for both individual and multi-listener scenarios using the same recording/measurement and reproduction paradigm. Of particular interest is the potential to utilize current commercially available technology to achieve this.

In supporting this hypothesis, two main strands of research are defined, termed 'virtual acoustic recording' and 'virtual acoustic reproduction'. 'Virtual acoustic recording' outlines the direct-field capture of the musical event, as well as its subsequent convolution with Spatial Room Impulse Responses (SRIRs). The process considers the perceptual differences between convolution-based recordings and real recordings, in particular the removal of coloration effects in the convolution chain, as well as the synthesis of the directional properties of the source using commercially available loudspeakers. A novel interpolation algorithm, as well as a perceptual analysis of source localization in reverberant environments is then used to reduce the number of measurements required in the SRIR dataset.

'Virtual Acoustic Reproduction' then refers to the spatial audio presentation of an acoustic event in both single and multi-listener scenarios in a VAE. To this end, an in-depth study of spatial audio reproduction strategies is presented, where the localization accuracy of systems such as Vector Based Amplitude Panning, Ambisonics, and Wave Field Synthesis is assessed through both objective and subjective measures. The computational challenges involved in headphone reproduction of spatial audio presentations are simplified using a novel factorization algorithm with reduces the length of head-related impulse response filters significantly, without loss of spatial information.

Finally, conclusions from this thesis work are drawn pertaining to the research hypothesis, and future directions are presented, which include the incorporation of the signal processing techniques presented for computational-based auralization.

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,

Gavin Kearney

March 30, 2010.

Acknowledgments

I would first like to thank Prof. Frank Boland for allowing me to explore the possibilities of audio signal processing applications to virtual recording technology. His constant guidance, support, and enthusiasm, as well as his endless DSP knowledge have been a major inspiration throughout all phases of this work. I would also like to express my gratitude to Dr. Fionnuala Conway for providing the financial (and moral!) support over the last three years, without which this work would not have been accomplished. My thanks also go to Dr. Dermot Furlong for his wealth of knowledge of all things psychoacoustic and his good humor during my constant barrage of questions!

A huge thank you goes to my colleague and collaborator Claire Masterson, whose input was invaluable to the direction of this work. My heartfelt thanks also go to Jeff Levison of Euphonix Inc., for helping me investigate some of the underlying concepts of this work in the early stages, as well as Enda Bates for collaborating with me on the initial perceptual evaluation. I am also grateful to Ken Giles of Soundfield Ltd., Dr. Helmut Wittek of Schoeps Gmbh, Gordon Kapes of Studio Technologies, Inc. and Deirdre Costello of Westland Studios Dublin for allowing me the use of their equipment and facilities. Many thanks also to Prof. Anil Kokaram for giving me much needed travel finances, and to Science Foundation Ireland for also supporting this work.

A huge thanks to all my colleagues and friends in Sigmedia, including Deirdre 'the rocker' O'Regan and (D.J.) Dan Ring (your honorary degrees in Ph.D counseling are in the post!), Damien Kelly, Daire Lennon, Riccardo 'Please fix my equalizer' Veglianti, Dr. David Corrigan, and Dr. François Pitié. Thanks also to the members of the Music Technology Research Group, TCD including Liam O'Sullivan, Mark Linnane and Les Keys. A big thank you to Stephen Adams who had important input in the final stages of this work, as well as during the construction of the Wavefield Synthesis array, and to John 'Flash' Squires, for his fantastic photography skills. Also thanks to the technical support staff in the Department of Electronics, in particular, Conor Nolan, Robbie Dempsey and Bernadette Clerkin, as well as Nora Moore for all her help throughout the years!

Constant inspiration for this thesis has come from the students of the Music and Media Technologies course in TCD. I am forever grateful to the efforts of Marcin Gorzel, Laura Kilty, Paul Donnan, Michael Kearns, Marek Olik, Sheila Rodgers, Emma Brennan and Conor Hinfey for their assistance in the experimental setups. A huge thanks also goes to all the participants of the listening tests who gave up their valuable time to contribute to this work.

This work could never have been accomplished without the unwavering support and love from my family, for which I am eternally grateful. This is as much the fruit of their labour as it is mine. Finally and most importantly to Roz, thank you for all the love and support you have given me throughout this work...You make life's journey a much easier road to travel!

List of Acronyms

ADC Angle Dependent Component

ADReSs Azimuth Discrimination and Re-Synthesis

AIC Angle Independent Component

ASW Apparent Source Width

BIR Binaural Impulse Response

CAVE CAVE Automatic Virtual Environment

DirAC Directional Audio Coding

DTW Dynamic Time Warping

FOA First Order Ambisonics

HOA Higher Order Ambisonics

HPTF Headphone Pinna Transfer Function

HRIR Head Related Impulse Response

HRTF Head Related Transfer Function

IACC Interaural Cross Correlation Coefficient

IACF Interaural Cross Correlation Function

ILD Interaural Level Difference

IPD Interaural Phase Difference

IR Impulse Response

ITD Interaural Time Difference

JND Just Noticeable Difference

KDE Kernel Density Estimate

OCT Optimized Cardioid Triangle

ORTF Office de Radiodiffusion Télévision Française

PHAT PHAse Transform

RIR Room Impulse Response

RT Reverberation Time

SRIR Spatial Room Impulse Response

VAE Virtual Auditory Environment

VBAP Vector Based Amplitude Panning

SIRR Spatial Impulse Response Rendering

SSR Source Simulation Radiator

WFS Wave Field Synthesis

1 Introduction

1.1 Introduction

Over the last 30 years, technology intent on creating realistic auditory scenes has found its way into many diverse areas of modern living, ranging from architectural engineering to the entertainment industry. The presentation of auditory cues which can fool the brain into thinking it is perceiving actual reality is the primary goal of such systems. For the cinematic presentation, sound designers and sound engineers can place sounds around the audience for increased dramatic effect. In music production, recordings can be presented such that the listener has the impression that they are in the original acoustic space where the recordings were taken. In architectural design, an acoustic consultant can show their clients alternative acoustic responses as a result of different architectural decisions and/or internal finishes. The primary objective in all cases, is the synthesis of an auditory scene which attempts to present as faithfully as possible, soundfields perceptually equivalent to those experienced naturally. That is, for any auditory scene reproduction system to give the feeling of "you are there", it must have a good psychoacoustical foundation. This is an issue of simulating the listening experience through knowledge of:

1. The directional characteristics of the source.
2. The propagation of the source in the acoustic space.
3. The interaction the source and acoustic space with the head and ears.

The listening experience can therefore be considered as an acoustic communications channel, where the properties of the source, receiver and acoustic medium are modeled [Savioja et al., 1999]. This process is what is commonly referred to as 'auralization' [Kleiner et al., 1993]. There are many different ways to implement auralization ranging from mathematical to physical modeling of room acoustics and source-ear interaction. However, it is well known that highly convincing results can be achieved by processing source signals using real acoustic measurements, resulting in an impression that the virtual source was recorded in the room where the measurement was taken. The many significant challenges for the recording and reproduction of room acoustics in an auralization system for convincing spatial audio presentations are the main research focus of this thesis. The purpose of this chapter is to give a background into this area and to set specific goals pertaining to contributions to the field.

1.2 Analysis and Synthesis of Auditory Scenes

An auditory scene can be defined as the perceptual organization of sound. Auditory scene analysis then is the evaluation of the physical properties of sound that lead to perceptual effects. It is related not only to the properties of the sound source, but also the environment in which it propagates and the subsequent reaction with a listener's head and ears. Bregman [Bregman, 1994] describes this process in humans as:

> *"The primitive process of scene analysis seems to employ a strategy of first breaking down the incoming array of energy into a large number of separate analyses. These are local to particular moments of time and particular frequency regions in the acoustic spectrum. Each region is described in terms of its intensity, its fluctuation pattern, the direction of frequency transitions in it, an estimate of where the sound is coming from in space, and perhaps other features".* Bregman, A.S., *Auditory Scene Analysis: The Perceptual Organization of Sound*, The MIT Press, p.p 641.

This statement relays not only the importance of both physical acoustic and psychoacoustic knowledge in the formation and analysis of auditory scenes, but also that there remains a large amount of phenomena pertaining to the perception of auditory scenes that are, as yet, still unknown. It follows, that any novel work in the creation of auditory scenes synthesis must first have a relevant psychoacoustical basis.

Auditory scenes can be formed using two main mechanisms: recording or auralization. In the first approach, recording for scene synthesis is usually implemented in a studio environment. For example, in popular music production, a multi-track approach is taken, whereby instruments are layered temporally, spectrally and spatially. Another example is in cinema, where Foley artists create the auditory scene by adding in everyday sounds synchronized to the actor's movements. In the second approach, the creation of auditory scenes through auralization involves the processing of recorded audio (preferably anechoic) with acoustic responses taken in real

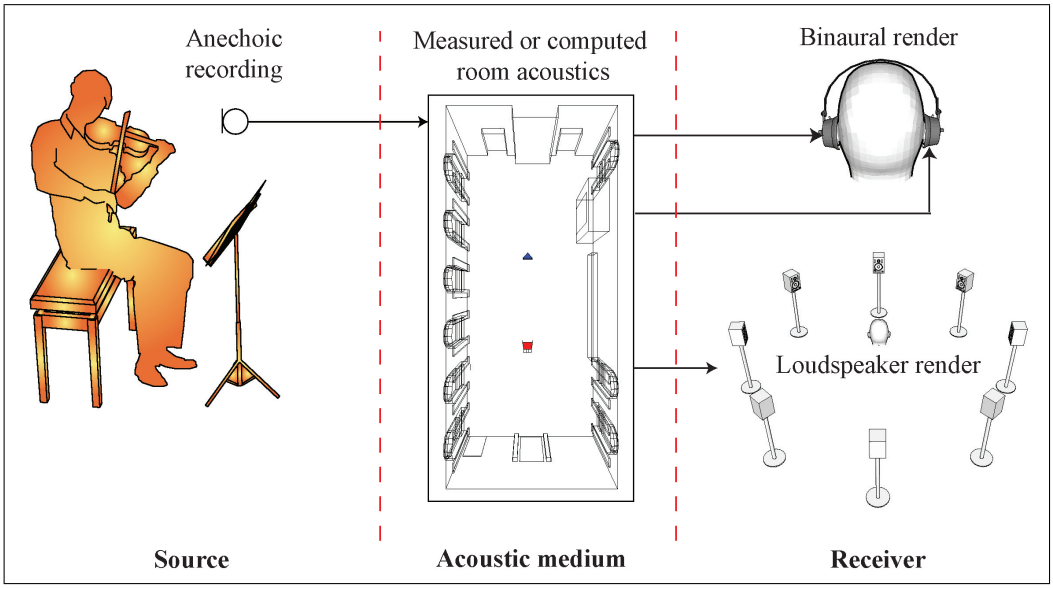
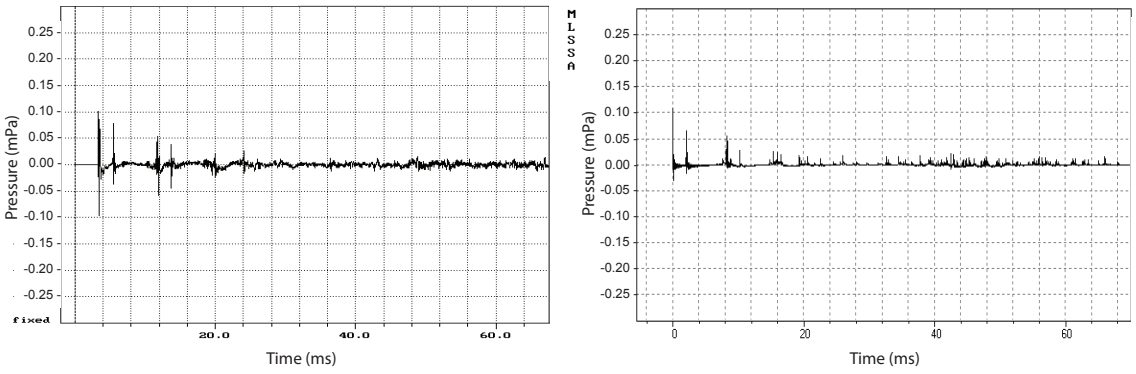


Figure 1.1: *The auralization concept: Source audio taken in an anechoic environment is filtered with measured or computed room acoustic responses, and the results rendered for headphone or loudspeaker listening.*

rooms or computed with auralization software. This concept is illustrated in Figure 1.1. This filtering method is the convolution of the source audio with a Room Impulse Response (RIR) that describes how the direct sound and reflections (due to multiple reflective paths in the room) behave at a particular point in space. This can then be rendered for loudspeaker reproduction or for headphone listening, delivering the perception that one is in the measured room listening to the source audio. An analysis of the quality of the final auralization depends not only on the RIR, but also on the methodology used for playback and evaluation [Lokki and Savioja, 2005]. In order to deliver realistic soundfields to the ears, spatial audio reproduction must be employed. For headphone listening, this is achieved by generating Binaural Impulse Responses (BIRs) which describe the interaction of the source in the measured/simulated acoustic environment with the head and ears. For loudspeaker reproduction, we measure Spatial Room Impulse Responses (SRIRs) such that the sound pressure and velocity at a particular listening point in the room can be reconstructed over loudspeakers.

1.2.1 Auralization of Room Responses

There are two main ways of obtaining RIRs: by computation using acoustic response modeling software or by measurement in real rooms. As a consequence, much debate is made regarding the nature and definition of auralization. In this work, the aforementioned methods are considered as two separate classes: *computational-based auralization* and *data-based auralization* respectively. Kleiner et al define auralization in the context of computation as

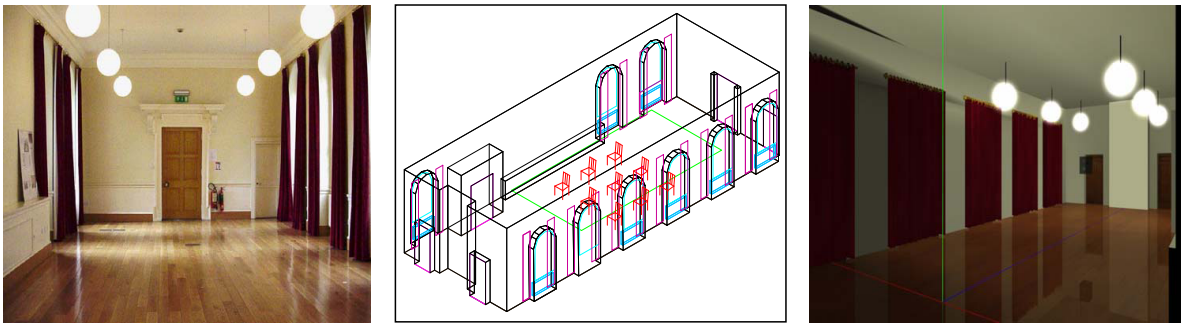


(a) (b)

Figure 1.2: *Comparison of impulse responses: (a) measured and (b) computed with the same geometrical parameters.*

> *"..the process of rendering audible by physical or mathematical modeling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modeled space"*, 'Auralization – An Overview', Kleiner, M., Dalenback, B.I., and Svensson, P., J. Audio Eng. Soc., 41(11), Nov. 1993, pp.861-875.

There are many different methods for computing impulse responses based on geometric or wave-based acoustic simulation, which have various advantages and disadvantages. Highly accurate results can be achieved using wave-based methods, such as the Finite Element Method (FEM) or the Boundary Element Method (BEM), and a good overview of these approaches can be found in [Pietrzyk, 1998]. In such methods, sound propagation is numerically modeled through interacting elements governed by wave propagation theory. The higher the density of elements in the simulation, the greater the spectral accuracy, with a resultant trade-off in computational expense. For this reason, such methods are generally limited to low frequency RIR estimation. The Finite Difference Time Domain (FDTD) method [Savioja et al., 1994] is also suited to low-frequency auralization, where a discrete approximation of wave propagation using finite time differences is simulated. Geometric-based solutions to calculating RIRs, such as the image-source method [Allen and Berkley, 1979] and the ray-tracing method [Krokstad et al., 1968] are well-suited to the mid-to high frequency regions. However, they do not consider phenomena such as diffraction or scattering, although hybrid image source-ray tracing methods have been proposed in [Rindel and Christensen, 2003] to overcome this. Furthermore, geometric methods assume that the wavelengths involved are much less than the dimensions of the room, and so are not suited to low frequency propagation models. Consequently, current auralization practice utilizes geometric-based simulations for mid to high frequencies and wave based simulations at low frequencies, with the ultimate aim of giving impulse responses equivalent to those measured



(a) (b) (c)

Figure 1.3: *Equivalent acoustic environments: (a) Real Acoustic Environment, (b) Acoustic model of real environment and (c) Virtual Environment.*

in real rooms. An example of a measured RIR in comparison to a computed RIR (using the ray-tracing method) is shown in Figure 1.2. The RIR in Figure 1.2(a) is the real room response, measured with a known stimulus in the room of Figure 1.3 (a). The RIR in Figure 1.2(b) is computed with auralization software using the acoustic model of Figure 1.3 (b), which has similar geometric and acoustic surface parameters as the real environment. A textured virtual environment, such as shown in Figure 1.3 (c) can then be constructed for the visual modality at playback, and the measured or computed impulse responses used in the final auralization. We note that whilst the arrival time of the reflections in each case is somewhat similar, there are discrepancies between the magnitudes of the measured and computed responses, partly due to uncertainties in the impulse measurement [Lundeby et al., 1995], but mainly due to limitations in the computational software. As a result, research has continuously strived to improve the field of computational-based auralization and to create better auralization. Good summaries of experiments comparing the different auralization approaches to real measurements can be found in [Vorländer, 1995], [Bork, 2000] and [Katz, 2004].

Source directivity in computational based auralization can be modeled through two main ways. The first is via directional filters, where the frequency dependent directional response of the source is known, and source propagation in different directions in the RIR computation is modeled through filtering of the source audio in the different directions [Savioja et al., 1999]. The second method involves multichannel recording of the source material in an anechoic environment using an array of microphones, where the resultant auralization reproduces the recorded audio at the source point in different directions simultaneously [Otondo and Rindel, 2005].

Data-based auralization allows us to create highly convincing auditory scenes yielding listening experiences over headphones and loudspeakers equivalent to those experienced in measured reverberant spaces. In measuring real room responses, a known stimulus is played out over a (typically omnidirectional) loudspeaker, recorded back into a digital computer and then post-processed such that the RIR is obtained. The result, when used as a filter on dry source audio

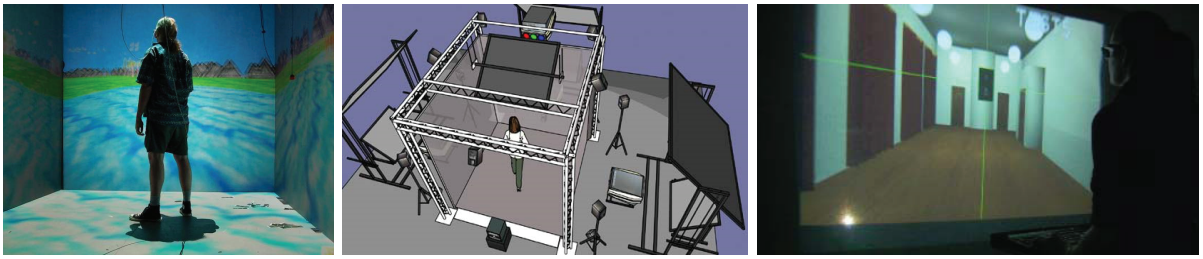
is a reasonable impression of the source in the acoustic space. There are however, several significant limitations with data-based auralization. First, the directionality of real sources (i.e. how their sound spatially radiates) is mis-represented by using commercial loudspeakers as the measurement source [Patynen et al., 2008b]. Secondly, the spectral character of each element in the measurement and reproduction chain leads to audible effects in the final auralization that differs from the true acoustic event [Kearney and Levison, 2008].

A further consideration for both computational and data-based auralization is the method of reproduction. For individual listening, headphone reproduction is desirable, since the acoustics of the reproduction environment can be ignored. However, 3-D audio reproduction over headphones is non-trivial. The common method for headphone rendering is to utilize the head-related transfer functions (HRTFs) that describe the interaction of an acoustic source from a given direction with the head and ears. The downfall is that each individual's HRTFs are unique and different auralizations are required for different listeners. Furthermore, headphones will also lead to timbral coloration of the final auralization, particularly at high frequencies, since the headphone response changes each time the headset is put on, due to minor positional shifts [Møller, 1992] [Hiekkanen et al., 2009]. For loudspeaker rendering, assuming the acoustics of the reproduction environment are not problematic, then the transparency (i.e. the naturalness) of the auralization is dependent on the spatialization method used to render the final audio, in particular if localization accuracy of virtual sources over an audience area is to be considered. 3-D spatial audio rendering techniques exist which can be applied to multiple audience scenarios, including Vector Base Amplitude Panning (VBAP) [Pulkki, 1997], Ambisonics [Gerzon, 1980] and Wave Field Synthesis (WFS) [Berkhout, 1988]. A comparison of the advantages and disadvantages of such systems in the context of auralization still remains to be explored.

1.2.2 Interactive Virtual Auditory Environment Generation

In recent years, commercially available music and film production software has utilized auralization processes for the creation of convolution reverberation effects, with notable examples in [Kessler, 2005] and [Merimaa and Pulkki, 2005]. Auralization for architectural acoustics has also shown significant commercial success, resulting in numerous competitive software packages which render room acoustic responses based on computer models, with leading examples in [Ahnert, 1992], [Rindel et al., 1994], and [Dalenbäck and Strömberg, 2006]. Conversely, multi-track recordings taken in anechoic rooms such as in [Denon, 1995] and [Patynen et al., 2008a] have proved extremely useful in enabling acoustic consultants, psychoacoustic analysts, and audio engineers to analyze the perceptual attributes of spatial sound with auralization processes.

However, an important aspect in the quest for realism in auditory scene synthesis is *user interaction*. That is, how the movements of a person listening to an auditory scene directly influence the scene presentation. Such interactive auditory scenes are often given the term 'Virtual Auditory Environments' (VAEs), and can be used in conjunction with 'Virtual Visual



(a) (b) (c)

Figure 1.4: *Virtual Visual Environments : (a) CAVE virtual reality system at the Electronic Visualization Laboratory, Chicago [Cruz-Neira, 1995], (b) EVE virtual reality system at the Helsinki University of Technology [Hiipakka et al., 2001] (c) Head-tracked virtual reality system of the Printing House Hall, Trinity College Dublin.*

Environments' (VVEs) such as in the CAVE (CAVE Automatic Virtual Environment) [Cruz-Neira, 1995], and EVE (Experimental Virtual Environment) structures [Hiipakka et al., 2001], as shown in Figure 1.4(a) and (b).

A common element used in VVE and VAE control is head-tracking. For example, in personalized headphone reproduction, binaural filters are adjusted according to the orientation of the head, so that the source localization remains stable. System latency is a key issue in such cases, and the head-tracking should update the filter kernel within perceptual limits. This often results in compromises in the HRTF filter length, leading to localization errors [Sandvad and Hammershøi, 1994]. Another desirable feature of VAEs is to give the user freedom to move about or 'walk-through' the virtual acoustic environment. For this, multiple SRIRs are required where the SRIR corresponding to the instantaneous position of the listener in the VAE is used at run-time. This involves the measurement of multiple SRIRs at different points in the room and is in general a tedious process. It is mainly for this reason that most VAEs are computational-based, rather than data-based. However, the use of real acoustic measurements in VAEs has the potential to yield highly convincing results. Furthermore, any computational based auralization process, utilized for interactive VAEs must be first benchmarked against real acoustic measurements, if its effectiveness is to be assessed. Currently, there is no strategy for the measurement of SRIRs for application to interactive VAEs. Consequently, no framework exists for capturing real acoustic events such as live concert performances for application to interactive VAEs.

1.3 Focus of the Thesis: Virtual Acoustic Recording and Reproduction

It is the purpose of this thesis to explore possible strategies for recording and reproduction of real acoustic events in interactive VAEs. More formally, it is concerned with the application of

multi-track recording strategies to data-based auralization in VAE structures. It is proposed that this methodology be known as *virtual acoustic recording and reproduction*. The main hypothesis of this work is as follows:

> **A plausible formation of natural auditory scenes can be constructed efficiently using data-based auralization in a virtual auditory environment, and that said scenes can be rendered to either individual or multiple listeners using the same recording and reproduction paradigm.**

This hypothesis is supported through results of investigations into

1. The psychoacoustical limitations of spatial audio reproduction systems, with particular focus on localization accuracy under real listening conditions,
2. The creation of signal processing strategies that support the virtual recording and reproduction approach in a practical manner,
3. A perceptual analysis of virtual recording strategies in comparison to real acoustic recordings.

The work developed will also inform the realm of computational based auralization.

1.3.1 Structure of the Thesis

The chapters that follow will investigate and assess the validity of the aforementioned hypothesis through a formal programme of research:

Chapter 2 outlines the relevant psychoacoustics for investigation of spatial sound reproduction and evaluation in real listening rooms as well as for binaural listening. Anechoic conditions are first investigated and the localization cues for single source listening are discussed. The effect of room acoustics on localization accuracy is then investigated, and a formal listening test is presented which demonstrates issues inherent in objective and subjective localization of sources in reverberant rooms.

Chapter 3 presents an in-depth analysis of previous work in the domain of spatial audio reproduction through extensive simulations. It traces the development of time and intensity based stereophonic techniques through to VBAP, Ambisonic and WFS systems.

Chapter 4 presents an objective and subjective study of VBAP and Ambisonic systems, under real listening conditions for distributed audiences. Of particular interest is the perceptual effects of psychoacoustically optimized Ambisonic decoders of low and high orders in comparison to VBAP. Based on this analysis a spatial audio reproduction method is recommended for horizontal reproduction that utilizes commercially available technology.

Chapter 5 outlines the virtual acoustic recording strategy. First, an analysis of the perceptual attributes of actual versus virtual acoustic recordings is presented. Particular attention is paid to coloration effects in the convolution chain. Next, a method for the approximation of source

directivity in data-based auralization, using measurements made with commercially available loudspeakers is presented.

Chapter 6 addresses the measurement of SRIRs. The perception of localization blur within reverberant environments is assessed with the aim of reducing the number of SRIR measurements required for walk-through auralizations. A novel interpolation algorithm for the early part of the SRIR is then presented, allowing for more sparse measurements to be taken. The limitations of the method are discussed, and example topologies derived from the performance of the method, as well as the perception of localization in reverberant environments, are shown.

Chapter 7 outlines a strategy for low computational expense in spatial audio reproduction over headphones using a simplification method for reducing binaural filter lengths. The method is compared to full-length binaural filters and relevant psychoacoustical parameters are investigated. A method of converting 1^{st} order SRIRs to higher order representations is also shown.

Chapter 8 summarizes the work accomplished in the thesis and draws conclusions pertaining to the research hypothesis. Future directions in which the work can be taken are indicated.

Further work supporting the main body of the thesis can be found in the the appendices. A DVD containing RIR measurements, relevant auralizations and acoustic model designs also accompanies this thesis.

1.3.2 Contributions to the Field

The following contributions to the field are presented throughout the course of this work:

- A novel binaural localization method for ITD estimation in reverberant environments incorporating use of the Phase Transform weighting function for interaural cross correlation (Chapter 2)
- An investigation into the localization accuracy of stereophonic and low and high order Ambisonic systems for distributed audiences in real listening rooms (Chapter 4)
- An investigation into the perceptual attributes of virtual versus actual multichannel acoustic recording (Chapter 5)
- A least squares approximation for horizontal source directivity using commercially available loudspeakers (Chapter 5)
- An algorithm for the interpolation of the direct sound and early reflections of SRIRs based on Dynamic Time Warping (Chapter 6)
- Development of an algorithm for the simplification of HRIRs (Chapter 7)

1.3.3 List of Publications

The following is a list of publications by the author pertaining to the work accomplished in this thesis:

1. 'Dynamic Time Warping for Acoustic Response Interpolation: Possibilities and Limitations', Gavin Kearney, Claire Masterson, Stephen Adams, Frank Boland, in Proc. of the 17th European Signal Processing Conference (EUSIPCO), August 2009.

2. 'Approximation of Binaural Room Impulse Responses', Gavin Kearney, Claire Masterson, Stephen Adams, Frank Boland, in Proc. of the Irish Signals and Systems Conference, Dublin, Ireland, June 2009.

3. 'Towards Efficient Binaural Room Impulse Response Synthesis', Gavin Kearney, Claire Masterson, Stephen Adams, Frank Boland, in Proc. of the EAA Symposium on Auralization, Espoo, Finland, June 2009.

4. 'Virtual vs. Actual Multichannel Acoustical Recording', Gavin Kearney and Jeff Levison, in Audio Engineering Society 134th Convention, May 2008.

5. 'Audio-Visual Processing Tools for Auditory Scene Synthesis', Gavin Kearney, Rozenn Dahyot and Frank Boland, in Audio Engineering Society 134^{th} Convention, May 2008.

6. 'Source Localisation Mapping using Weighted Interaural Cross-Correlation', Gavin Kearney, Damien Kelly, Enda Bates, Frank Boland and Dermot Furlong, in Proc. of the Irish Signals and Systems Conference, September 2007.

7. 'A Comparative Study of the Performance of Spatialization Techniques for a Distributed Audience in a Concert Hall Environment', Gavin Kearney, Enda Bates, Dermot Furlong and Frank Boland, in 31st International Conference of the Audio Engineering Society, June 2007.

Parts of the novel research developed in this thesis have also appeared in the following co-authored work:

1. 'A Method for Head Related Impulse Response Simplification', Claire Masterson, Stephen Adams, Gavin Kearney, Frank Boland, in Proc. of the 17th European Signal Processing Conference (EUSIPCO), August 2009

2. 'Localisation Quality in Source Separation Based Up-mixing Algorithms', Dan Barry and Gavin Kearney, in Audio Engineering Society 35th Conference, February 2009.

3. 'Acoustic Impulse Response Interpolation for Multichannel Systems using Dynamic Time Warping', Claire Masterson, Gavin Kearney and Frank Boland, in Audio Engineering Society 35th Conference, February 2009.

4. 'Monophonic Source Localization for a Distributed Audience in a Small Concert Hall', Enda Bates, Gavin Kearney, Frank Boland and Dermot Furlong, in Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07), September 2007.

5. 'Localisation Accuracy of Advanced Spatialisation Techniques in Medium-Sized Concert Halls', Enda Bates, Gavin Kearney, Dermot Furlong and Frank Boland, in 153rd meeting of the Acoustical Society Of America, June 2007.

2

Spatial Hearing

2.1 Introduction

The ability to discriminate spatially and localize sounds around us is a vital part of human life. It follows that in order to design effective VAEs, we must first have a thorough understanding of how humans perceive real world sounds. It is generally taken for granted that our ability to localize sound sources is directly related to the environment that surrounds us. For example, one can appreciate that it is usually easier to localize a sound source outdoors, than it is in a reverberant hall. What is often unclear however is how the physics of sound propagation relates to such perception. It is only through thorough physical, neurological and psychometric evaluation that we can draw conclusions on how we localize and discriminate sounds, as well as the environment around us. This is an area of study known as *psychoacoustics* and is the focus of this chapter.

We will begin our analysis of auditory perception of source signals by investigating sound propagation and localization in non-reverberant (anechoic) environments. The main cues for the human perception of localization will then be introduced, which will then form a basis for investigation of localization perception in real rooms and auditoria. Psychoacoustic metrics relevant to this thesis work are then introduced. Finally, a new method for source localization in reverberant environments using binaural microphones is presented.

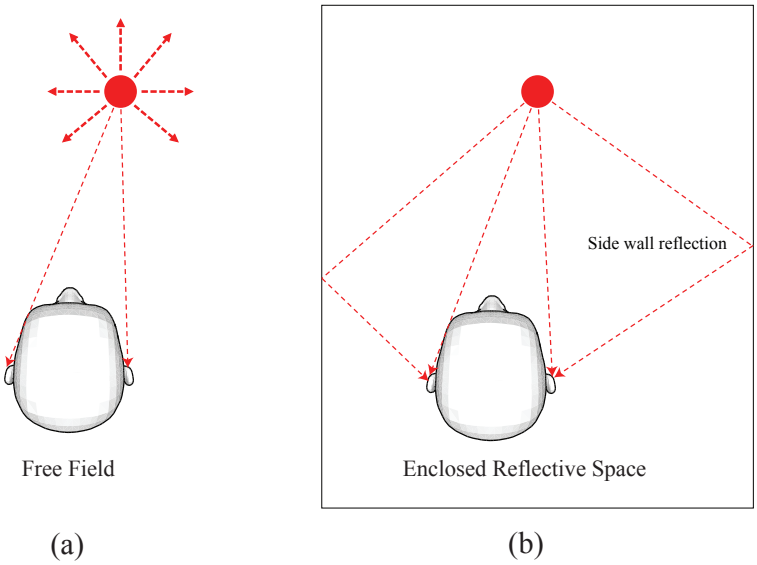


Figure 2.1: *Source localization: (a) In free-field conditions only the direct sound is heard (b) In enclosed spaces, the soundfield at the ears is made up of the direct sound and reflections off the room boundaries.*

2.2 Sound Localization in a 'Free Field'

Human beings are very much 'visual animals', and strong cases have been made that suggest our visual modality is the dominant localization cue. For example, Colavita [Colavita, 1974] has published a series of experiments that suggest that when visual and auditory stimuli are presented simultaneously, humans consistently respond to the visual cue and are often unaware that the auditory cue has occurred. However, any visual cue is limited quite literally to the 'field of vision', whereas the auditory cue is spatially aware of sounds from all directions. In other words, the auditory cue has a constant omnidirectional sensitivity. Localization is a remarkable feat, if one considers the fact that in reverberant rooms, multiple delayed versions of the source signal appear at the ear due to acoustic reflections off the room boundaries, as shown in Figure 2.1(b), resulting in highly complex ear signals. In fact, it is the acoustic environment which determines the *spatial impression* of the source and strong reflections can increase the perceived source width, making sources harder to localize. Conversely, a source which propagates in an environment with no reflections is said to exist in a 'free-field', as shown in Figure 2.1(a). As sound propagates in a free-field, only the direct sound is heard by the listener. Wide open outdoor spaces are the closest most people get to experiencing this, but anechoic chambers have been designed so that they have no reflections whatsoever, requiring large amounts of absorbing material to produce a free-field over a wide frequency range.

Study of sound propagation in a free-field reveals the idealistic behavior of sound and sound systems. We can outline the cues that affect the human auditory mechanism without considering multiple reflections. Likewise, knowledge of the behavior of sounds and sound systems in a free-

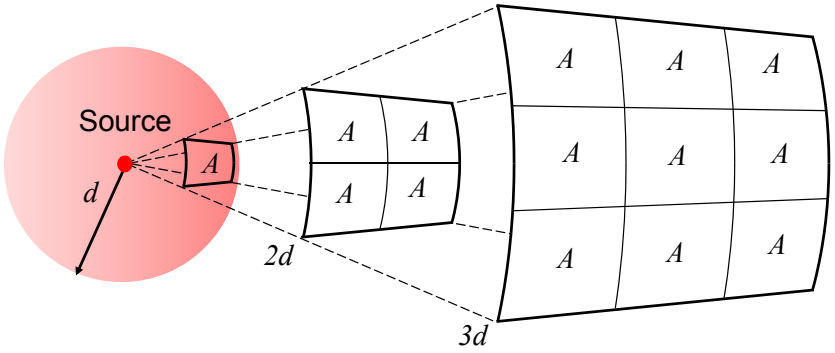


Figure 2.2: *Effective unit area for intensity measurement with increasing distance away from the source in free-field propagation. The same sound energy is distributed over increasingly wider areas and so intensity falls inversely proportional to the square of the distance from the source.*

field makes it a lot easier to understand the effect of room reflections on the same sources in real environments. Before we can investigate these issues however, we must first gain insight into the spatial, spectral and temporal characteristics of the source signal itself.

2.2.1 Source Propagation in a Free-Field

An ideal 'point source' radiates spherically in a free field and the effective surface area of this spherical radiation increases exponentially with distance from the source. This is illustrated in Figure 2.2. If the source radiates with a power P, then the intensity on the surface of the sphere is given by

$$I = \frac{P}{4\pi d^2} \tag{2.1}$$

where d is the distance from the source. Sound intensity is typically measured relative to the intensity threshold of human hearing I_0 which is 10^{-12}watts/m^2, and a measured intensity I can therefore be expressed in decibels as

$$I(\text{dB}) = 10 \log_{10}\left(\frac{I}{I_0}\right) \tag{2.2}$$

We can express this in terms of the sound pressure p by

$$\text{SPL(dB)} = 10 \log\left(\frac{p^2}{p_0^2}\right) \tag{2.3}$$

$$= 20 \log\left(\frac{p}{p_0}\right) \tag{2.4}$$

where $p_0 = 2 \times 10^{-5}(\text{Newtons}/m^2)$, the threshold of hearing. This yields an SPL reduction of 6dB per doubling of distance from the source. At large distances from the source (in comparison to the source wavelength, λ) the wave curvature becomes flat enough to be considered a *plane wave* as illustrated with the simple single-frequency wave in Figure 2.3.

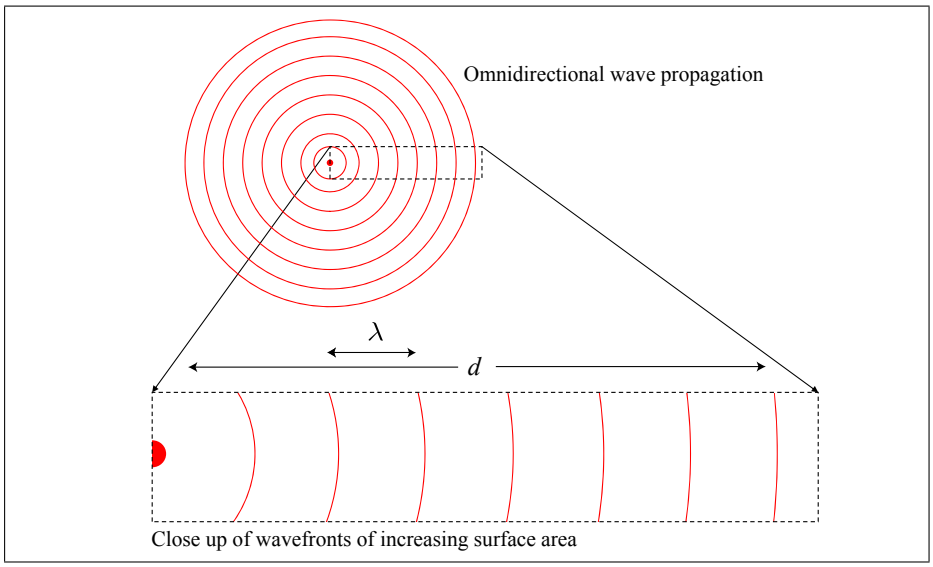


Figure 2.3: *As a point source propagates in a free-field, the wave curvature changes from spherical to planar ($\lambda << d$). The illustration shows the peaks in the SPL of a single frequency and the wavelength with respect to the overall distance from the source.*

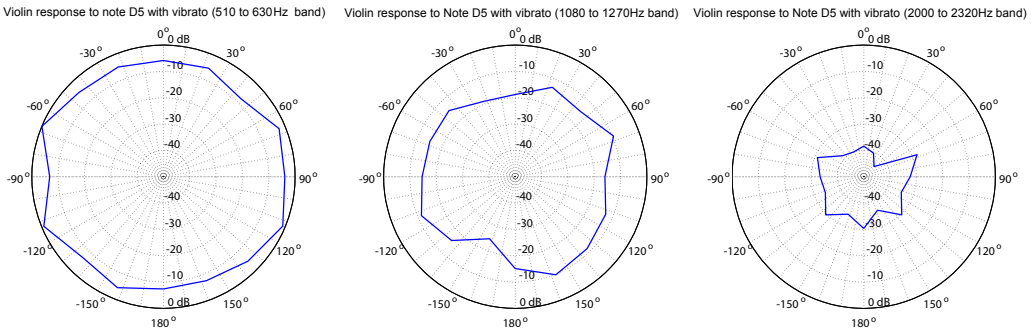


Figure 2.4: *Violin directivity characteristics for note D5 with vibrato at three different frequency bands, 510-630Hz, 1080-1270Hz and 2000 to 2320Hz. Each band clearly displays different directivity characteristics.*

The propagation behavior of Figure 2.3 is that of an omnidirectional source, i.e. where sound pressure is created uniformly in all directions. However, most real world sources have varying radiation patterns. For example, the sound pressure level of a human voice at the back of the head is less than that at the front. The directivity of a source relates how much sound will be directed towards a specific area, compared to the total sound energy being generated by the source. This is known as the 'Directivity Factor', D, and is equal to 1 for an omnidirectional source. Directivity can be depicted clearly on polar patterns such as those shown in Figure 2.4. In this example, the directivity (from a violin) is depicted in 3 frequency bands, 510-630Hz, 1080-1270Hz and 200-2320Hz. The method of directivity measurement is shown in Section 5.5.2. We

see that the radiation patterns in each band are quite different, exhibiting different directional intensities.

2.2.2 The Nature of the Source

The spectral and temporal characteristics of the sound source are also extremely relevant to source localization accuracy. The duration of the sound, as well as how quickly it rises and falls in amplitude, can benefit or hinder localization. In this context, sound sources can be divided into two broad categories:

- Steady-state sources: stationary sounds such as pure tones and
- Transient sources: sounds with onsets and offsets.

Hartmann [Hartmann, 1983] has presented a thorough study of the subjective localization of sound sources for both steady-state and transient signals. He found that localization for narrow-band steady-state signals is almost impossible and that this localization improves monotonically with increasing spectral density of the source. In fact, of all the test signals Hartmann used, localization of full bandwidth white noise in a free field yielded the best results (although it is noteworthy that under reverberant field conditions the localization of white noise is dependent on the room reverberation time). Hartmann attributed this fact to the brain interpreting the noise as a continuous series of small impulses. Hartmann also varied the onset time of the signals in a free field and used impulsive tones and tones with slow-onsets (approx. 7 seconds). Hartmann found that the localization performance was significantly poorer for the slow-onset signals. He concluded that the abrupt-onset transient signals provide an envelope cue which is absent from the slow-onset tone and that it provides a signal which is spectrally broadened, and excites more neurons, if only briefly, than does a slow-onset tone. Lastly, he observed that the broadened spectrum of the abrupt onset tone is modified by the direction-dependent filtering of the ear. Based on this work, we can conclude that if we are to assess the localization accuracy within a VAE, then

1. The source signals should have abrupt-onset transients and broad spectral content to be correctly localized.
2. The sound system used must be able to cater for broad spectral content for correct localization to occur.

These characteristics are therefore of extreme importance to the localization studies in this thesis. Although these properties do not tell us directly how the brain processes the directional information of sound, they do infer that such processing is spectrally and temporally dependent. In light of this, we will now discuss the spectral and temporal resolution of the auditory system.

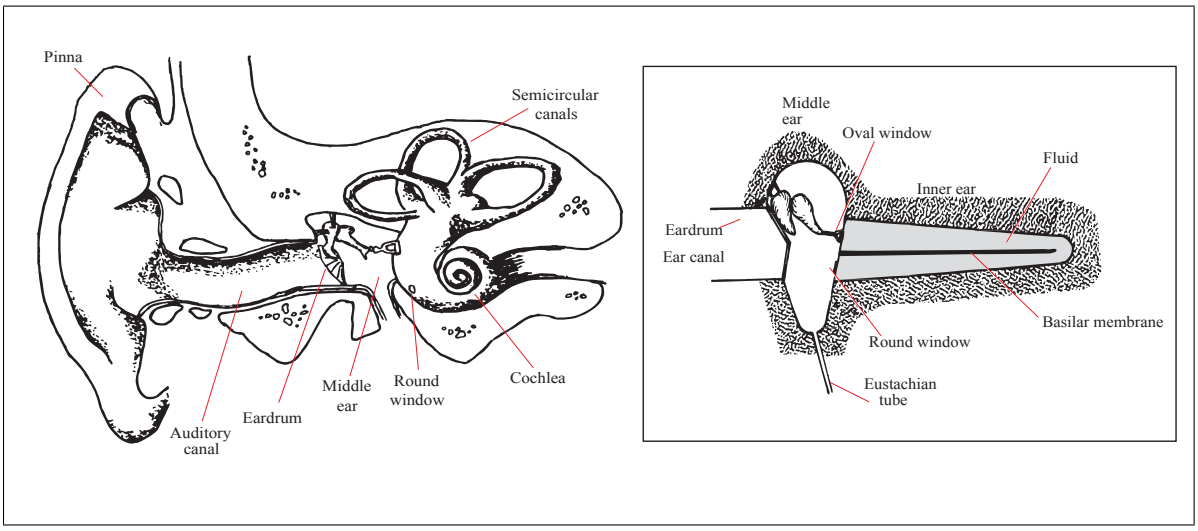
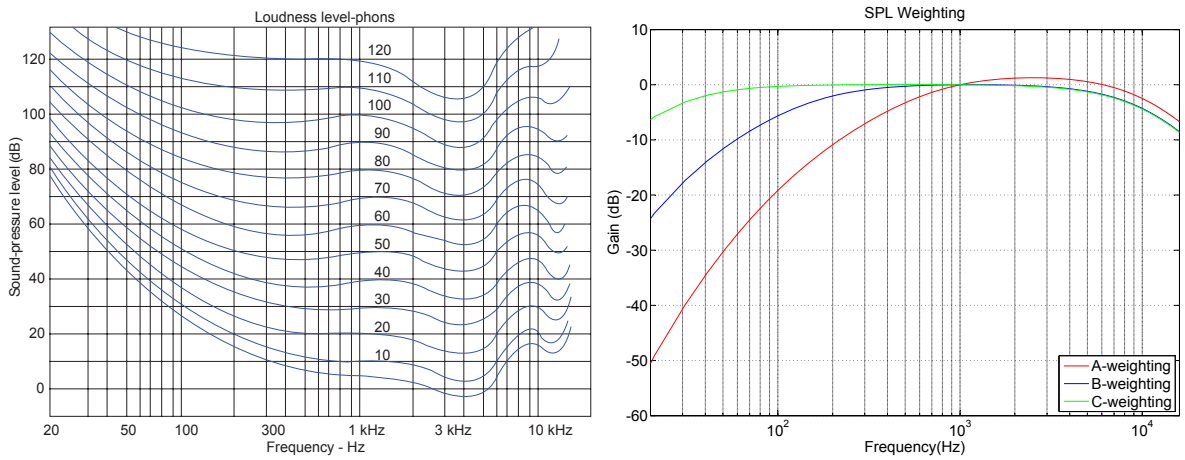


Figure 2.5: *Physiological components of the ear and (inset) the inner ear. [Everest, 1994]*

2.2.3 Resolution of the Auditory System

In any exploration of the resolution of human hearing, we must first have a basic understanding of the physiology of the ear. Figure 2.5 shows the four main components of ear physiology: the external ear, the auditory canal, the middle ear and the inner ear. The external ear consists of the pinnae, the ear canal (also known as the auditory canal or meatus) and the tympanic membrane (eardrum). It is here that spatial filtering begins, and the pinnae, as well as the size of the head play an important role in this process. The external ear can be viewed as a linear filtering system, and is represented by the so-called Head-Related Transfer Functions (HRTFs), described in Section 2.2.4. The middle ear acts as an impedance matching device, where the eardrum vibrates due to the air pressure fluctuations between the air-filled middle ear and the external sound pressure due to some stimulus. The vibrations are then transmitted via the oval window to the fluid-filled cochlea in the inner ear, and in turn, to the hair cells on the basilar membrane. It has been shown in [Von Bekesy, G., 1960] that stimuli of different frequencies causes different hair cells to vibrate, with high frequency stimuli vibrating cells near the oval window, and low frequency stimuli causing movement along much of the basilar membrane. These hair cells allow us to hear the full auditory range, and to perceive the differences in timbre and pitch of sound.

As a consequence of our ear physiology, the perceived loudness of sounds is frequency dependent. For example, we can appreciate that the quarter-wavelength resonance of the ear canal improves the overall sensitivity of our hearing in the region of 2-4kHz, a frequency range particularly important for speech. Furthermore, this perception changes with the intensity of the source. Fletcher and Munson [Fletcher and Munson, 1933] derived the now famous equal-loudness curves, which demonstrated that as the sound intensity increases at the ear, the perceived difference in loudness between tones at different frequencies decreases. The ISO-338 standardized version of



(a) (b)

Figure 2.6: *Perceptual loudness measures: (a) ISO-226 equal loudness contours [ISO, 2009a] and (b) A, B and C weighting filters.*

these measurements, shown in Figure 2.6(a), form the basis of the A, B and C-weighting sound pressure measurement weights utilized in this thesis and shown in Figure 2.6 (b).

A further consequence of our physiology is that the response of the basilar membrane to two sinusoidal tones depends on their frequency separation. For tones with frequencies that are close together the basilar membrane will fail to resolve each sinusoidal component due to physical limitations in its motion. Thus the ability of the ear to resolve spectral components is limited, and as a result, sounds can be made inaudible by other sounds simultaneously presented to the ear which are particularly close in frequency. This is known as *spectral masking* and forms the basis for 'critical band' theory, first introduced by Fletcher [Fletcher, 1940], where the detection of partial sounds within the same critical band is perceived as a single entity. Fletcher hypothesized that the ear could be modeled as a bank of continuous bandpass filters, known as 'critical-band filters'. This has led to development of several critical band scales useful in auditory filter design such as the Bark scale [Zwicker, 1961]. This scale is derived from listening tests in which the bandwidth of narrow-band noise is adjusted with constant energy until its loudness is perceived to increase. The relationship between Barks (z) and frequency is approximately [Moore, 1995]

$$z = [26.8/(1 + 1.96/f)] - 0.53 \tag{2.5}$$

The result is that the auditory spectrum can be divided into 24 critical bands. The ERB (Equivalent Rectangular Bandwidth) scale [Patterson, 1976] is another critical band scale which results in smaller bandwidth filters. Here the ERB scale is derived by using notched noise masking listening tests, where the passband bandwidth of masking noise is varied and the

threshold of audibility of a sinusoid is determined. The ERB scale is related to frequency by

$$R_{ERB} = 21.3 \log_{10}(1 + f/228.7Hz) \tag{2.6}$$

This results in a filterbank of approximately 42 bands. Both ERB and Bark filterbank implementations are used throughout the course of this thesis.

Temporal masking or 'non-simultaneous' masking occurs when one stimulus masks another temporally. When the masked signal occurs first, and its threshold of detection must be raised due to some masking stimulus (which occurs a short time later), this is known as 'backwards masking'. Likewise, if the masking stimulus occurs first, and the threshold of detection of the masked stimulus is to be raised, this is known as 'forward masking'. Forward masking can persist up to 100ms after the masker sound, and is dependent on both the level and spectral content of the source [Moore, 1995]. However, much debate still exists as to the effects of backward masking. Models of temporal masking have been proposed in [Plack and Oxenham, 1998], [Festen and Plomp, 1981], [Festen and Plomp, 1983], [Moore et al., 1988], and [Plack and Moore, 1990], and the reader should consult these papers for further details. Temporal masking in the context of the 'precedence effect' in room acoustics is further discussed in Section 2.3.2. From the succint review presented here on the limitations of the auditory system, we will now consider aspects of directional hearing that are dependent on the spectral and temporal content of sound sources.

2.2.4 Localization Cues

Throughout this thesis, directional hearing is described in terms of a three dimensional coordinate system with the human head at the origin. This is shown in Figure 2.7. Horizontal directions are termed *azimuth* and vertical directions *elevation*. The ability to discriminate sources in both azimuth and elevation is based on the differences we perceive in the signals presented to the auditory system. These differences can be described in terms of the HRTF, previously introduced. HRTFs characterize both the *monaural transfer function* at a single ear and the *interaural transfer function* between both ears [Blauert, 2003]. The *monaural transfer function* relates the sound pressure at the ear canal for any position of the sound source, to sound pressure measured at the same point but with the sound source at a reference angle and distance (in general, a plane wave with azimuth $\theta = 0^\circ$, and elevation $\phi = 0^\circ$ is used as a reference). In this way, we can characterize the spectral changes that occur at one ear canal due to different source positions. The *interaural transfer function* relates sound pressure levels measured at corresponding points at the two ear canals, with the reference sound pressure being the ear facing the sound source (the 'ipsilateral' ear). This allows us to characterize the binaural (two-ear) localization cues due to arrival time and intensity difference of sound at the ears, commonly referred to as Interaural Time Difference (ITD) and Interaural Level Difference (ILD) respectively. The concept of ITD and ILD was first introduced by Lord Rayleigh in 1887

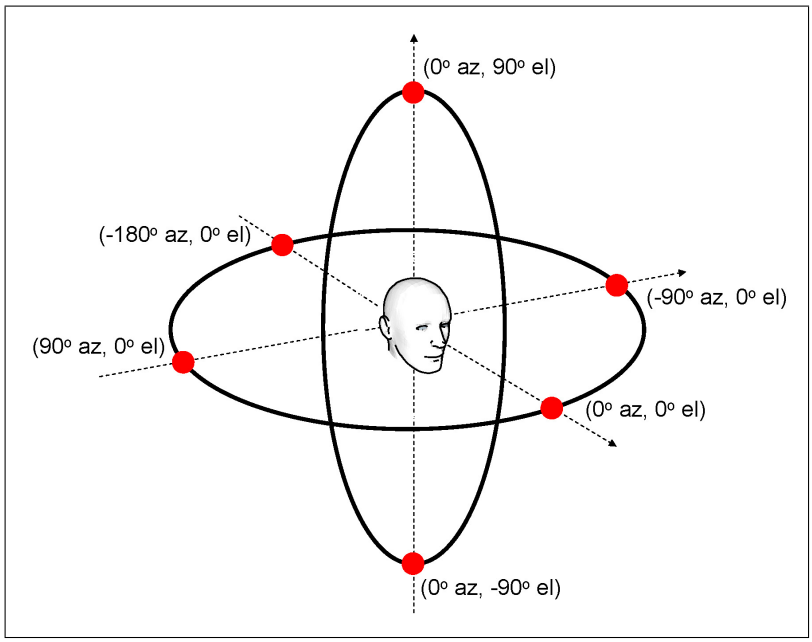


Figure 2.7: *Coordinate system for directional hearing: The convention in this thesis is that sources to the left of the head have negative angles, as do sources below the horizontal plane.*

---

and is known as 'Duplex Theory' [Rayleigh, 1945].

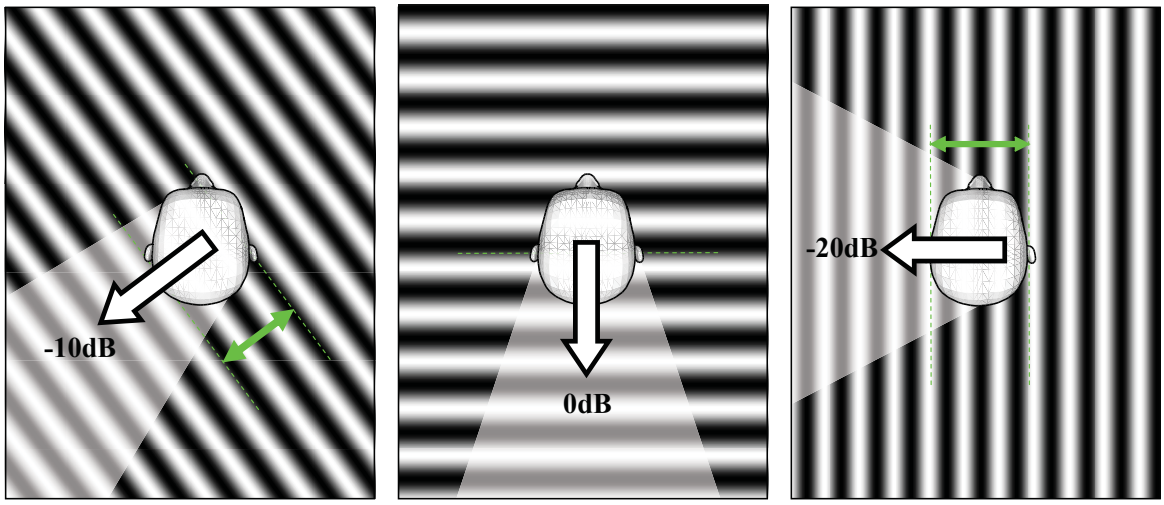
2.2.4.1 Interaural Level Difference

ILD can be described in terms of the ratio of the sound pressure levels at the ears by

$$ILD = 20 \log \left| \frac{P_l(f)}{P_r(f)} \right| \tag{2.7}$$

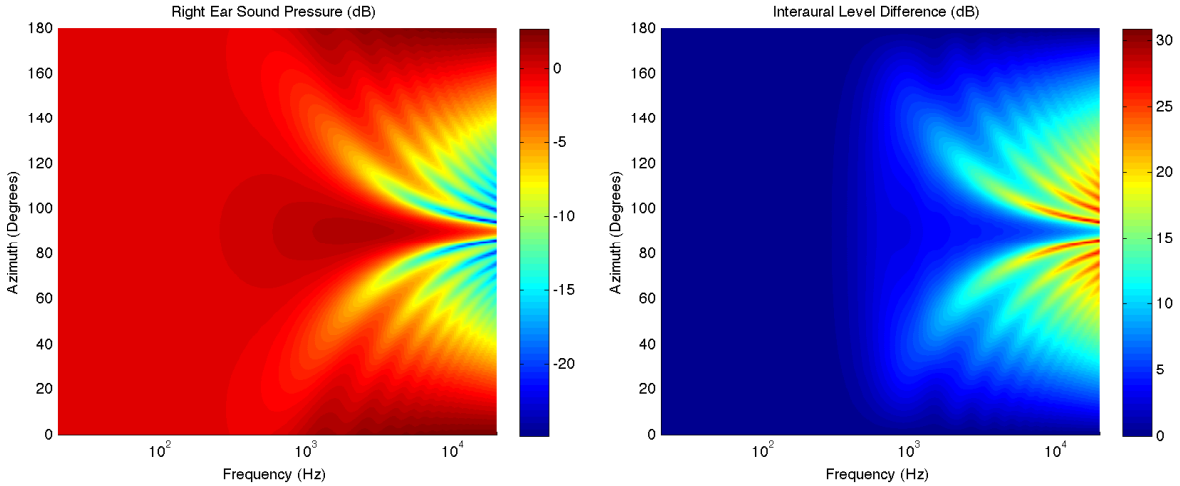
where $P_l(f)$ and $P_r(f)$ are the Fourier domain representation of the sound pressure measured at left and right ears respectively. ILDs arise due to the shadowing effects of the head for spectral components above approximately 500Hz. Below this frequency, the source wavelengths are comparable to the size of the head, and diffraction and head shadowing no longer occur. The effect of ILD is illustrated in Figure 2.8. As the source position changes, so too does the interaural level difference. Maximum ILD occurs for sources to the side of the head with values reaching approximately 20dB.

A simple computational spherical head model was adapted to obtain an estimate of the size of the ILD with respect to frequency. This was achieved by calculating the sound intensity at opposite points of the surface of a sphere of diameter 18cm (given an incident plane wave). The 'earpoints' were located at $\pm 90°$ from the front of the sphere. Such an approach has been previously adopted by different authors for estimation of HRTFs, for example in [Rabinowitz et al., 1993], [Brungart and Rabinowitz, 1999], and [Duda and Martens, 1998]. The sound pressure



(a) (b) (c)

Figure 2.8: *ILD changes due to incident plane waves at varying angles: (a) Plane wave at 45°, (b) Plane wave at 0° and (c) Plane wave at 90°.*



(a) (b)

Figure 2.9: *Sound pressure level at the ears: (a) Right ear sound pressure in relation to sound pressure at centre of sphere with sphere removed, based on a spherical head model with ear points ±90° relative to centre front and (b) ILD as a function of frequency and source angle*

level at the ear varies with frequency, as shown in Figure 2.9(a), whilst the interaural attenuation also varies with azimuthal angle, as shown in Figure 2.9(b). Significant head shadowing effects in the ILD can be seen around 4kHz. Below 1.5kHz the ILD reduces significantly, making it difficult to localize low frequency signals solely based on the ILD cue.

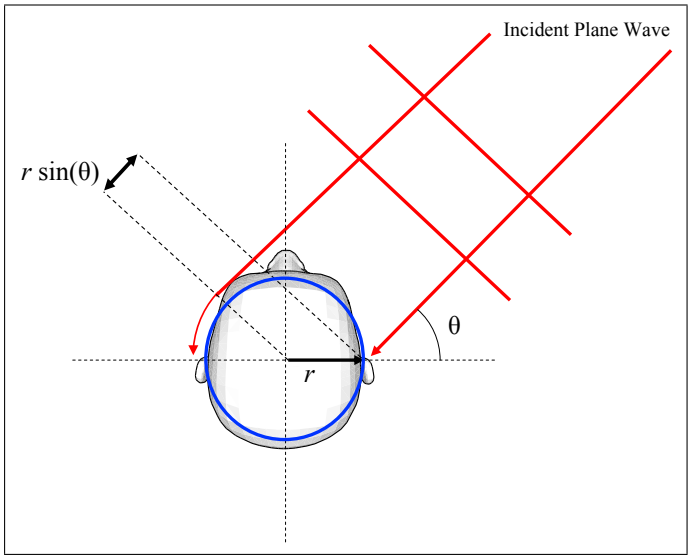


Figure 2.10: *Source angle dependence of the ITD.*

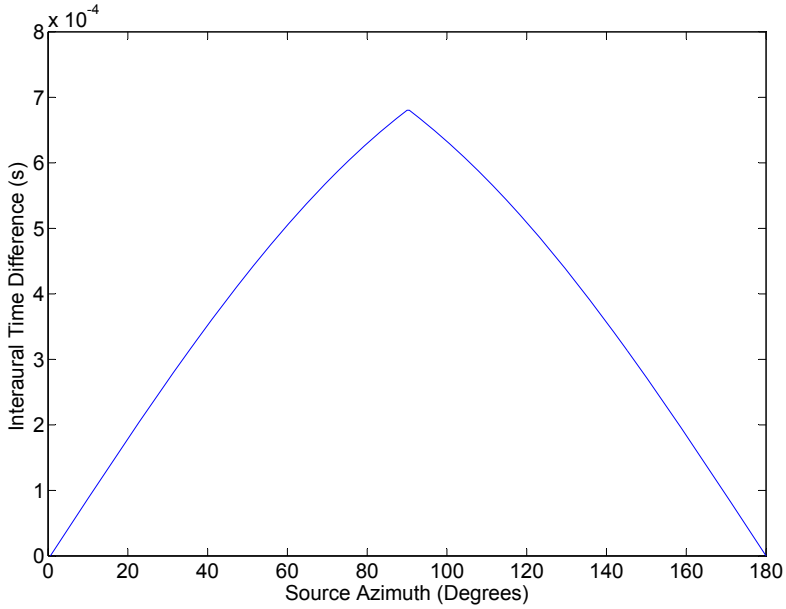
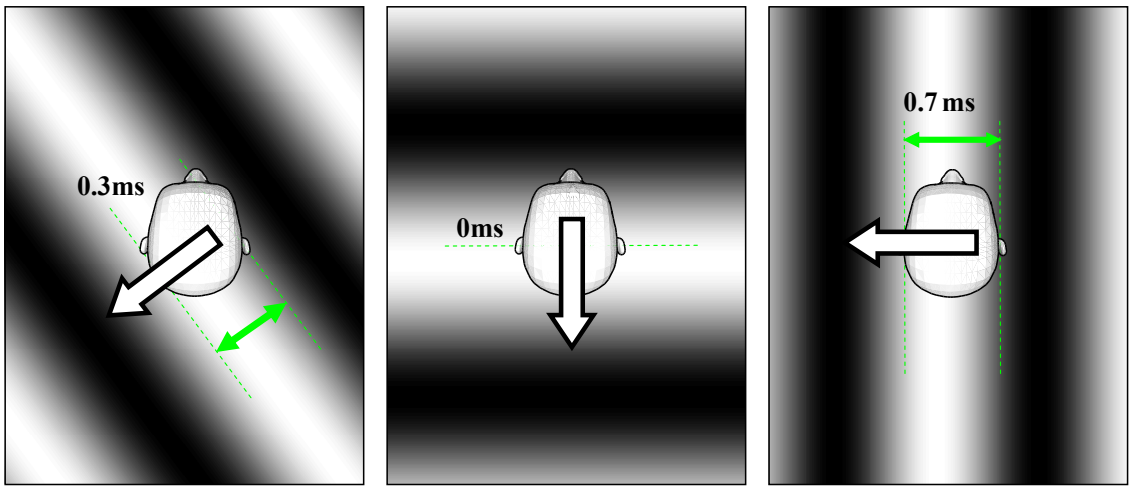


Figure 2.11: *Interaural Time Difference with varying source azimuth. The ITD reaches its maximum for sources at side of the head.*

2.2.4.2 Interaural Time Difference

The ITD describes the time difference between the same wavefront arriving at each ear from a given relative source-head angle. ITD is particularly relevant for wavelengths greater than the size of the head (< 700Hz) and under the low frequency assumption, can be modeled using the simple circular head model (with ear points at $\pm 90°$) shown in Figure 2.10. As the head rotates



(a) (b) (c)

Figure 2.12: *ITD changes due to incident plane waves at varying angles: (a)* 45°*, (b)* 0° *and (c)* 90°*.*

(or the angle of the source is changed), a path length difference occurs between the source and the ears. Here the ITD can be expressed as

$$ITD = \frac{r(\theta + \sin\theta)}{c} \tag{2.8}$$

where c is the speed of sound (approx. 343m/s), r is the average radius of the human head, taken as 9cm approximately, and θ is in radians. The resultant ITD with varying source azimuth is documented in Figure 2.11. We see that the ITD is maximum for sources at the side of the head, where the interaural delay is in the region of 0.65ms -1ms. For sources to the front (0°) and rear (180°) the ITD becomes zero. This physiological interaction is illustrated in Figure 2.12.

At low frequencies, i.e. below 700Hz, the ITD is particularly relevant. Here, the ILD is very low since the wavelengths are greater than the size of the head and shadowing effects are negligible. ITD cues are particularly strong for sounds with sharp attacks or transients, resulting in broader spectral content. This was shown by Klumpp and Eady [Klumpp and Eady, 1956] who investigated the just noticeable difference (JND) for source movement (lateralization blur) using artificially generated ITDs for different broadband noise sources. They found that source movement is detectable in the 150 to 1700Hz band with only 9μs delay. This becomes larger for higher frequency bands, such as in the 2400 to 3400 band, where the just noticeable difference is 62μs as well as when the bandwidth is reduced. Furthermore, lateralization blur from ITDs decreases when the duration and level of the signal increases [Klumpp and Eady, 1956]. This is attributed to an increased number of elicited nerve impulses [Blauert, 2003].

Steady state (continuous) sounds can also be localized through ITD, although this is only achievable up to around 700 to 1200Hz. Above this, the phase of each of the ear signals becomes

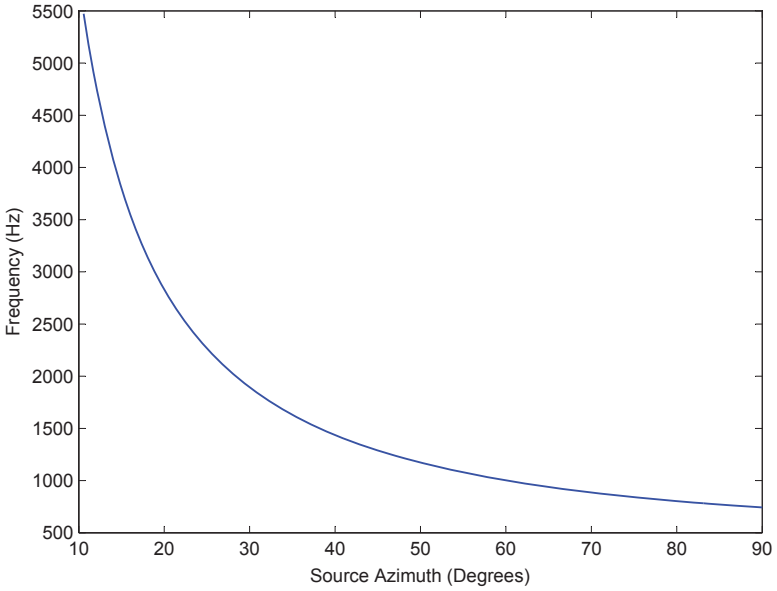


Figure 2.13: *Upper frequency limit of phase ambiguity.*

ambiguous, and it is impossible to tell the correct delay between the wavefronts. This is due to the fact that the basilar membrane decomposes the incoming waves into spectral components, and when we attempt to localize sounds, we are in fact analyzing spectral pairs at different frequencies. Thus, the phase for signals with wavelengths shorter than the size of the head, may have gone through several cycles before reaching the ear. In this case, the signal becomes ambiguous, and interaural phase is no longer a reliable cue.

However a simple analytical analysis with a circular head model shows us that these ambiguities are largely directional dependent. For example, consider the case of Figure 2.10 with a sound source at $\theta = 45^\circ$ azimuth to a listener. The path length difference between the source and each ear can be approximated from equation 2.8 by

$$d = r\theta + r\sin(\theta) \tag{2.9}$$

For a 45° angle, with $r = 0.09$m, we then have

$$d = 0.09\left(\frac{\pi}{4}\right) + 0.09 \times \sin\left(\frac{\pi}{4}\right) \approx 0.135\text{m}$$

The resulting ITD (for $c = 343$m/s) is approximately 0.4ms. We therefore have unambiguous ITD information up to approximately 1.2kHz, i.e.

$$f = \frac{c}{\lambda} = \frac{343}{2 \times 0.135\text{m}} = 1.270\text{kHz} \tag{2.10}$$

If we repeat this calculation for all azimuths we see that the limiting frequency for unambiguous ITDs decays as the angle subtends towards the ear. This is shown in Figure 2.13 for 10° to 90°. We see that the worst case scenario occurs for the side of the head with the frequency limit reaching approximately 700Hz.

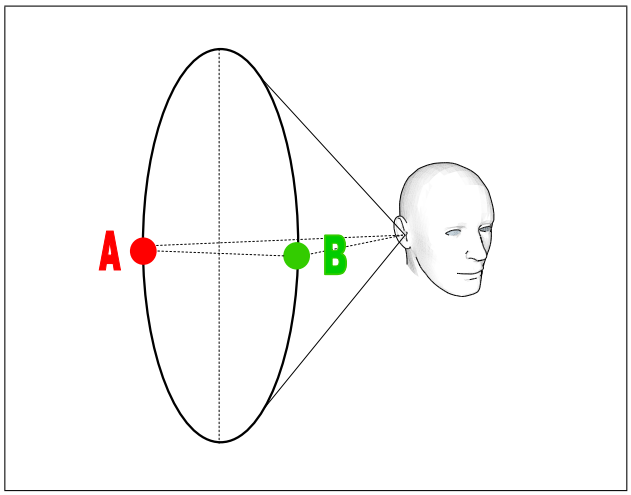


Figure 2.14: *The 'Cone of Confusion' where source positions A and B (and all other points on the cone) have the same ITD and ILD values.*

2.2.4.3 Monaural Spectral Cues

Our ability to localize sounds is strongest when the different auditory cues are non-contradictory, i.e. when the ITD and ILD indicate the same direction. However, situations occur when the ITDs and ILDs are non-contradictory and the direction of localization is still ambiguous. In fact, for each ITD and ILD pair, there exists a 'cone-of-confusion' where the same ILD and ITD values can give rise to multiple source positions as shown in Figure 2.14.

For real world sounds there is an important difference between the source wavefronts that arrive from behind us in comparison to those in front, namely the spectral signature due to the filtering of the pinnae. We must bear in mind that asides from proprioception (i.e. neural feedback from ear muscles), we also use head orientation (derived from the vestibular system) to detect the location of sound sources. When we do this, we alter not only ITD and ILD, but also the monaural spectral cue. In this way, the front-back ambiguities that arise in ITD and ILD can be overcome. This interaction with the ear causes pinnae reflections, which result in spectral variations that depend on source angle. What is more, since this filtering is dependent on the shape of the pinnae, it is dependent on the individual listener, i.e. the spectral variations required to overcome front-back confusion in one listener may not work with another listener.

Such spectral differences amongst individuals can be seen in Figure 2.15. Here, the spectral deviations for a source at an azimuthal angle of 45°, but with varying elevation are shown for 4 subjects from the CIPIC HRTF database [Algazi et al., 2001]. The readers attention is drawn to the changes in sound pressure level that occur particularly in the 2kHz and 7-8kHz regions.

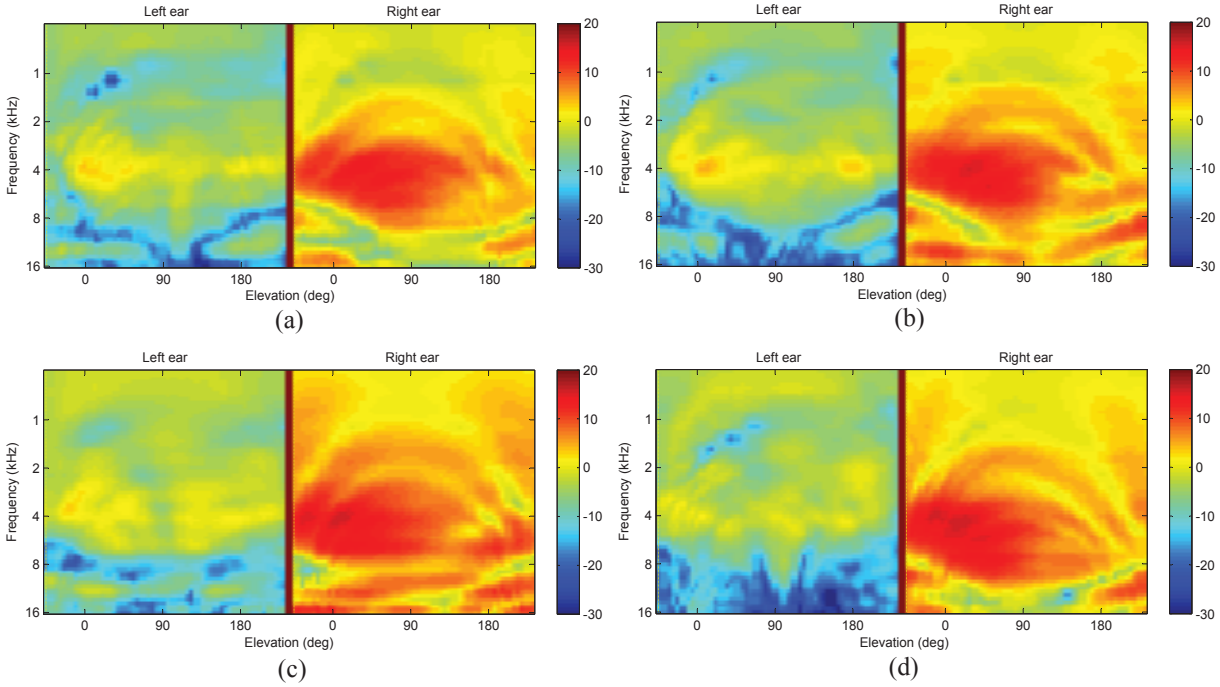
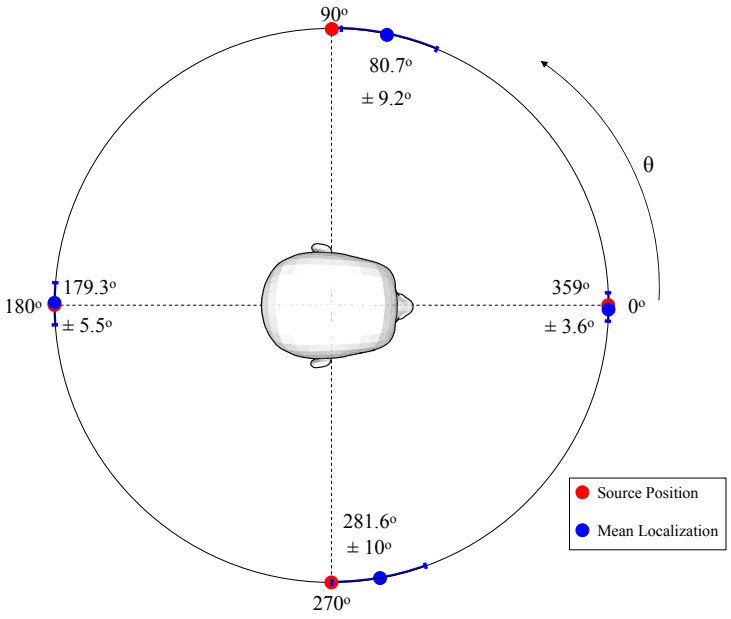


Figure 2.15: *HRTFs of 4 different subjects extracted from CIPIC HRTF database [Algazi et al., 2001] with source at azimuthal angle of* 45° *and changing elevation (a) Subject 3, (b) Subject 8, (c) Subject 9, (d) Subject 21.*

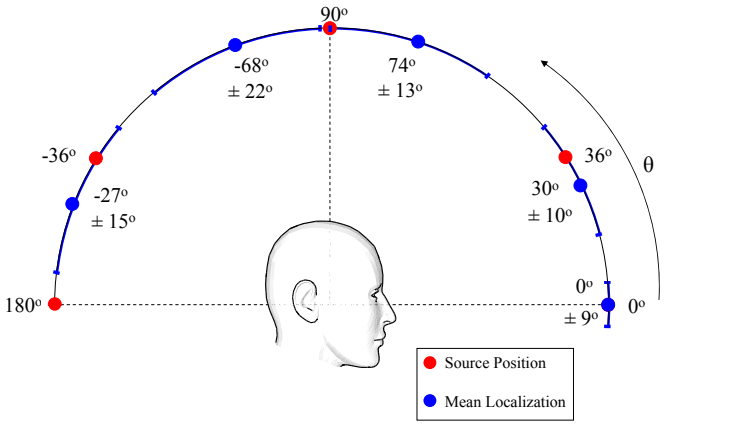
2.2.5 Localization Blur

Given all these cues it may be surprising to know that we do not localize sounds equally well in all directions. Large scale localization experiments in the azimuthal plane have been conducted in [Preibish-Effenberger, 1966] and [Haustein and Schirmer, 1970], using 600 and 900 untrained subjects respectively. The subjects used an 'acoustic pointer' i.e. a movable loudspeaker, to indicate the direction of the source. The source stimulus used was 100ms white noise bursts. The deviations from the intended source position are shown in Figure 2.16(a). We see that in the forward direction, the localization blur is in the region of ±3.6°, with a larger deviation from behind of approximately ±5.5°. However, localization blur is greatest at the sides of the listener, with deviation values of approximately ±10°. It is interesting to note that the mean localization is also approximately 10° offset from its intended position for sources directly at the sides. According to [Blauert, 2003], it is unclear whether the aforementioned deviations are a result of systematic error, or that they reflect true localization blur. However, similar results have also been found by Wilkens [Wilkens, 1972] and de Veer [Van de Veer, 1957].

Directional hearing is quite different in the median plane, where localization is based primarily on spectral cues, since the ITDs and ILDs remain constant. Figure 2.16(b) shows the results of median plane localization experiments conducted by Damaske and Wagener [Damaske and Wagener, 1969] for continuous speech. The localization blur due to changes in source elevation



(a)



(b)

Figure 2.16: *Localization Blur, after [Preibish-Effenberger, 1966], [Haustein and Schirmer, 1970] and [Blauert, 2003] . (a) Azimuthal localization blur, (b) Elevation localization blur.*

---

is approximately 9° for speech of a familiar person.

2.2.6 Distance Perception in a Free Field

Aside from our ability to localize sounds horizontally and vertically, we are also capable of discriminating the distance at which a source is placed. Whilst this ability is not fully understood, there are several key factors which are known to contribute to distance perception. We have

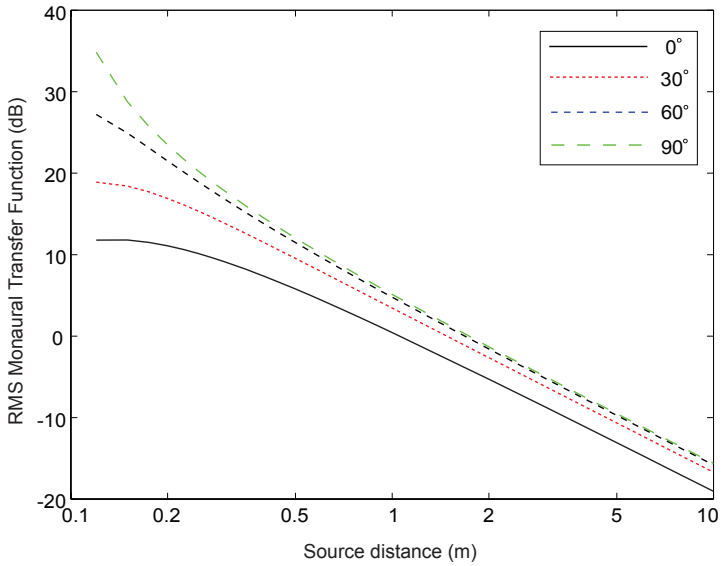


Figure 2.17: *RMS monaural transfer function at the left ear for broadband source at different angles with varying source distance (reference = plane wave at* $(0°, 0°)$*).*

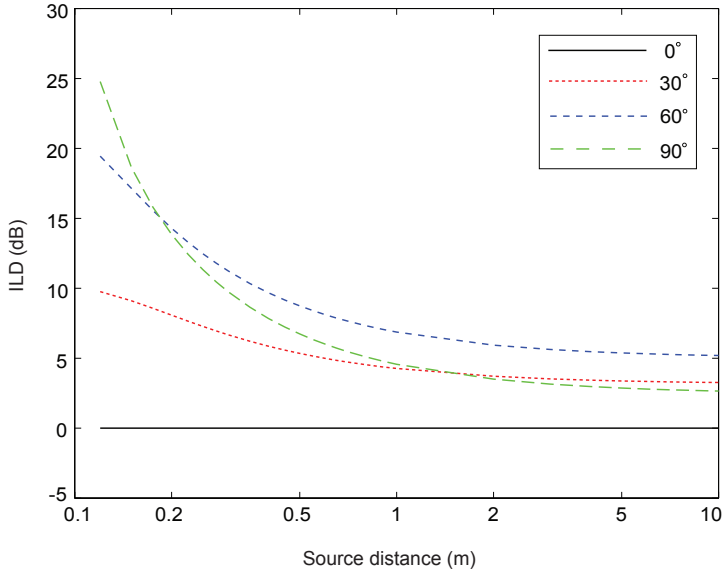


Figure 2.18: *Interaural level difference of spherical head for broadband source at different angles with varying source distance.*

already shown that as the source propagates in a free field, its level drops 6dB per doubling of distance. Thus, changes in distance also lead to changes in the monaural transfer function. This is shown in Figure 2.17 for a spherical head. We see that for sources of less than 1m distance, the sound pressure level varies depending on the angle of incidence, due to the shadowing effects of the head. Beyond 1m however, the intensity of the source decays according to the inverse square law.

It is interesting to note that for sources in the median plane, the level at distances less than 1m does not change as dramatically as sources located at the ipsilateral point. This will not affect the low frequency ITDs, but it is reflected in the ILDs as shown in Figure 2.18. We note that the most extreme ILD is exhibited at the side of the head (90°), due to the maximum head shadowing effect. Thus, near-field ILD cues exist which aid us to discriminating source distance. However, these cues will only work if we have some prior knowledge of the source level i.e., how familiar we are with the source. In other words, a form of *semiosis* occurs, where the perception of localization is based on anticipation and experience [Blauert, 2008]. For example, for normal level speech (approximately 60dB at 1m), we expect nearer sources to be loud, and quieter sources further away. However, this is more difficult to assess for synthetic sounds or sounds that we are unfamiliar with. Air absorption also contributes to such distance dependent filtering, where sources further away will sound low pass filtered. Analytical expressions defining atmospheric absorption can be found in [Bass et al., 1972], and digital filter implementation of same in [Huopaniemi et al., 1997].

Despite such cues, distance perception and in particular 'externalization' (i.e. out of head localization) is significantly aided by room acoustic reflections and in particular, the ratio of the direct to reverberant sound. This is discussed in detail in Section 2.3.3.

2.2.7 Free-Field Localization Summary

We have seen how monaural cues contribute to the perception of height and, when coupled with head-movement, how they allow us to discriminate between front and back sources, due to the spectral signature. It has also been shown that the ITD and ILD are highly frequency dependent and that this dependence manifests itself physically when dealing with the size of the human head. For a low frequency sinusoidal signal, such as that shown in Figure 2.19(a), where $\frac{\lambda}{2}$ is greater than the diameter of the head, the wavelength is so large that the resulting ILD is practically zero. However, the phase difference between the signals at the ears results in measurable ITDs. Conversely, a higher frequency signal, such as that shown in Figure 2.19(c) loses a significant amount of energy due to the shadowing effects of the head, resulting in a large ILD. Figure 2.19(b) shows the point where the wavelength of the sinusoid becomes comparable to the size of head (approx. 1.5kHz). Here the ITD becomes less reliable and ambiguous and the phase difference no longer corresponds to the unique source location. However, at higher frequencies, the head begins to attenuate the sound, and less energy arrives at the shadowed ear than at the un-shadowed ear. ITDs and ILDs are useful in allowing us to numerically assess the location of a particular sound source in relation to the head. They allow us to measure the localization accuracy of a particular sound system in relation to real sources, as well as the localization accuracy in reverberant rooms when compared to free-field conditions. We will now investigate the latter situation in detail.

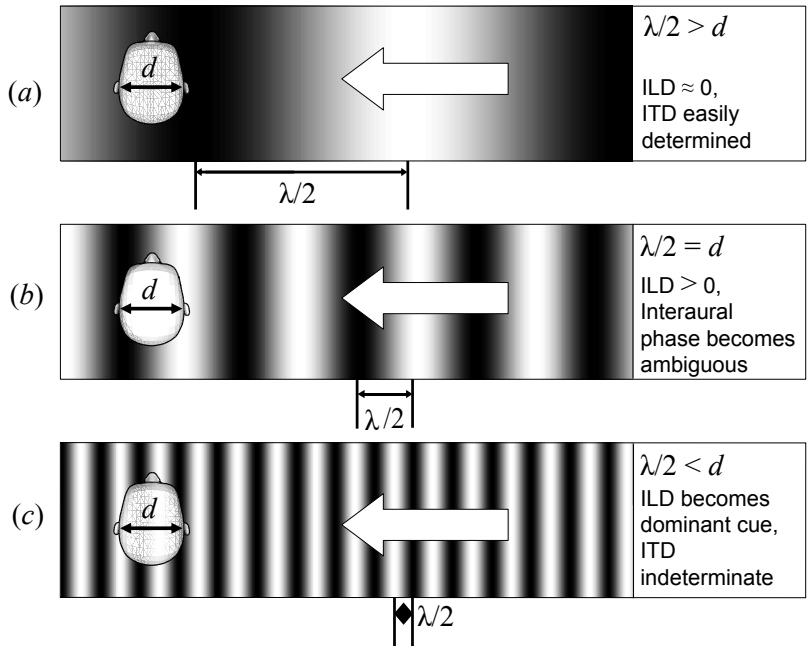


Figure 2.19: *Summary of interaural transfer function cues.*

2.3 Sound Localization in a Reverberant Field

In the previous section, we analyzed the perception of a single source in non-reflective conditions. Now we will generalize this study to consider the perception of multiple sound sources in enclosed spaces. From this analysis we will outline several physical acoustic measures which pertain to the work in this thesis, as well as the derivation of a new measure for enhanced source localization using binaural microphones. We will begin by first considering the effect of the room itself.

2.3.1 Source Propagation in a Reverberant Field

It is well known that in enclosed spaces a proportion of the sound radiated from a sound source is reflected from the surfaces of the environment, and the rest absorbed by the surfaces or by air. Thus, the application of transient stimulus in any given reverberant space will create a soundfield that will change at the listeners ears from one that is coherent, to partially coherent, to non-coherent. Such propagation of sound can be described as a linear-time invariant system, where the resultant ear signals are a superposition of the direct sound and reflections from the room boundaries. The number of times a source wavefront reflects off a boundary is known as the reflection order, illustrated in Figure 2.20. The density of the reflections per unit of time increases as the square of the elapsed time after the sound source radiates the pulse or

$$\text{no. of reflections/s} = \frac{4\pi c^3}{V} t^2 \tag{2.11}$$

where t is the measurement interval, c is the speed of sound (343m/s) and V is the volume of the room [Kuttruff, 1979]. Eventually, the room reflections become so dense that a diffuse

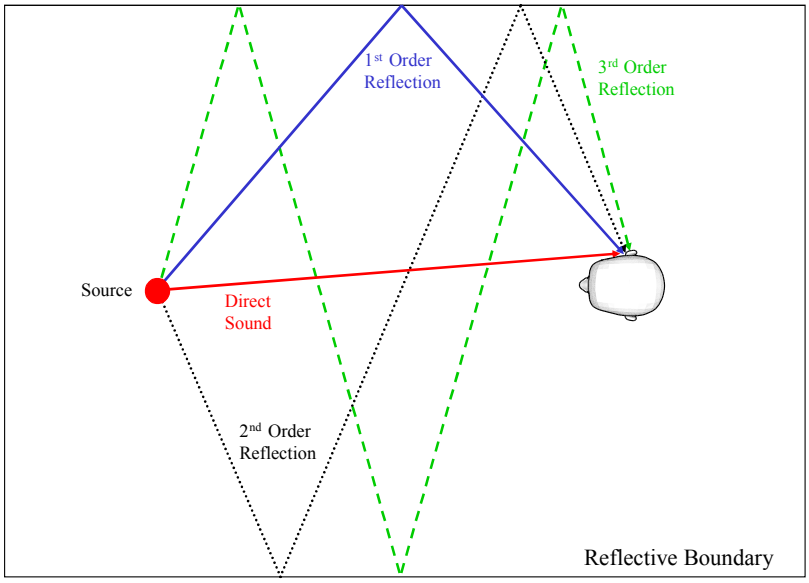


Figure 2.20: *Impinging reflections of different orders at the right ear.*

reverberant field is built up, and reflection density can only be described by statistical signal theory.

The resultant linear superposition of the direct sound and room reflections is known as the acoustic Room Impulse Response (RIR). This is the response of the room to a Dirac delta function $\delta(t)$ stimulus at particular source and receiver points. If the receiver is a binaural mannequin (or alternatively a human subject), with microphones at the ear canals then the resultant impulse responses are known as Binaural Room Impulse Responses (BRIRs). Such impulse responses will have three important regions of interest; the direct sound, which is due to the direct path between the source and receiver, the early reflections due to the wavefronts that have propagated and reflected off between 1 to approximately 4 surfaces and have arrived at the measurement position, and the diffuse decay which contains a higher reflection density and can generally be considered stochastic. The sound field is fully diffuse when the mean squared sound pressure is uniform and where any reflection angle is equally probable. This field exists when the reflection density becomes so high that the impulse response can be considered to become stochastic.

The reverberation time (considered the dominating acoustic characteristic), RT_{60}, is the time it takes an impulsive sound to decay by 60dB. Sabine [Sabine, 1964] has presented an equation for a room's reverberation time as

$$RT_{60} = 0.163 \frac{V}{\sum a_k S_k} \tag{2.12}$$

where V is the room volume and a_k and S_k are the frequency dependent absorption coefficient and surface area of the k^{th} room boundary, respectively. When the reverberation time is extremely short the room is said to be acoustically 'dead'. Only in the near vicinity of the sound

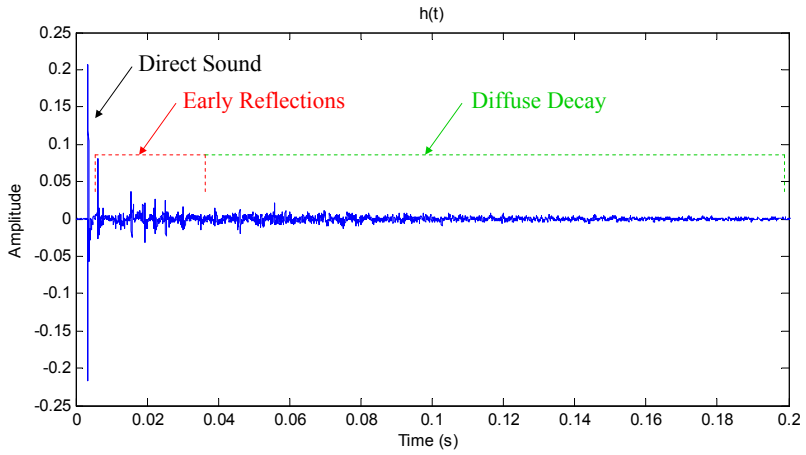


Figure 2.21: *Components of a room impulse response: The direct sound is followed by sparse early reflections and the diffuse decay.*

	Volume (m^3)	RT_{60} (Occupied)	D_{Crit}(m)
Vienna, Gr. Musikveceinssaal	15000	2.0	4.94
Amsterdam, Concertgebouw	18780	2.0	5.52
Boston, Symphony Hall	18750	1.85	5.74
Washington, Kennedy Centre	19300	1.8	5.9
Berlin, Philharmonic Hall	26000	1.95	6.58
London, Royal Albert Hall	86650	2.4	10.83

Table 2.1: *Critical Distance of several world famous concert halls [Okano et al., 1998].*

source does the level of the primary sound exceed that of the reflections. The point at which the direct and reverberant sound pressure levels are equal is know as the *critical distance*, given by

$$D_{\text{Crit}} = 0.057\sqrt{\frac{V}{\text{RT}_{60}}} \tag{2.13}$$

where V is the volume of the room in m^3 and RT_{60} is the reverberation time. Typical critical distance values for world-famous music halls are shown in Table 2.1. Reverberation times and room volumes are taken from [Okano et al., 1998].

The localisation of transient signals within reverberant environments has been investigated by Hartmann [Rakerd and Hartmann, 1985]. He found that localization of sound sources did not change when the room reverberation time was increased from 1 second to 5 seconds. However, it was found that the room geometry had a significant effect on the initial reflections (<30ms), and this played an important role in successful localization. However, even in the presence of strong reflections, it is the first arriving wavefront which dominates the localization of the sound. This is known as the 'Precedence Effect', or 'Law of the First Wavefront'.

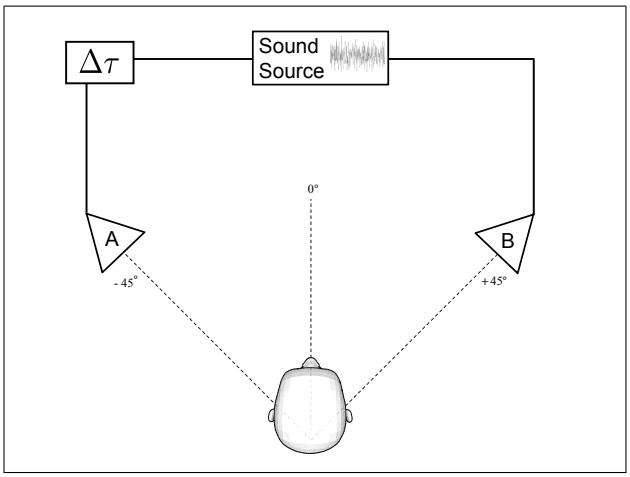


Figure 2.22: *Stimulus configuration for localization discrimination experiment.*

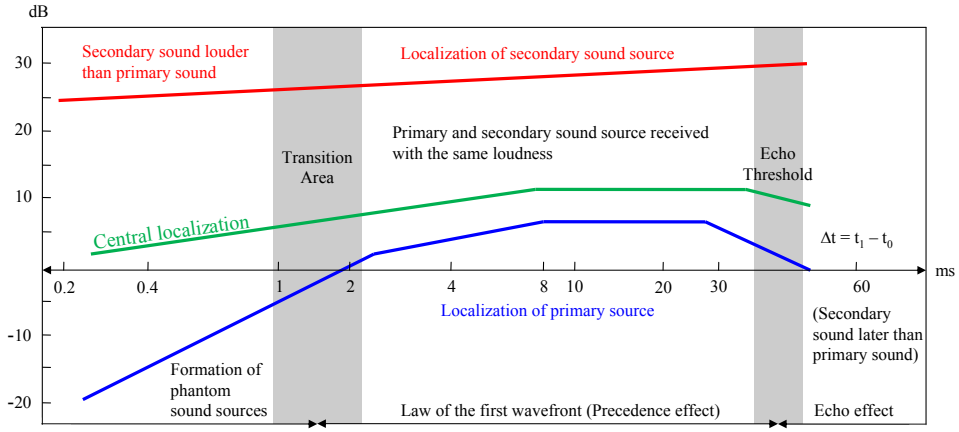


Figure 2.23: *Validity domain of the 'Law of the First Wavefront', after [Hoeg et al., 1983].*

2.3.2 The Precedence Effect

As a demonstration of the 'Precedence Effect', consider the loudspeaker arrangement shown in Figure 2.22. Both loudspeakers are in an anechoic environment, and the same source stimulus is applied to each. Loudspeaker A, however, is delayed by τ seconds. Thus, we can consider this a simplified model of a direct sound (loudspeaker B) with a single reflection (loudspeaker A). This setup is useful for analyzing the perceived location for fused sounds[1] or early reflections.

When the lag time τ is zero, the perceived location of the source of the sound is at 0° in the azimuthal plane. As the lag time is increased a 'phantom source' is formed, and the perceived sound moves towards loudspeaker B, reaching its location between 0.65ms and 1ms. For delays between 1 and 5ms, it seems as if only one sound occurs, not two, and it is located in the direction of the leading speaker (between 0° and +45° in the case of Figure 2.22). This is known

[1] We will consider a 'fused' sound to be two sounds played within 20ms of each other.

as 'summing localization', as the two sounds are effectively summed together at the ear canal to give the direction of localization (this is explained further in the context of two-channel stereo in Section 3.3). This defines the lower boundary of the range of delay times for which the 'Law of the First Wavefront' is applicable. Further explanation of this phenomenon can be seen in Figure 2.23 [Hoeg et al., 1983], illustrating the perceptual effects of incorporating delay and SPL differences between the two loudspeakers. It shows the upper boundary of the precedence effect at around 30ms. However, this is highly dependent on the nature of the source sound, the delay time, the level of the sound, and the angle of incidence. Blauert [Blauert, 2003] defines the threshold at which the reflection becomes "absolutely audible" as the point where one can begin to notice changes in the spatial width of the sound, the tone and the perception of the auditory event shifting towards the reflection. It is the point at which we *detect* the presence of the lag signal.

In its simplest form, a reflection is an echo of the source. The delay at which we can hear this echo (absolutely) is known as the 'echo threshold'. It is important to note that the echo threshold is not the detection of the lag, but rather an estimation of the delay at which the perceptually fused sound splits into two separate entities. At short delays, 'localization dominance' occurs. This is where the leading source dominates the image location due to fact that it is masking the stimulus parameters (such as the onset) of the lag source. It can also be found that the precedence effect holds true when the direct sound is applied solely to one ear, and the reflected sound to the other. Again, the lower boundary lies somewhere between 630μs and 1ms, and the upper boundary is determined by the echo threshold. This is known as the 'Haas Zone', due to its description by Helmut Haas [Haas, 1972]. Thus, the reception of a signal in one ear inhibiting the signal in the other ear indicates that contralateral inhibitory processes are being used. In both cases the echo threshold is larger for steady-state signals than for impulses. Also, high frequency signal components lead to shorter delay times, and thus shorter echo thresholds. Figure 2.23 also shows the effects of level difference between the direct and reflected sound. The bottom area of the graph (below the blue line) illustrates the SPL difference for correct localization at the leading speaker. Conversely, the upper shaded area of the graph shows the level at which suppression of the primary sound will occur and localization found at the delayed loudspeaker. This is the 'backward masking' effect, previously introduced in Section 2.2.3.

This analysis of the precedence effect has considered a single auditory event and a single reflection. Whilst this is useful in the analysis of localization, it does not reflect true soundfield propagation in a reverberant room. We will now generalize our investigation to consider multiple reflections and their impact on source localization, depth perception and envelopment.

2.3.3 Distance Perception in Reverberant Environments

We have already identified several key factors in the discrimination of source distances in a free-field, namely the source loudness and the ILD. However, in reverberant rooms, the ratio of

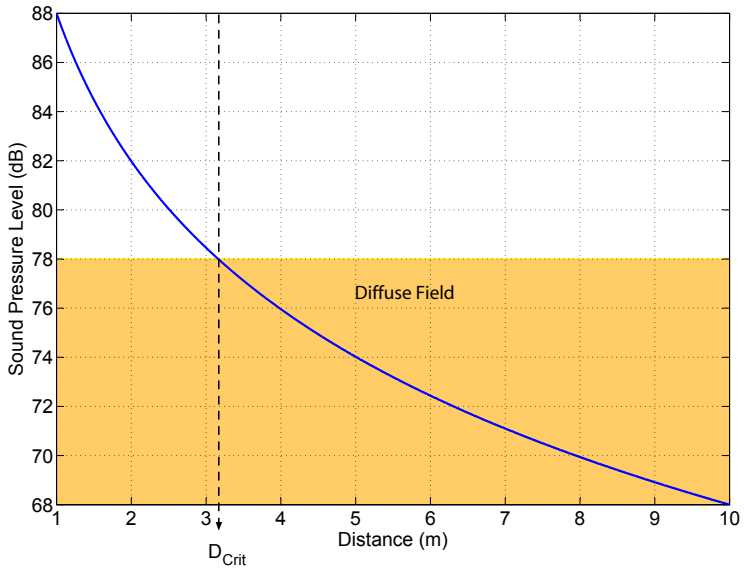


Figure 2.24: *Level of direct sound with respect to a diffuse field as a function of distance.*

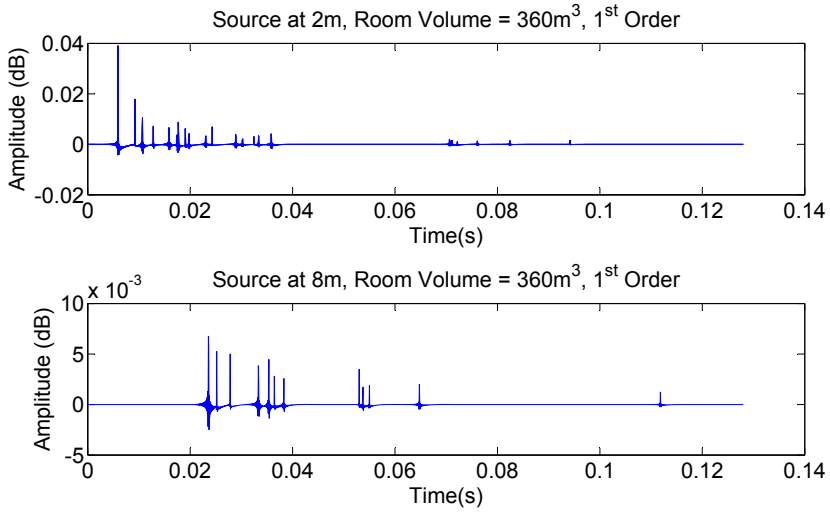


Figure 2.25: *Comparison of impulse responses at source-listener distances of 2m and 8m respectively using 2-D image method simulation.*

---

the direct to reverberant sound plays an extremely important part in distance perception. For near sources, where the direct field energy is much greater than the reverberant field, the level approximately changes in accordance to the free-field conditions. However, for source-listener distances greater than the critical distance, the level of reverberation is in general independent of the source position due to the homogeneous level of the diffuse field. Thus, at distances greater than 1m the direct to reverberant ratio changes approximately 6dB per doubling of distance as shown in Figure 2.24.

However, the level and time of arrival of the early reflections is of extreme importance. The

closer the sound is to the listener, the greater the initial time delay gap between the direct sound and first reflections. This is shown in Figure 2.25. Here we compare the level of the direct sound and first order reflections for a sources located at 2m, and also at 8m from the receiver in a simple 2-D mirror image room simulation. We clearly see that the initial time delay gap decreases with the 8m case, and that the level of direct sound has become comparable to that of the (decreased) first order reflections. Thus the dominating acoustic characteristics in distance perception can be viewed as an accumulation of near-field cues such as ILD and far-field cues such as the arrival-time/level of early reflections and the direct to reverberant ratio. Consequently as the source moves further into the reverberant field, the level of reflections will also affect the perceived source width. This is now discussed in detail.

2.3.4 Apparent Source Width in Reverberant Rooms

The Apparent Source Width (ASW) is the auditory width of the soundfield created by a source as perceived at a particular listener position. It is attributed to the level and time of arrival of the early reflections. It has been shown in [Keet, 1968] that the auditory system utilizes a correlation process in determining the source width, related to the level of coherency of the signals received at each ear. This is quantified using the Interaural Cross Correlation Function (IACF) which is defined as,

$$\mathrm{IACF}(\tau) = \frac{\int_{t_1}^{t_2} x_1(t)\ x_2(t+\tau)\ dt}{\sqrt{\int_{t_1}^{t_2} x_1^2(t)\ dt \int_{t_1}^{t_2} x_2^2(t)\ dt}} \tag{2.14}$$

The IACF is a function in the range [-1,1] and gives a measure of the correlation between the received signals in the integration limits t_1 to t_2 as a function of the time delay τ. 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 [Okano et al., 1998]. In general, the higher the value of the IACC, the narrower the perceived source width. This relationship is demonstrated in Figure 2.26 [Keet, 1968]. Here, Keet modeled the head using two opposing cardioid microphones placed perpendicular to a sound source in order to measure the IACC. The results show that the IACC, measured within the first 50ms of the acoustic response for different sound pressure levels, is linearly related to the subjective source width. Furthermore, the data shows that the ASW widens approximately 1.5° for each 1dB increase in level. Hidaka et al. have further quantified the relationship between the IACC and ASW as

$$ASW = |1 - \mathrm{IACC}_{E_3}| \tag{2.15}$$

where E denotes 'early' and 3 represents the number of octave bands (500, 1000 and 2000Hz) used in the averaging [Hidaka et al., 1995]. Typical values for $|1 - \mathrm{IACC}_{E_3}|$ are taken from the early part of the impulse response (< 80 ms). The IACC is therefore useful in measuring

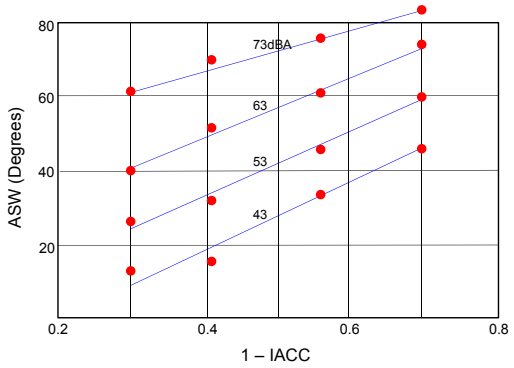


Figure 2.26: *Relationship between 1 – IACC (measured within first 50ms of response) and perceived ASW, with sound level (dBA) as a parameter, after [Keet, 1968].*

	Volume (m^3)	$\lvert 1 - \mathrm{IACC}_{E_3} \rvert$
Vienna, Gr. Musikveceinssaal	15000	0.64
Amsterdam Concertgebouw	18780	0.57
Boston, Symphony Hall	18750	0.65
Washington, Kennedy Centre	19300	0.60
Berlin, Philharmonic Hall	26000	0.46

Table 2.2: *ASW measured in several world famous concert halls [Hidaka et al., 1995].*

the acoustic quality of auditoria, and typical values of $\lvert 1 - \mathrm{IACC}_{E_3} \rvert$ are shown in Table 2.2 for well-known concert halls. Conventionally, this measure is implemented on sets of binaural impulse responses, but can also be used on other signals as well. For example, the IACC can also be used as a continuous measure over the length of the musical progression, as suggested by Mason [Mason et al., 2003]. The width of the main IACC lobe, $\mathrm{W}_{\mathrm{IACC}}$, can also be measured, as it is useful in detecting frequency dependent changes in the ASW when the IACC remains constant. This is defined as the interval of delay time 10% below the IACC peak. It is also informative to look at the late IACC values (from 80ms to ∞) which give insight into the diffuseness of the reproduced wavefield and have been shown to bear a strong relationship with the subjective impression of *envelopment* [Hidaka et al., 1995]. In this context, the three band IACC is known as IACC_{L_3}.

2.3.5 Properties of the IACC

The IACC is an extremely useful parameter, not only in determining the spatial width of the source, but also its location relative to the listener. This is possible since the time delay at which the IACC is maximum is representative of the ITD for the relative source-listener position, given

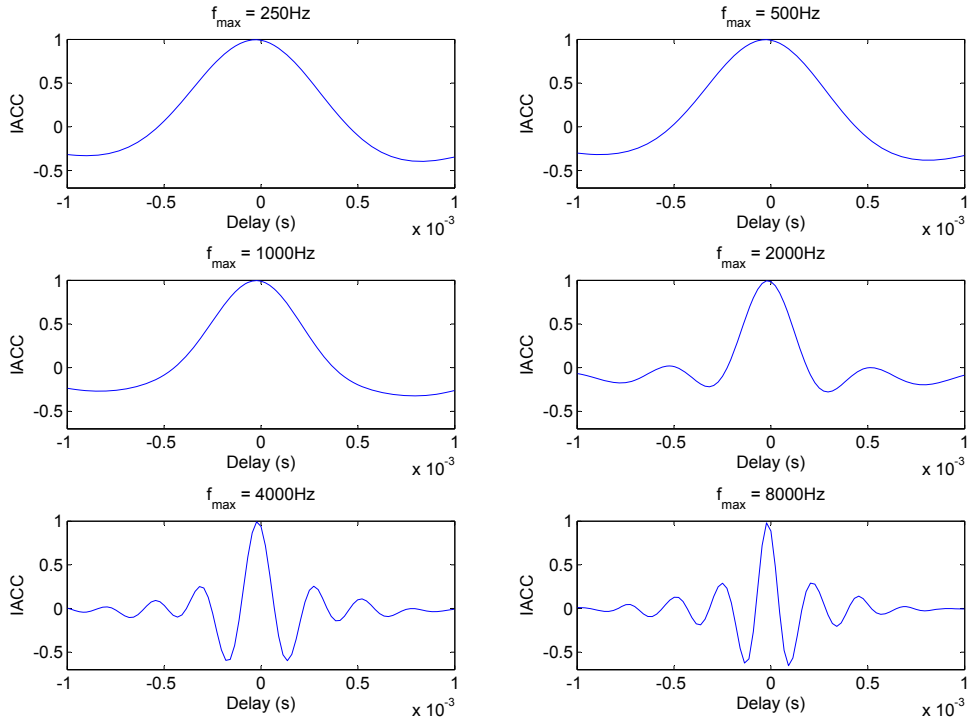


Figure 2.27: *Interaural Cross Correlation coefficient for KEMAR mannequin with source at (0°, 0°) with cumulative frequency.*

as

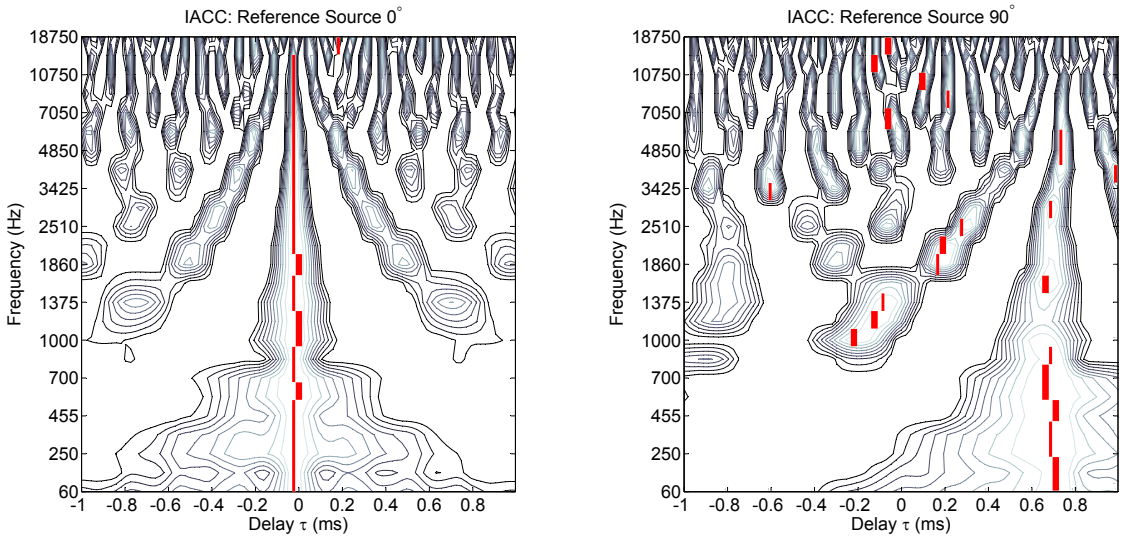
$$ITD = \arg\left(\max_{\tau} |IACF(\tau)|\right) \tag{2.16}$$

The IACC will change in different ways depending on three main factors:

1. The bandwidth of the source signal
2. The level and time of arrival of the reflections
3. The relative position of the source to the head

In the first case, the effect of increasing the source bandwidth on the IACF is demonstrated by the free-field IACC response of Figure 2.27. The responses shown are from a source located at 0° azimuth and elevation in front of a KEMAR mannequin[2]. It can be seen that the correlation for low frequencies results in a single broad main lobe. As we include frequencies whose wavelengths are smaller than the size of the head multiple lobes appear, and as a result it is often useful to visualize the IACC in frequency bands as shown in Figure 2.28(a). The red marks in the figures indicate the positive peak in the correlation, the ITD. We note that the although the ITDs in each band vary slightly, for a source located at 0° azimuth, the ITD is relatively consistent over all bands. However, this consistency is dependent on the angle of the source, which, as was

[2]KEMAR: Knowles Electronic Mannequin, a head and torso simulator made to the requirements of ANSI S3.36/ASA58-1985 [ANSI, 2006]



(a) (b)

Figure 2.28: *Full-spectrum IACC due to (a) source at* $0°$ *and (b) source at* $90°$ *(where accuracy degrades due to the frequency limit of phase ambiguity at around 800Hz).*

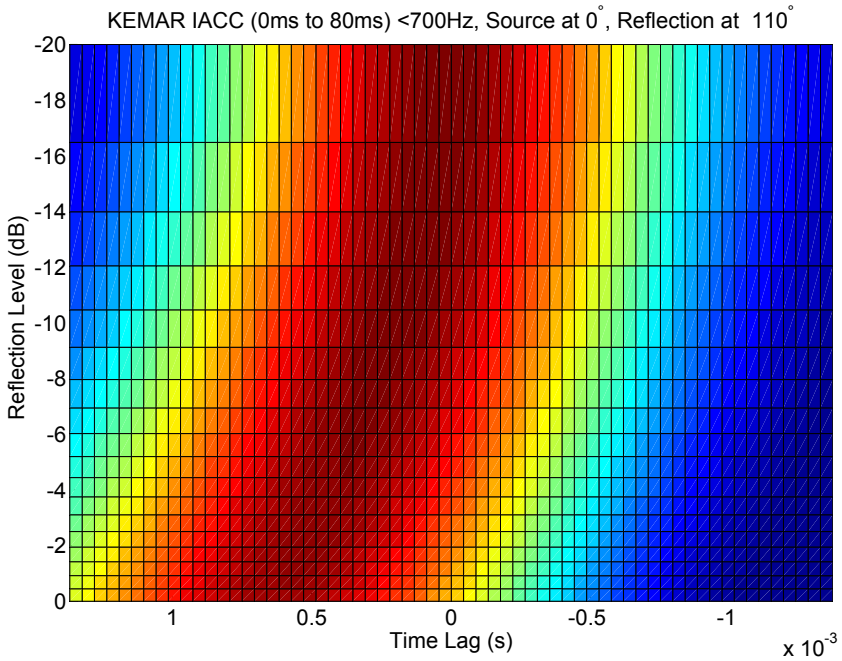


Figure 2.29: *IACC* ($< 700Hz$) *due to source at* $0°$ *with increasing level of broadband lateral reflection at* $110°$.

---

shown in Fig 2.13, dictates the frequency limit of phase ambiguity. For example, Figure 2.28(b) shows the case where a source at $90°$ yields consistent ITDs up to approximately 800Hz only. Beyond this point, the in-band ITD values are ambiguous.

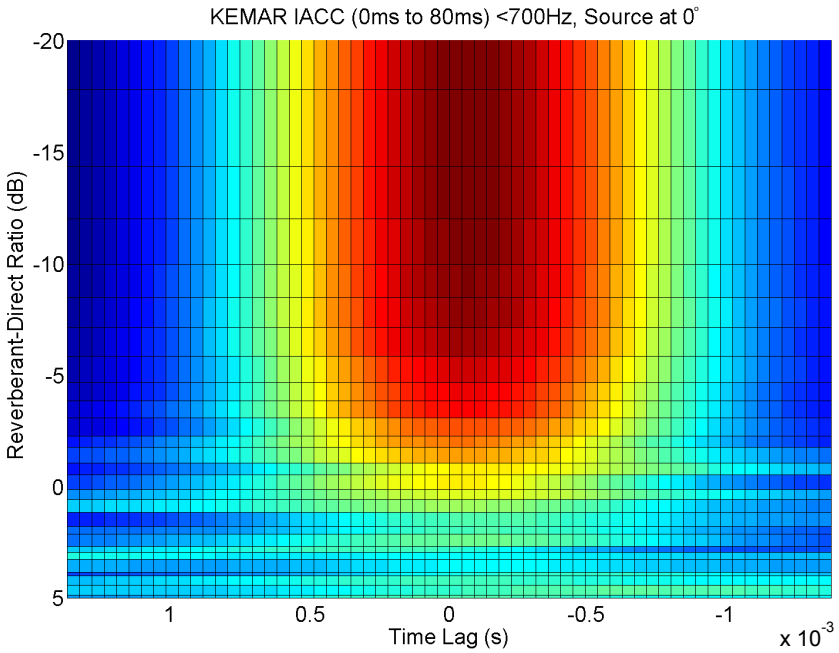


Figure 2.30: *IACC* < 700 *due to source at* $0°$ *with increasing level of diffuse field reverberation. The diffuse field is delayed by 30ms in the simulation. The plot shows the IACC values with increasing level of reverberation and hence a reverberant-to-direct field ratio is presented.*

In the second case, the level of reflections alters both the time-delay and the magnitude of the IACC. Consider the example shown in Figure 2.29 where a broadband lateral reflection (at 110° azimuth) is introduced to the ear signals at the same time as a broadband direct sound from 0° azimuth. These results show the IACC below 700Hz. We see that as the level of the lateral reflection is increased, the ITD shifts, indicating source movement in the direction of the reflection as expected. Thus the IACC is useful in determining source shifts due to extremely early reflections that occur within the cross-head delay.

Now let us consider the case where we wish to obtain estimates of the ITD in a more reverberant field. Figure 2.30 shows the effect of localization of a direct sound at 0° with an increasing level of diffuse field reverberation. In this simulation, the diffuse field reverb is delayed such that only the direct sound occurs in the first 30ms of the response. In this case, the precedence effect should hold, and the direct sound should be localized. However, this is not shown in the IACF model, and the accuracy of the ITD estimate falls significantly with the increased level of the reverberant energy. At a reverberant to direct ratio of 1 (0dB in the figure), the IACC indicates that the source is almost completely diffuse, contrary to what is perceived in real world scenarios. Thus, whilst the IACC gives good correlation between the subjective measures of spaciousness for early reflections, it does not take into consideration the precedence effect in ITD estimation i.e. reflections are given equal weight in the calculation. A method of improving the ITD estimate in this regard is shown in Section 2.4.

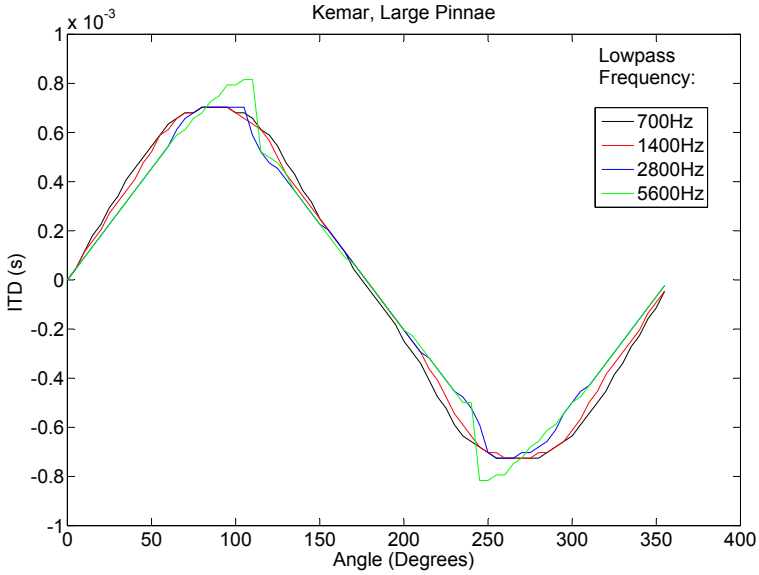


Figure 2.31: *Interaural Time Difference for KEMAR mannequin with increasing frequency and varying source azimuth.*

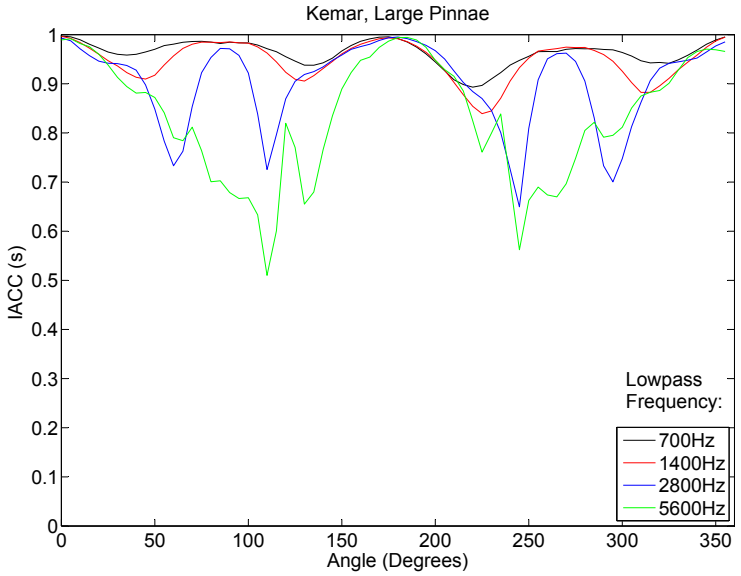


Figure 2.32: *Interaural Cross Correlation coefficient for KEMAR mannequin with increasing frequency and varying source azimuth.*

Finally, the ITD, estimated from the IACF will also change due to the relative position of the source to the head as shown in Figure 2.31. Here, we see the ITD values for all azimuthal source positions with cumulative frequency. It is interesting to note that as the bandwidth is increased, lobes in the ITD occur at the contralateral points. This can be clearly seen in the 5600Hz low pass filter curve. If we look at the magnitude of the correlation coefficient with varying source angle, shown in Figure 2.32, we see that these regions exhibit poor correlation. These are also

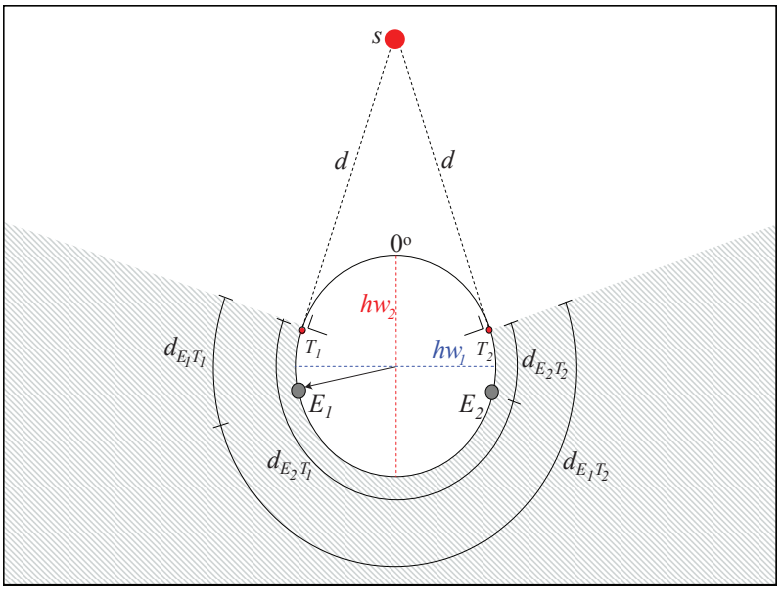


Figure 2.33: *Elliptical head model: Ear points are E_1 and E_2. hw_1 and hw_2 are the elliptical minor and major axis respectively.*

the angular regions where phase ambiguity begins at the lowest frequencies, previously shown in Figure 2.13. This effect is certainly due to high frequency content, but currently there exists no models to quantify this for full bandwidth signals. To this end, we will now attempt to derive such a model.

2.3.6 A Head Model for Broadband ITD to Angle Mapping

Let us consider an elliptical model of the head as shown in Figure 2.33, with ear points set at $\pm 100°$ from the $0°$ azimuth. The ear points E_1 and E_2 can be expressed in Cartesian form as,

$$E_1 = \left[\frac{hw_1}{2}\cos(\theta + 100), \frac{hw_2}{2}\sin(\theta + 100)\right] \tag{2.17a}$$

$$E_2 = \left[\frac{hw_1}{2}\cos(\theta + 260), \frac{hw_2}{2}\sin(\theta + 260)\right] \tag{2.17b}$$

where θ is the angle between the source and the head, $hw_1 = 19$cm and $hw_2 = 21$cm correspond to the length of the elliptical head minor and major axes, respectively. We will consider here full bandwidth signals, i.e. the ITD estimation is not band-limited below 700Hz. When the ear is in the shadow region of the head the distance from the source s to the ears E_1, E_2 is given by,

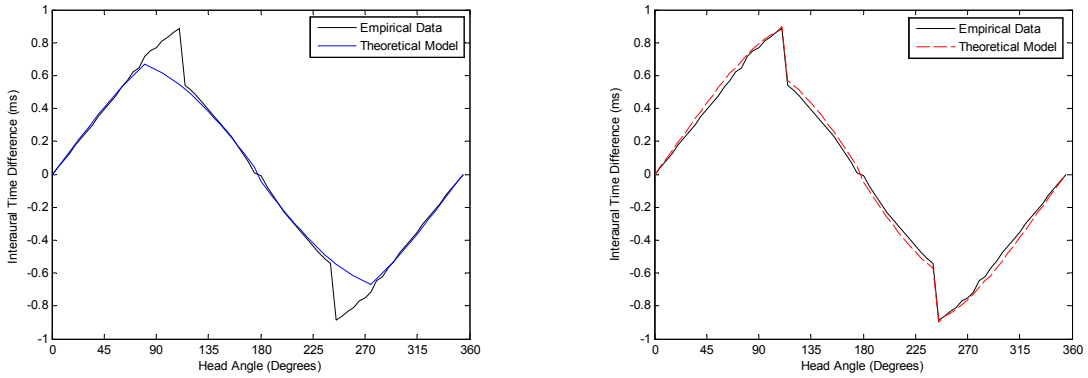
$$d_{sE_1} = d + \min(d_{E_1T_1}, \rho d_{E_1T_2}) \tag{2.18a}$$

$$d_{sE_2} = d + \min(\rho d_{E_1T_1}, d_{E_1T_2}) \tag{2.18b}$$

where d is the distance from the source to the non-shadow part of the head, and each distance $d_{E_aT_b}$, represents a diffraction path as shown in the geometry of Figure 2.33. The variable ρ is



Figure 2.34: *Studio setup for binaural recordings, consisting of a Genelec 1029a, Neumann KU100 head and rotation table.*



(a) Circular Head Model

(b) Elliptical Head model

Figure 2.35: *Comparison of circular and elliptical (with path weighting) mapping functions to empirically derived data.*

included here as an extra propagation factor in the diffraction path due to pinnae shadowing effects. When the ear is in the non-shadow region of the head in relation to the source the values of d_{sE_1} and d_{sE_2} are taken to be,

$$d_{sE_1} = |s - E_1| \tag{2.19a}$$

$$d_{sE_2} = |s - E_2| \tag{2.19b}$$

The theoretical time delay is then determined as,

$$T = \frac{d_{sE_1} - d_{sE_2}}{c} \tag{2.20}$$

where $c = 343$m/s is the speed of sound.

In order to investigate the accuracy of this head model, empirical measurements were taken with a Neumann KU100 dummy head with a Genelec 1029A loudspeaker (at 1m from the head) in a controlled studio environment using the setup in Figure 2.34. The theoretical mapping function obtained is shown superimposed over the empirical measurements in Figure 2.35. It is also useful to examine how this model compares to that of a circular head model without compensating for the shadowing effect of the ears. This corresponds to the case where $hw_1 = hw_2$ and $\rho = 1$.

We can see from Figure 2.35(a) that the theoretical and empirical time delay estimates match well for angles between 280° to 80° and 115° to 245°. However, localization accuracy breaks down around the contralateral points between 80° to 115° and 245° to 280°. This is attributed to the shadowing effect of the dummy head outer ear with respect to the ear canal affecting the intensity levels at the microphones, thus highlighting the need for the weight ρ in Equation (2.18b). Because the shortest propagation path has a smaller intensity at the ear, the dominant peak in the cross-correlation function corresponds to the longer path to the ear. At approximately ±115° the ear canal is no longer shadowed and the time delay estimate drops significantly. The resultant theoretical mapping function obtained using the elliptical head model with weighting $\rho = 2$ is shown in Figure 2.35(b), overlaying the empirically derived mapping function. The elliptical head model can be seen to explain well the empirical data due to path difference length, particularly at the contralateral points through the use of the path weighting ρ.

2.4 An Improved Model for ITD Estimation in Reverberant Environments

Whilst the full bandwidth IACF gives a good indication of source position by calculating the ITD, its use in reverberant environments can lead to localization that differs from subjective results. In the context of the work presented in this thesis, it is important to have a robust localization mechanism with which to assess the performance of virtual auditory environment creation under real non-anechoic (domestic, concert hall etc) conditions. In this section, it is shown that it is possible to improve the accuracy of the IACF by pre-filtering the correlated audio prior to normalization. This is demonstrated by comparing computational localization with real subjective listening tests.

2.4.1 Generalized Cross-Correlation and the IACF

It is interesting to note that current trends in computational non-binaural sound source localization utilize large numbers of microphones for direction of arrival estimation, for example in [Omologo and Svaizer, 1994], [Omologo and Svaizer, 1993], [Silverman et al., 1997]. At the heart of these computations is the estimation of the relative time of arrival of the source wavefront at the microphones. Like binaural microphones, this is commonly achieved by cross-correlation of the received signals at the microphone array in order to determine their relative time delays. Several frequency domain weighting functions have been explored which aim to enhance this estimation technique [Knapp and Carter, 1976], [Stuller, 1987]. Such a weighting function, or 'processor', acts as a pre-filter to the cross-correlation so as to accentuate the signals at frequencies where the signal to noise ratio is highest, whilst suppressing the noise components. However, the use of such processors for estimation of Interaural Time Differences (ITDs) from binaural recordings has not been sufficiently explored. Such an exploration is not without a neurophysiological basis: It has been shown that animals (such as the barn owl) incorporate a cross-correlation like process of the signals at the ears and that neural activation due to ITDs is not as sensitive to noise as would be predicted by normal interaural-cross correlation [Rucci and Wray, 1999]. This implies that a weighting process is employed, analogous to the one that will be presented here.

We will begin by modeling the signals received at the ears due to some stimulus as,

$$\begin{aligned} x_1(t) &= s(t) + n_1(t) \\ x_2(t) &= \alpha\ s(t+T) + n_2(t) \end{aligned}$$

where $s(t)$ is the source signal, $n_1(t)$ and $n_2(t)$ are the independent noise components associated with $x_1(t)$ and $x_2(t)$ respectively and T is the relative time delay between the received signals. The value of α is used to represent signal attenuation. Prior to the normalization of (2.14) pre-filters may be applied in the frequency domain so as to enhance the observed peak at the true time delay. This defines the generalized cross correlation (GCC) approach [Knapp and Carter, 1976] where the GCC function is defined as,

$$R_{x_1x_2}(\tau) = \int_{-\infty}^{+\infty} \Phi(\omega)\ G_{x_1x_2}(\omega)\ e^{j\omega\tau} d\omega \tag{2.22}$$

where $G_{x_1x_2}(\omega)$ is the cross-power spectrum of signals $x_1(t)$ and $x_2(t)$ and $\Phi(\omega)$ is a weighting function. In the case of the white, independent noise components $n_1(t)$, $n_2(t)$ and multiple delays T_i with corresponding attenuation factors α_i, the GCC function can be interpreted as,

$$R_{x_1x_2}(\tau) = R_{ss}(\tau) * \sum_i \alpha_i \delta(\tau - T_i) \tag{2.23}$$

where $R_{ss}(\tau)$ is the autocorrelation of the source signal $s(t)$. From this, it can be seen that a sharp peak at the delays T_i is desirable so as to avoid spreading of the delta functions making

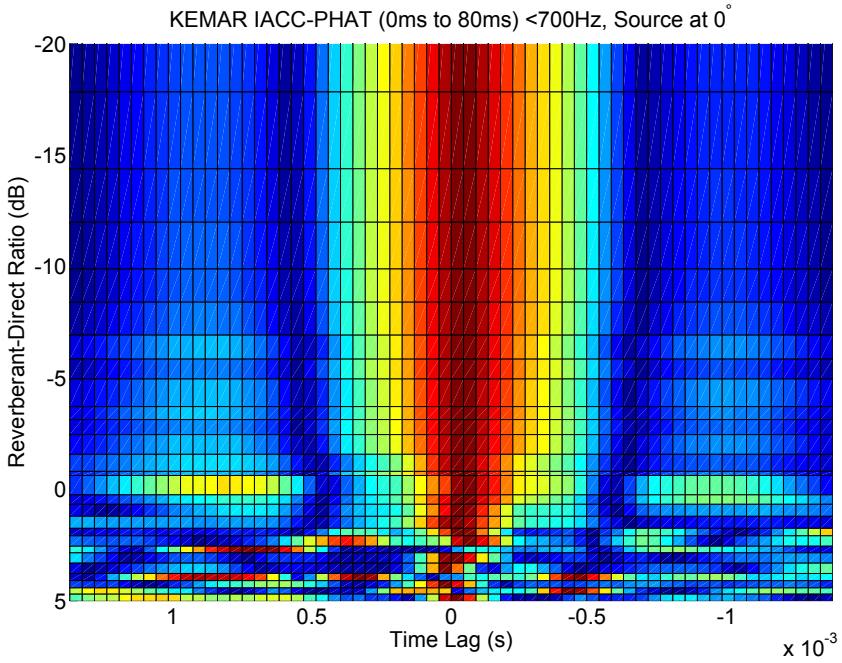


Figure 2.36: *IACC-PHAT (0 to 80ms) due to broadband source (< 700Hz) at 0° with increasing level of diffuse field reverberation. The diffuse field is delayed by 30ms in the simulation. The plot shows the IACC values with increasing level of reverberation and hence a reverberant-to-direct field ratio is presented. The equivalent IACF simulation is shown in Figure 2.30.*

---

individual peaks indistinguishable from one another. The use of a weighting function $\Phi(\omega)$ attempts to reduce this spreading. One such weighting is that of the Phase Transform (PHAT) given by,

$$\Phi(\omega) = \frac{1}{|G_{x_1x_2}(\omega)|} \tag{2.24}$$

This function essentially whitens the magnitude spectrum of the signals prior to cross-correlation. It has been shown to be the most suitable choice of processor for use in reverberant environments with microphone arrays [Gustafsson et al., 2003]. The PHAT weighting ensures that each frequency (bin) is given equal importance, since the signals are divided by their magnitude spectrum, resulting in constant energy over all frequencies. Thus, the correct direction of arrival is found based on the coherence between the de-emphasized signals. In the context of this work the non-weighted cross-correlation approach refers to the use of the IACF in ITD estimation, whose correlation coefficient is known as IACC, and the weighted cross-correlation approach refers to the use of the IACF with pre-filtering through the PHAT processor, whose correlation coefficient will be known as IACC-PHAT.

An example of the method applied to a binaural simulation of a broadband direct sound at 0° with increasing level of diffuse-field reverberation is shown in 2.36. In contrast to the equivalent simulation using the IACF, previously shown in Figure 2.30, we see that, even though the level of the reverberant field increases, it does not influence the peak value of the correlation to the

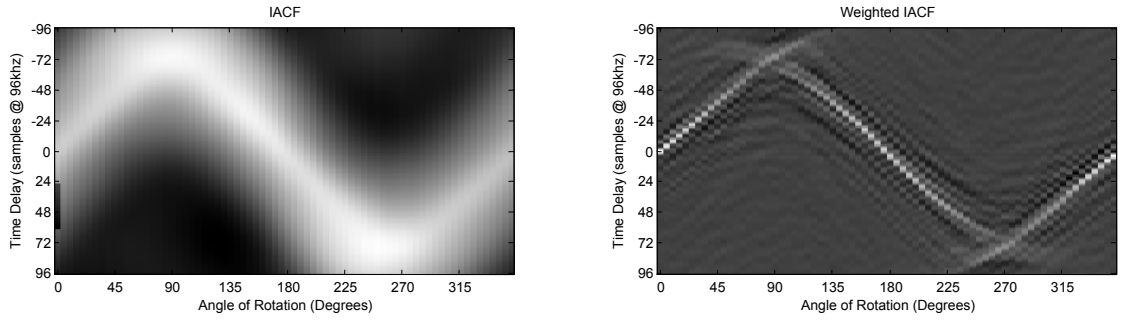


Figure 2.37: *Cross correlograms for simple cross-correlation and weighted cross-correlation using the Phase Transform.*

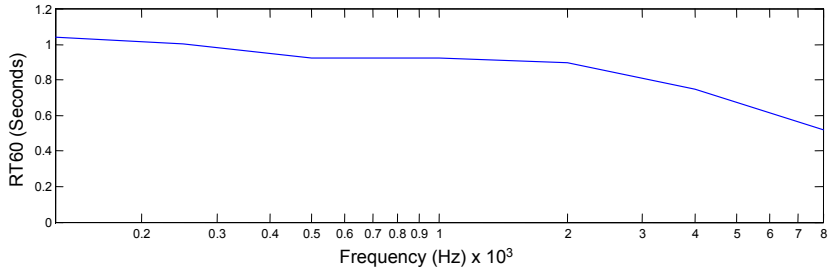


Figure 2.38: *Spatially averaged RT_{60} reverberation time for hall.*

same extent, due to the spectral de-emphasis.

2.4.2 Objective Evaluation of IACC-PHAT in a Reverberant Environment

Based on the empirical measurements taken in Section 2.3.5 the calculated time delay estimates to source angle mappings for both normalized cross correlation and weighted IACF can be found in Figure 2.37. These cross-correlograms show the magnitude of the cross correlation coefficient at the delays corresponding to each possible source angle. Gray-scale levels are used to visualize the magnitude of the coefficients increasing from black to white in the images. We can clearly see here how the use of the PHAT processor tightens the estimation of the delay as opposed to the smearing shown in the standard cross correlation. This smearing is due to the early reflections of the recording environment. Lookup tables were then created from the delays recorded at the maximum value of each correlation. Simple linear interpolation between the 5° increments was used in order to utilize the high resolution sample rates (96kHz) available in calculation of the ITD.

A second set of measurements was then taken in a small sized concert hall in Trinity College Dublin. This hall, termed *Test Environment 1*, is the main reverberant environment used throughout this thesis work. The acoustical parameters of this environment, corresponding to ISO-3382 standards [ISO, 2009b] are shown in Appendix A. The spatially averaged reverberation time of the hall, taken from the measurements shown in Appendix A is given in Figure 2.38.

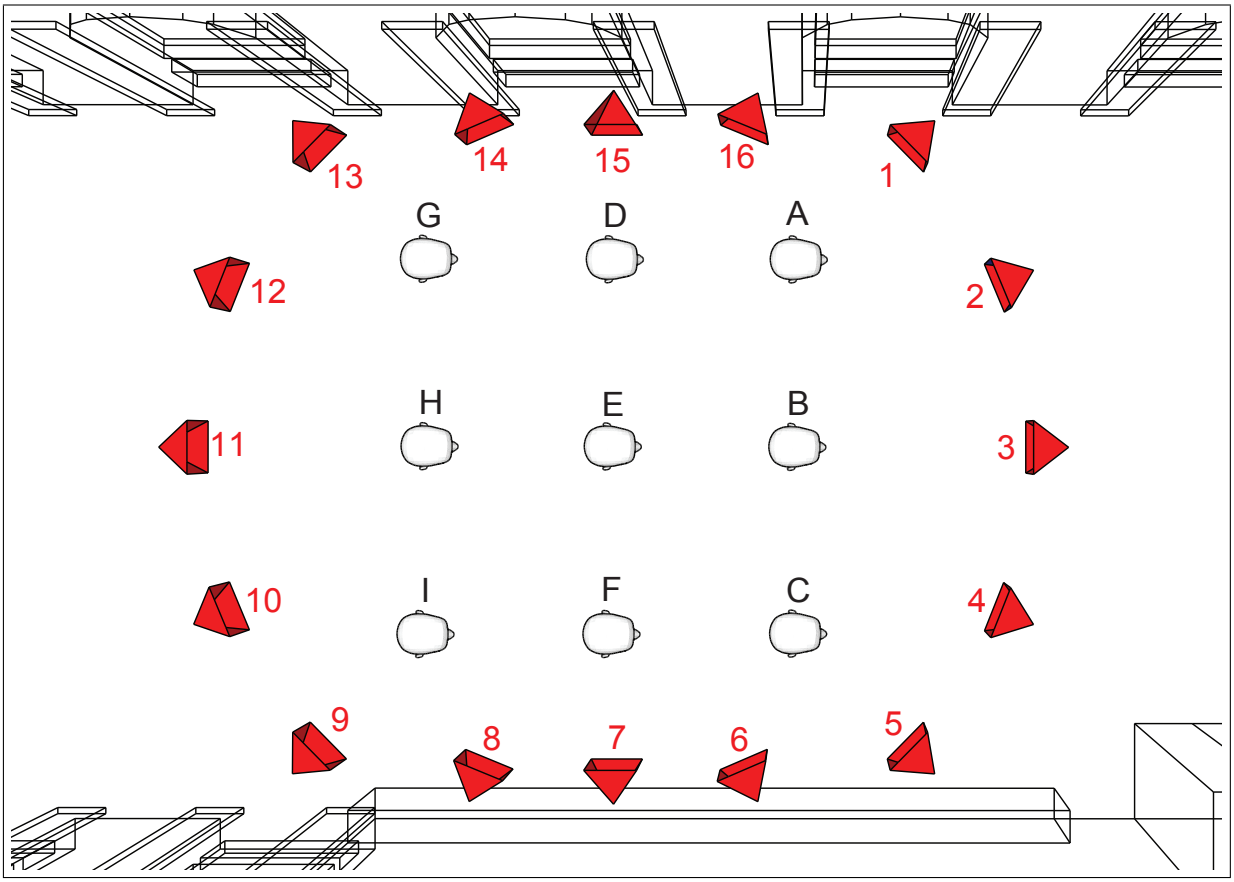
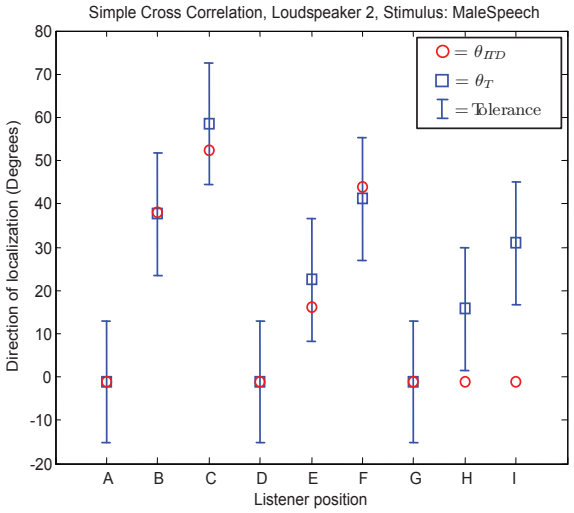
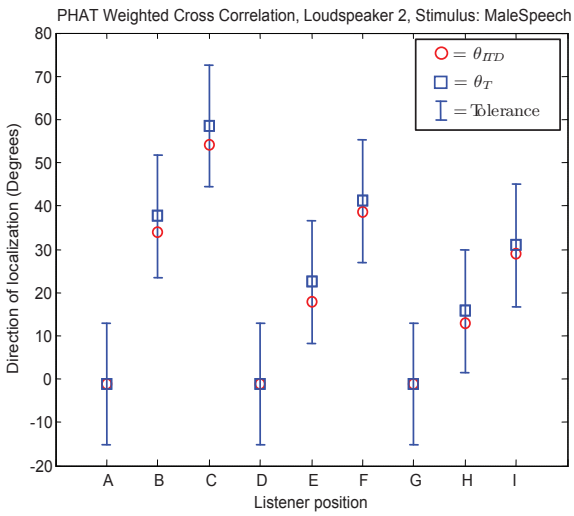


Figure 2.39: *Geometry of loudspeaker array and audience area for monophonic listening tests.*

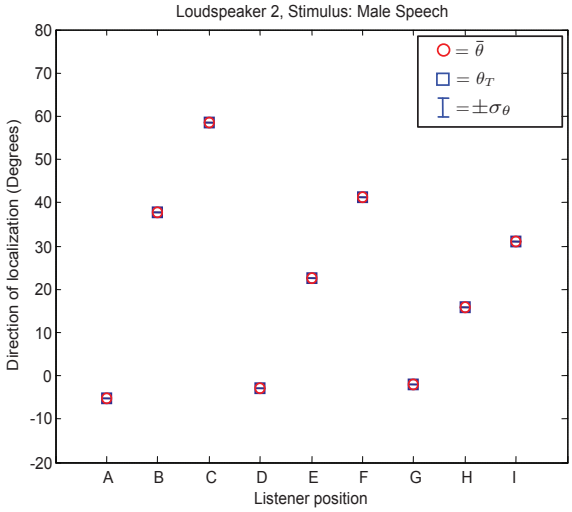
The test setup, shown in Figure 2.39 incorporates a 16 loudspeaker array of Genelec 1029A loudspeakers. Again, a Neumann KU-100 dummy head microphone was used, and a MOTU 896HD interface was employed to route the audio to a PC. Each loudspeaker was calibrated to 80dBA at 1m from the on-axis tweeter position and their axis lines were coincident with the centre listener position. The dummy head was placed at each listener position and was used to capture presentations of recorded male speech from each loudspeaker. Time delay estimates were made using IACC and IACC-PHAT from the resultant recordings. The ITDs were translated to source positions using the associated ITD lookup tables, with prior knowledge of the hemispherical position of the source, thereby avoiding front-back confusion. The objective results for front, front-lateral, rear and rear-lateral localization, corresponding to presentations from loudspeakers 2, 6, 10 and 14 respectively are shown in Figures 2.40 through 2.43 , as plots (a) and (b) in each case. It is important to note that in this set of plots the error bar $(\vdash\dashv)$ does not correspond to angular deviation $\pm\sigma_\theta$, but rather to the accepted tolerance of localization in the direction of a particular loudspeaker, before the source is localized at the another loudspeaker. This tolerance is set by the angles corresponding to the halfway points



(a) Computational localisation (No weighting)



(b) Computational localisation (PHAT weighting)

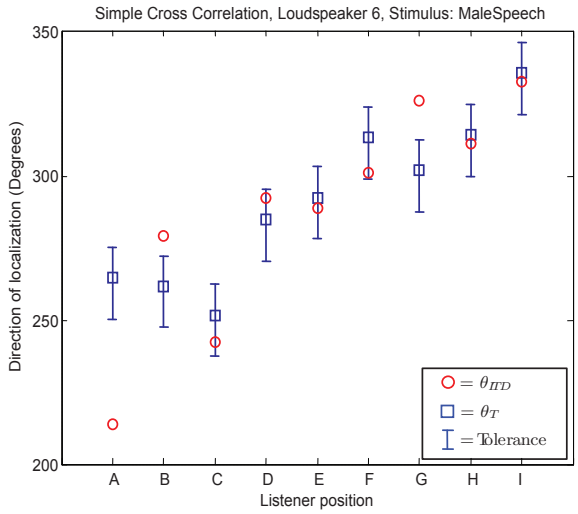


(c) Subjective localisation

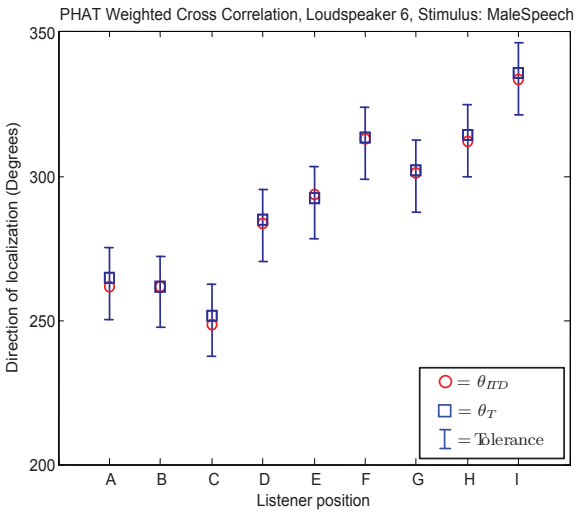
Figure 2.40: *Objective front localization using mapping function and ITD estimation using* **(a)** *IACF,* **(b)** *PHAT weighted IACF and* **(c)** *Subjective localization.*

between the loudspeakers on either side of the target location.

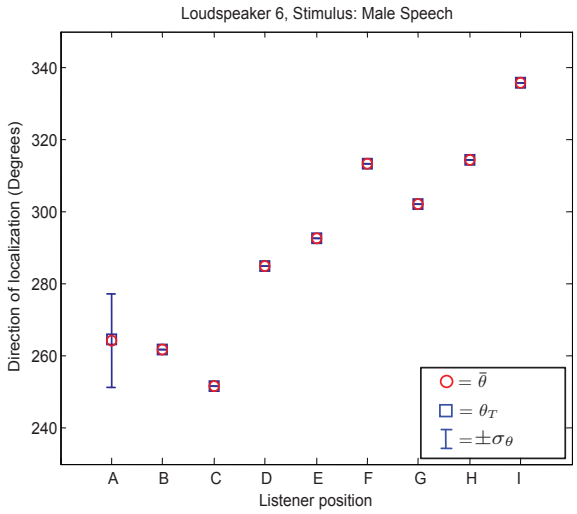
It can be seen that weighted cross correlation performs to a higher accuracy than standard IACF, in particular at seat positions where the reflection levels can be significantly higher than the direct sound. This can be seen for presentations from loudspeaker 2 (Figure 2.40) at seat positions H and I and loudspeaker 6 (Figure 2.41) at positions A and G. We also see that the PHAT significantly improves the localization accuracy at the contralateral points. This can be seen at loudspeaker 6 at listener positions A, B and C, and loudspeaker 14 (Figure 2.43) at listener positions H and I. We also note that all results using the PHAT method are within the



(a) Computational localisation (No weighting)



(b) Computational localisation (PHAT weighting)



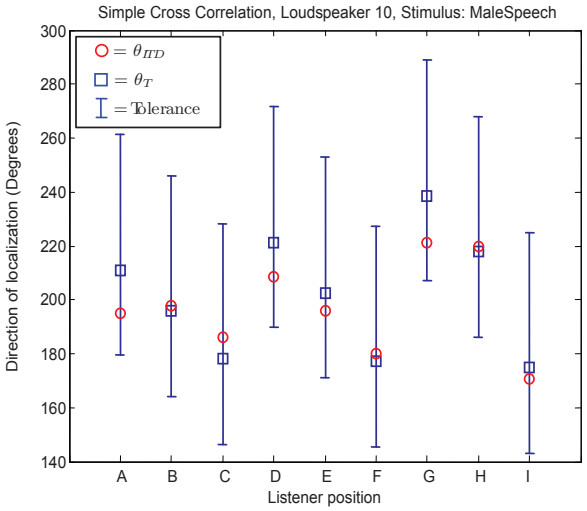
(c) Subjective localisation

Figure 2.41: *Objective front-lateral localization using mapping function and ITD estimation using* **(a)** *IACF,* **(b)** *PHAT weighted IACF and* **(c)** *subjective localization.*

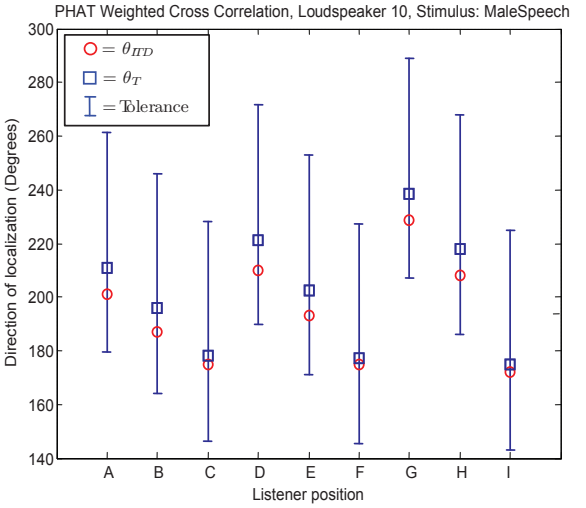
tolerance limits for loudspeaker localization in the array.

2.4.3 Localization Experiment I: Evaluation of IACC-PHAT Through Perceived Localisation in a Reverberant Environment

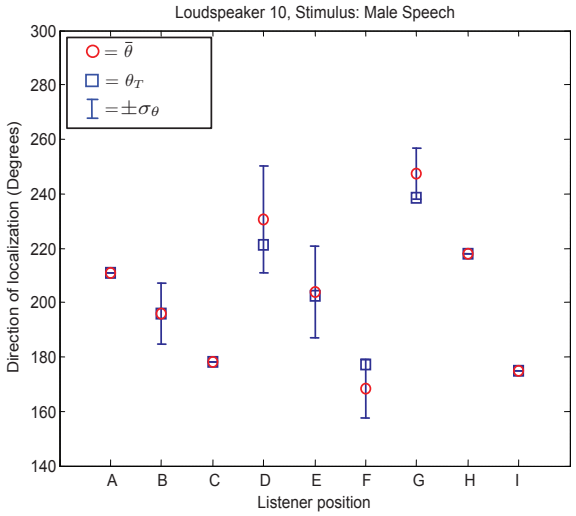
The objective results obtained with the theoretical model are also compared to the results of real audience listening tests. For this, an audience of nine test subjects were presented with the same male-speech samples as in the binaural measurements, from pseudo-random (pre-determined) positions located about the loudspeaker array. Subjects were asked to mark the loudspeaker



(a) Computational localisation (No weighting)



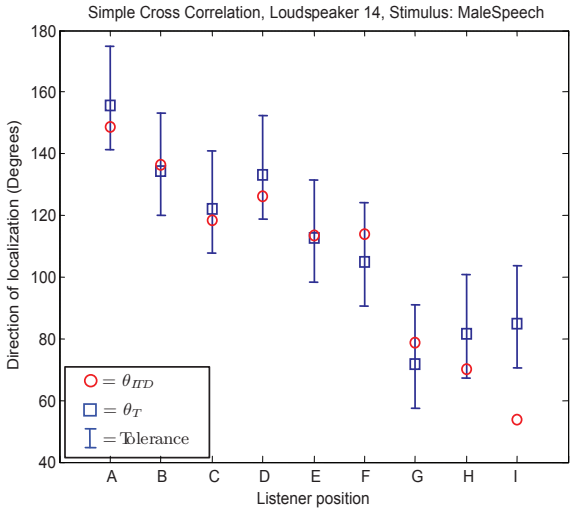
(b) Computational localisation (PHAT weighting)



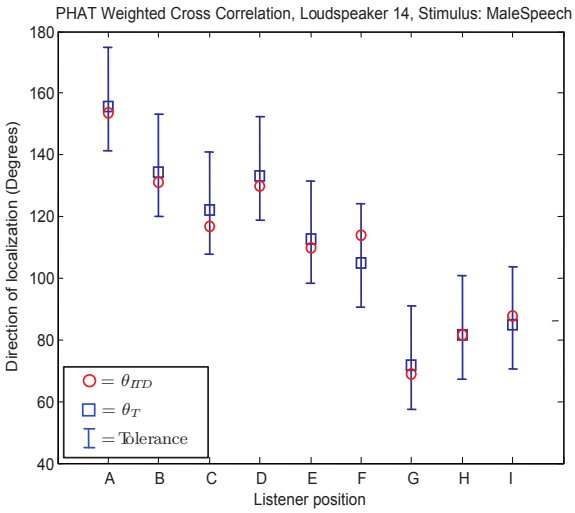
(c) Subjective localisation

Figure 2.42: *Objective rear localization using mapping function and ITD estimation using* **(a)** *IACF,* **(b)** *PHAT weighted IACF and* **(c)** *subjective localization.*

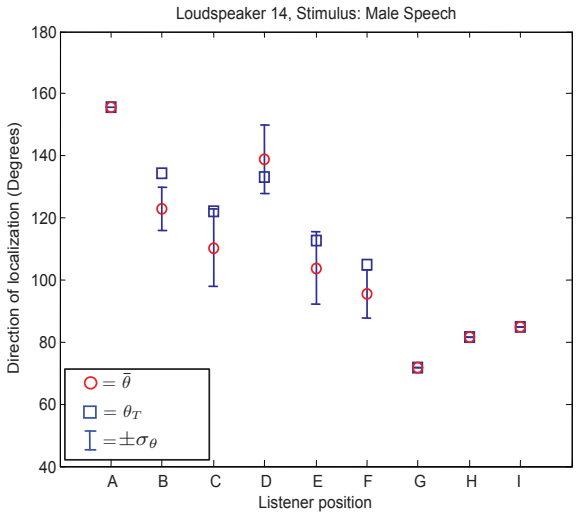
they felt the sound originated from on test sheets printed with a plan view of the test setup. This test was designed to assess large errors in subjective localization versus IACC. The randomized presentation method was used to negate any order effects during the tests. The same source locations used for the binaural measurements were used for the subjective testing. Each sample was presented twice, followed by a short interval before the next presentation and the listeners were asked to keep their heads in the forward direction. In total, data was collected from three groups of nine test subjects, and each subject repeated the test in each source position, yielding 243 votes per test round.



(a) Computational localisation (No weighting)



(b) Computational localisation (PHAT weighting)



(c) Subjective localisation

Figure 2.43: *Objective rear-lateral localization using mapping function and ITD estimation using* **(a)** *IACF,* **(b)** *PHAT weighted IACF and* **(c)** *subjective localization.*

Each of the listeners' answers were weighted, depending on the confidence level of the listener with their choice, with weightings of $1/i$, where i is the number (or range) of speakers that a listener felt the sound originated from. For example, if a listener felt that the source originated from both loudspeaker 2 and 3, then their vote is 0.5 for each, i.e. $h(\theta_2) = 0.5$ and $h(\theta_3) = 0.5$. From this, the histogram $\{h(\theta_n)\}_{n\in[1:16]}$ collecting all the listeners' answers is computed for each seat. The angular mean $\bar{\theta}$ and the unbiased standard deviation σ_θ at each listener position are computed as:

$$\bar{\theta} = \frac{\sum_{n=1}^{16} h(\theta_n)\ \theta_n}{\sum_{n=1}^{16} h(\theta_i)} \tag{2.25}$$

$$\sigma_\theta = \sqrt{\frac{\sum_{n=1}^{16} h(\theta_n)(\theta_n - \bar{\theta})^2}{(\sum_{i=1}^{16} h(\theta_n)) - 1}} \tag{2.26}$$

The results of these tests can be found as part (*c*) in Figures 2.40 through 2.43 for sources at speakers 2, 6, 10 and 14. The plots show the mean $\bar{\theta}$ (circle), deviation σ_θ (whiskers) and presented localization angle, or ground truth (square) from the perspective of each listener position. It is important to note that the subjects were tested using a forced speaker identification method which explains the high degree of correlation between the mean results and presented angle, and in general, a higher accuracy over the objective results. The deviation varies considerably for different listening and source positions, which is unsurprising considering the non-ideal listening conditions. Nonetheless, we can see the negative effect of lateral reflections on localization accuracy, since the test room contained a number of hard flat surfaces, in particular the two side walls, which generates significant reflections.

Overall, the results show that monophonic sources can be reasonably well localized by a distributed audience under reverberant conditions. Moreover, they show that the weighted cross-correlation was found to match the mean values in the subjective results to a higher degree of accuracy than the unweighted estimates. These results validate the use of PHAT weighting for ITD-based localization in reverberant rooms and its subsequent adoption in this thesis work.

2.5 Conclusions

In this chapter we have outlined the principal factors of localization and perception of sound in both anechoic and reflective environments. It was shown how localization of source stimuli can be achieved through interaural time and level difference as well as monaural spectral cues. Localization in a reflective environment revealed how initial reflection amplitudes and arrival times must be within the boundaries of the 'precedence effect' if correct localization is to occur. Distance perception was found to be related to near field ILD effects as well as the direct to reverberant ratio and arrival time and level of the reflections. The use of the interaural cross correlation function on the early part of binaural impulse responses was found to be a good measure of apparent source width. However, the use of the IACC as an estimator of ITD in reverberant environments was shown to be prone to localization errors. A theoretical mapping function was derived based on the elliptical shape of the head and the weighting of reflection paths was shown to improve the mapping of empirically measured broadband data over a simple head model. The shadowing effect of the ear on the ear canal was found to play a significant role in this mapping. Finally, it was shown that the PHAT processor improves the estimation of the interaural delay and the results complement the subjective listening tests.

3

Spatial Audio Systems for VAE presentations

3.1 Introduction

Having studied the psychoacoustical properties of sound propagation in rooms, we now investigate superimposed soundfields created by spatial audio reproduction systems. The purpose of this investigation is to find an appropriate solution to meet the demands of both the individual listener as well as a distributed audience in terms of localization accuracy and spatial impression for audio reproduction in VAEs. This analysis will span both this and the following Chapter. In this Chapter, we will review the state of the art in surround sound spatialization systems, such that we can identify potential effective and practical solutions for spatial audio reproduction in a VAE. First, we must clearly define the physical reproduction mechanisms for spatial audio, i.e. the transducer configuration. We will term these 'reproduction classes' which can be broadly divided in to headphone and loudspeaker listening. Coupled to the reproduction classes are 'spatialization techniques', used to form the transducer driving signals. Both reproduction classes and spatialization methods will be assessed according to their suitability for VAE creation, dictated by:

1. The psychoacoustical capabilities of the reproduction format
2. Localization capability over a listening area
3. The practicalities of system implementation.

The primary objective of this theoretical, analytical and practical assessment is to arrive at a spatialization format which gives plausible spatial cues over both loudspeaker and headphone

listening, thus catering for the demands of the individual listener and the distributed audience, as well as being a practicable storage and reproduction format. We will begin this review, by first investigating the limitations of the spatial reproduction classes, after which the spatialization schemes will be assessed.

3.2 Spatial Reproduction Classes

Before we proceed, we must clearly make the distinction between spatial and non-spatial audio reproduction. Given that monaural (one-ear) listening inhibits the correct localization of real sources, it is no surprise that monophonic playback is insufficient for delivering to the ears the correct wavefronts required for spatial separation of virtual sources, thereby classifying it as a *non-spatial reproduction* class. *Spatial reproduction*, where the spatialization system attempts to deliver the correct cues for auditory localization at the ears, can only be achieved through *multichannel* (two or more audio channels) reproduction. Such systems can be categorized into two main classes as shown in Figure 3.1:

- *Binaural reproduction*: Here, the transfer functions which describe the interaction of a listeners head and pinnae on impinging wavefronts are incorporated into headphone reproduction signals. In this way the correct cues (ILD, ITD and spectral cues) for effective localization and externalization of sound sources can be achieved. One should note that 'Biphonic' reproduction, i.e. the reproduction of recorded material over headphones that does not contain HRTF information, is not equivalent to binaural reproduction and the resultant sound image is localized entirely 'in-head'.

- *Multi-loudspeaker reproduction*: Here virtual sources are formed by introducing time or intensity biases (or complex frequency dependent combinations of both) to the loudspeaker signals. Spatial audio presentations over loudspeakers have developed from two channel-stereo for single listener reproduction to sound systems consisting of hundreds of loudspeakers, with the aim of wave field reproduction for distributed (large-area) audiences.

3.2.1 Binaural Reproduction

There are three main ways in which binaural audio can be generated:

1. By recording acoustic events using binaural microphones.

2. By convolving dry source audio with computed BRIRs.

3. By convolving dry source audio with measured BRIRs.

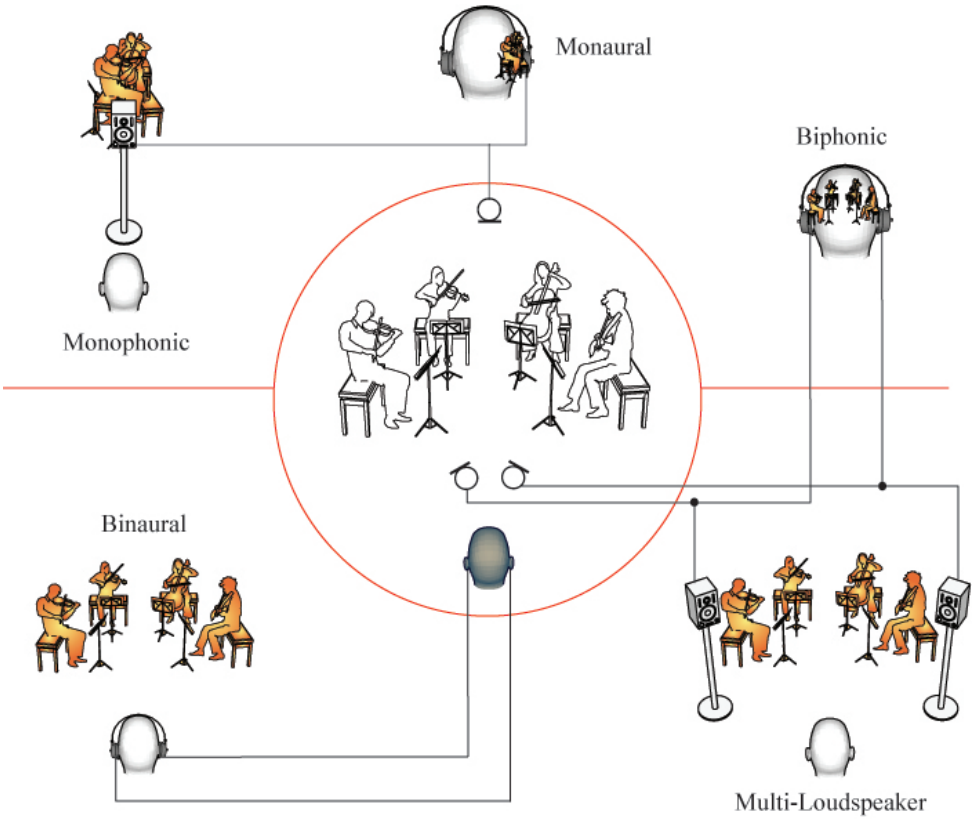


Figure 3.1: *Reproduction classes: The non-spatial reproduction class consists of monophonic loudspeaker reproduction, and monaural and biphonic headphone reproduction. Spatial reproduction consists of binaural headphone reproduction and multichannel loudspeaker arrays.*

In case 1, binaural microphones such as the Neumann KU100 or Schoeps KFM sphere, shown in Figure 3.2 are popular choices. A sensation of externalization (out of head localization) is usually achieved, but it is generally the case that recordings made with such binaural microphones do not translate to good localization on playback. This is because HRTFs change significantly from individual to individual, as we have seen in Section 2.2.4, and so the free-field transfer functions which characterize the interaction of sound with such binaural microphones are unique. To combat this, binaural recordings can also be made using in-ear microphones with individual listeners, but this is not a practical solution. It is also important to note that if the recording is taken in the ear canal, then a resonance at around 3kHz due to the size of the cavity will also be recorded. On playback via headphones, the ear canal resonance will effectively be doubled, unless the signals are equalized. One method to avoid this situation is to block the ear canal (known as 'blocked meatus') during the measurement/recording process [Kleiner, 1978].

For computational binaural auralization, good results can be achieved through simple methods which incorporate anechoically measured HRTFs and acoustic modeling techniques. For example, monaural room impulses responses can be computationally derived using the mirror-



(a) (b) (c)

Figure 3.2: *Binaural recording methods: (a) Neumann KU100 (b) Schoeps KFM sphere (c) In-ear capsule recording.*

image method [Allen and Berkley, 1979] alongside direction of arrival information and convolved with the free-field transfer functions of measured HRIRs [Ahnert, 1992]. Dry source audio can then be convolved with these responses to spatially place the source for the listener. The plausibility of reproduction i.e. the degree of convincing spatialization, is of course directly related to the computation of the RIRs. However, it is widely accepted that more convincing results are obtained from true BRIR measurements. Here, a known stimulus, such as maximum length sequence noise or exponential sine-tone sweeps is played from a (typically omnidirectional) loudspeaker in a reverberant space and then recorded back with a binaural microphone or with human subjects with microphones placed at their ear canals [Farina, 2000]. Deconvolution of the recorded signals results in the measured BRIRs.

In utilizing the aforementioned methods for auralization there are several important considerations. Firstly, it has been shown that for effective externalization and localization to occur, head-tracking should be employed to control the spatialization process [Begault et al., 2001]. However, the switching of the directionally dependent HRIRs can lead to auditory artifacts caused by wave discontinuity in the convolved binaural signals [Otani and Hirahara, 2008]. Furthermore, high latencies ($>$ 50ms) should be minimized between head motion and HRIR switching. One method to reduce this is to shorten the finite impulse response (FIR) filter length, but as has been shown in [Sandvad and Hammershøi, 1994], lower order filtering can also lead to detectable artifacts. Furthermore, BRIRs taken in a particular reverberant environment can only ever be used to simulate sources in that environment. It is therefore clear that to have effective binaural listening experiences *for all listeners* we must consider a two-stage process, whereby directional sound fields are imposed onto measured *individual* anechoic HRIRs. This provides the flexibility for a personalized set of HRIRs to be used with other auditory scenes.

However, synthesizing a set of BRIRs from real room measurements is non-trivial. There are two ways in which this can be accomplished. The first is to decompose the RIR into individual reflections, where each reflection is combined with its corresponding HRTF, and the result

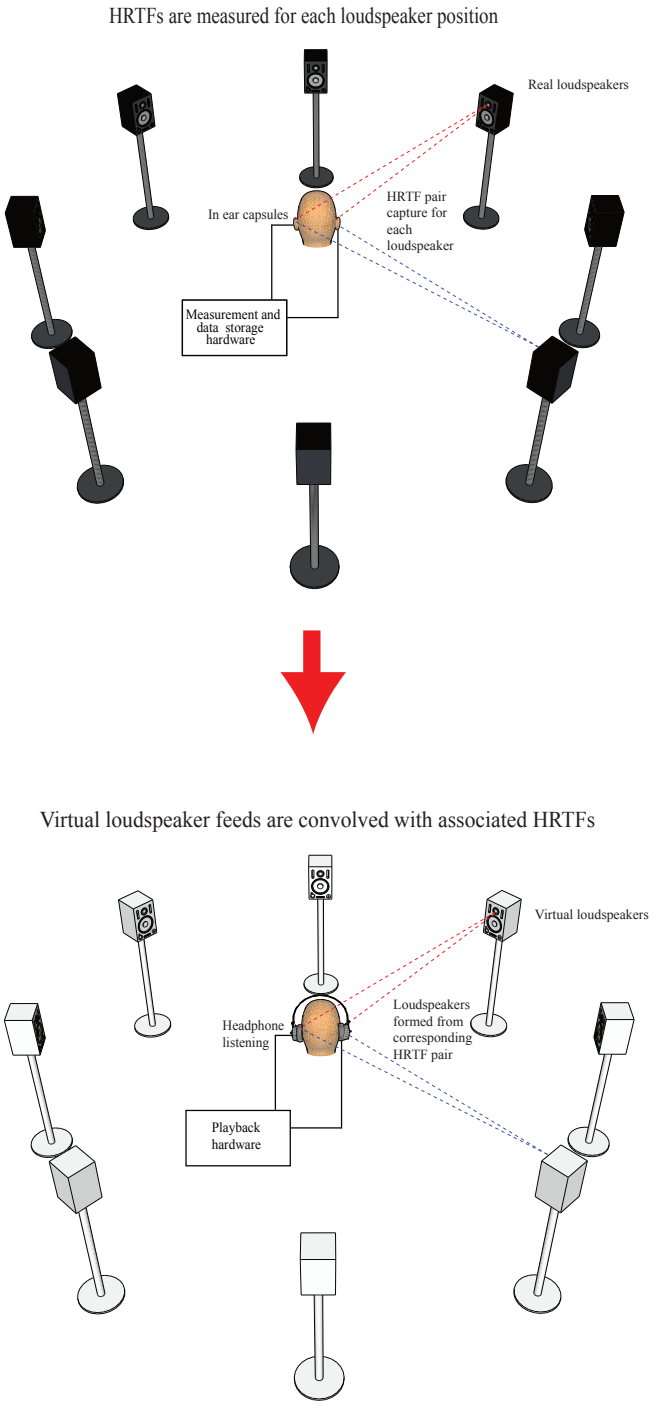


Figure 3.3: *The virtual loudspeaker reproduction concept.*

convolved with the dry source audio. This leads to a complex and inefficient filter kernel, in particular if head-movement is to be supported. An effective solution to this problem is to employ the 'virtual loudspeaker' approach, whereby HRTFs are measured at the 'sweet-spot' (the limited region in the centre of a reproduction array where an adequate spatial impression is generally

guaranteed) in a multi-loudspeaker reproduction setup and the resultant binaural playback is formed from the loudspeaker feeds to the virtual loudspeakers. This concept is illustrated in Figure 3.3. This method was first introduced by McKeag and McGrath [McKeag and McGrath, 1996], [McKeag and McGrath, 1997] and examples of its adoption can be found in [Noisternig et al., 2002], [Leitner et al., 2000] and [Dalenbäck and Strömberg, 2006]. This approach has major computational advantages, since a complex filter kernel is not required and head rotation can be simulated by changing the loudspeaker feeds as opposed to the HRTFs. This approach caters well for the needs of the individual listener and will be adopted in this thesis. Whilst the HRTFs in this case play an important role in the spatialization, ultimately it is the soundfield creation over the virtual loudspeakers which gives the overall spatial impression. Furthermore, the spatialization method adopted for loudspeaker reproduction can also be applied. This process is explored further in Chapter 7.

In general, binaural rendering is aimed at the individual listener through headphone reproduction, although there are spatialization methods which attempt to reproduce the binaural effect over loudspeakers. This is known as 'transaural' synthesis, and was first proposed by Atal et al. in the 1960s [Atal, 1966]. The concept is to remove the crosstalk created at the ears during loudspeaker reproduction, so as to emulate headphone listening. It will be shown in Section 3.3 that stereophony relies on the lateral separation of the ears, resulting in multiple delayed versions of the signals from the loudspeakers occurring at each ear. This does not happen in binaural listening, since the audio signals are applied to each ear separately. Thus, if binaurally recorded audio reproduced over loudspeakers is to have the same effect as headphone listening, this crosstalk must be eliminated. The solution is to use cross-talk cancellation filters. Consider, for example, the case where we have a standard stereophonic setup with the loudspeakers spaced at $\pm 30^{\circ}$. If we emit a pulse from the left loudspeaker, then with respect to the left ear a delayed version of this signal will occur at the right ear. This is the crosstalk component. In order to eliminate this crosstalk, we need to introduce another signal at the right ear, from the right loudspeaker, with the opposite polarity and equivalent magnitude to the crosstalk signal. This will eliminate the crosstalk signal at the right ear, but will create another crosstalk signal of lesser magnitude at the left ear. Thus, the process must be repeated for the left ear. The result is a set of filters which attempt to eliminate the crosstalk in this recursive manner as shown in Figure 3.4. The filters shown here are taken from a simple model of the head consisting of only two microphones, but they do illustrate the frequency response problems that occur due to the multiple delayed versions of the source signal at the ears. The comb-filtering exhibited can be pushed higher in the spectrum, if the loudspeakers are moved closer together. Figure 3.5 shows the crosstalk cancellation filters for loudspeakers at $\pm 5^{\circ}$ relative to the median plane. We see that the excessive comb-filtering is now further up the spectrum and the low frequency boost is exhibited due to phase coherent signals from the loudspeakers. Furthermore, moving the loudspeakers together also enlarges the effective area of crosstalk cancellation, increasing

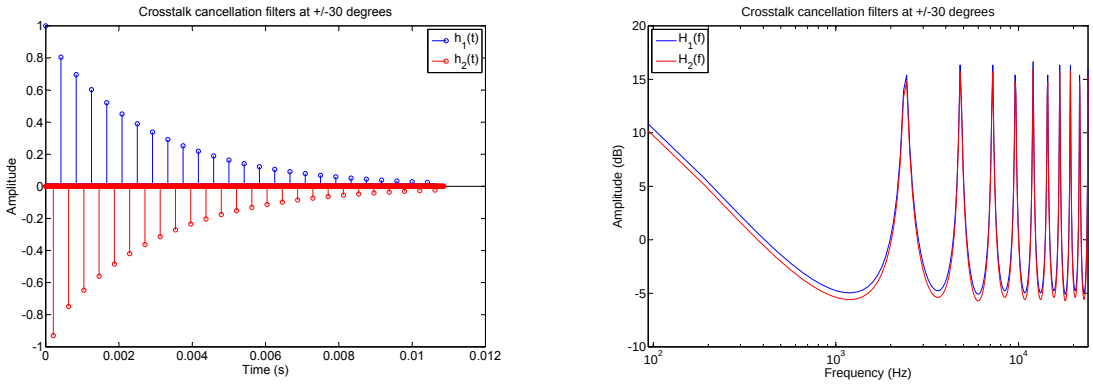


Figure 3.4: *Simple crosstalk cancellation filtering for* 30° *loudspeaker separation.*

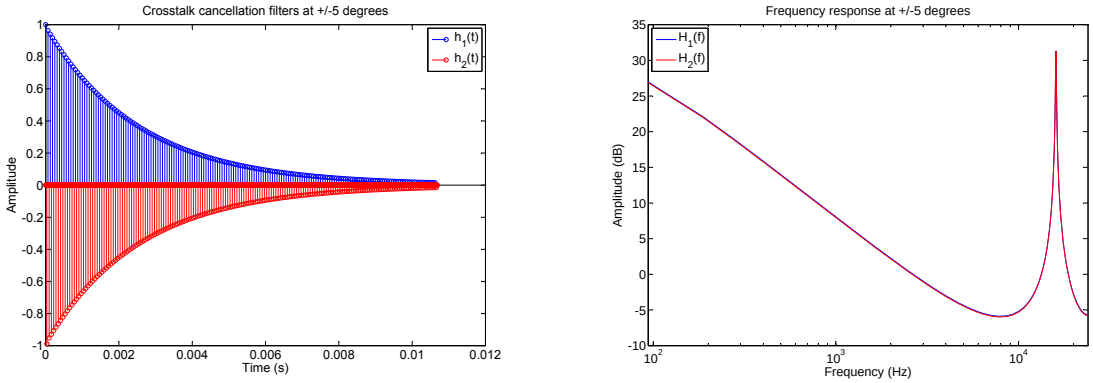


Figure 3.5: *Simple crosstalk cancellation filtering for* 5° *loudspeaker separation.*

the 'sweet spot'. Although the examples shown here are trivial they demonstrate effectively the main concepts of transaural reproduction. In reality, inverse filtering based on HRTF data is required.

One spatialization system which uses the transaural approach is Ambiophonics, developed by Glasgal [Glasgal, 2001]. The system aims at reproducing an enhanced frontal image using transaural techniques, whilst using dedicated loudspeakers to create lateral reflections. It is capable of extending the width of the stereophonic image up to 140° as opposed to 60° in standard two-channel stereo. Recordings made for Ambiophonics can be achieved using two-channel methods but the microphones are baffled such that lateral reflections are not recorded. The recorded signals are then convolved using impulses from the Ambiophonic Convolver (Ambivolver). Ambiophonics is limited however, since it always assumes that the stage is in the front of the listener. Whilst this is acceptable for reproduction of concert acoustics, this may not be case for all auralization scenarios, and in this regard Ambiophonics does not offer the potential provided by other spatialization systems such as Ambisonics.

We therefore conclude that reproduction of binaural signals over loudspeakers is not a practical solution to large audience presentations in a VAE. Conversely, virtual loudspeaker repro-

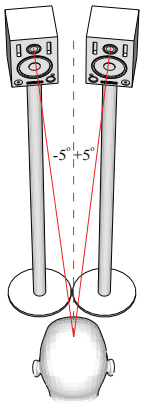


Figure 3.6: *Loudspeaker configuration for transaural reproduction.*

duction via binaural synthesis over headphones allows for flexible, efficient and effective *personal reproduction*. Therefore, in this thesis, we classify binaural synthesis as a reproduction method, and the spatialization scheme adopted for large audience listening will also be adopted for virtual loudspeaker reproduction over headphones. Thus, the major advantages of binaural synthesis as a reproduction class are:

- Fully immersive 3-D auralization is possible
- Any spatialization method is possible
- Any loudspeaker reproduction format is possible, and therefore content reproduction is not limited.

The challenges associated with this reproduction class are:

- Individualized HRTFs are desirable
- Head-tracking is required

3.2.2 Multi-Loudspeaker Reproduction

Auralization over loudspeakers offers the potential for VAEs to be created for multiple listeners. In this respect, there is a requirement that the spatialization method and loudspeaker layout should satisfy the localization requirements of a distributed audience, as opposed to a single listener sitting in an acoustic sweet spot. It follows, that any solution which caters for large audiences in this regard, will also work for single listeners. Thus, an appropriate spatialization method that satisfies both on and off-centre localization should also be applicable to the binaural virtual loudspeaker approach. It is also important that the loudspeaker layout is viable, and can be achieved in both domestic and non-domestic environments. This has always been a major factor in spatial audio system design and has led to significantly varied loudspeaker layouts over

the last century. It is fitting then, that in discussing potential loudspeaker reproduction formats for VAEs, we should approach this topic from a historical perspective.

In the early 1930's, significant developments in stereophonic capture and reproduction were made at EMI studios in London, by a young engineer known as Alan Dower Blumlein. His patent, "Improvements in, and relating to, sound recording and sound reproduction systems", outlined the basic concepts of what would eventually be known as 'intensity stereo' [Blumlein, 1931]. Blumlein's patent described a system where the a stereophonic illusion could be created to place virtual sources anywhere between two loudspeakers, and outlines its application to music and cinematic reproduction. However, domestic introduction of two-channel stereo was delayed for more than 20 years, partly due to the war, but also due to the transition from 78-r/min shellac records to 33-r/min vinyl [Torick, 1998]. To this day, two-channel stereophonic loudspeaker reproduction remains the main format for audio presentations of music production. However, over the last 60 years, the entertainment industry has invested significant commercial interest into multichannel loudspeaker formats, ranging from domestic surround sound systems for music and home cinema presentations to large scale cinematic systems. Early attempts at bringing surround sound to the consumer resulted in Quadraphonic reproduction, which doubled the number of loudspeakers to four (as shown in Figure 3.7), but extended the principle of intensity stereo operation in an unjustifiable manner. The idea was that each pair of adjacent quadraphonic loudspeakers be regarded as a stereo pair, and thus a phantom image could be created between a pair by simply dividing the source signal between the speakers in the correct ratio. This is, however, invalid, since for stereo to work the listener must be facing the speaker pair, thus ruling out the formation of stable images at the sides. Furthermore, as has been shown by Theile [Theile and Plenge, 1977], pair-wise panning with speakers greater than 60° apart leads to a hole-in-the-middle effect, so even the frontal stage is compromised in Quadraphonics.

The introduction of the quad formats was principally driven by hardware manufacturers, such as JVC, who were trying to expand the market after it reached near saturation with stereo records and tape players. Unfortunately, the whole quad era was a commercial disaster, principally brought about by far too many incompatible storage media and a very confused public, and also because of the poor psychoacoustical foundations of quadraphonic formats. Furthermore, the design of highly popular two-track media such as vinyl records and FM broadcasting meant that Quadraphonic systems had to use matrix encoding (sum and difference techniques) to convert four channels to two for compliance with these consumer media, resulting in limitations in bandwidth, dynamic range and the introduction of crosstalk [Ratliff, 1974].

Whilst the Quadraphonic era was declining, Dolby laboratories focused on cinematic presentations, and developed Dolby Stereo, a system designed to matrix encode four channels of audio (left, centre, right and rear) onto the two channels of standard 35mm film. Sum and difference techniques are then used to decode the L, C, R and S channels at playback. Dolby

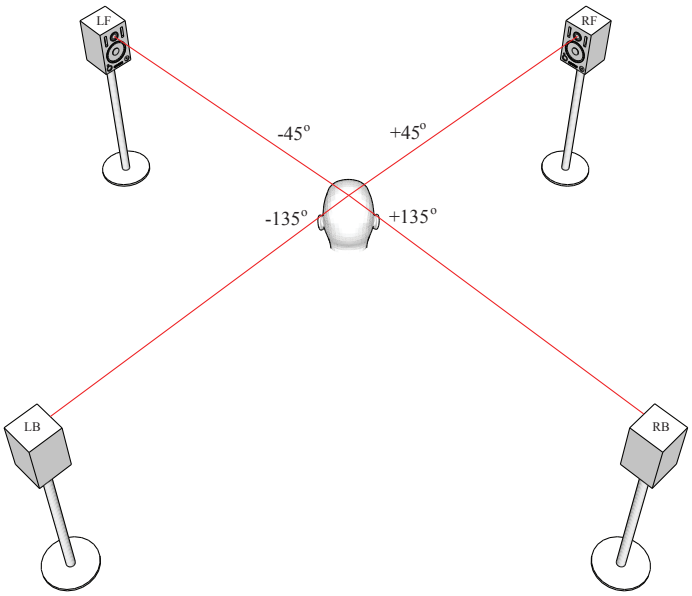


Figure 3.7: *Quadraphonic Layout.*

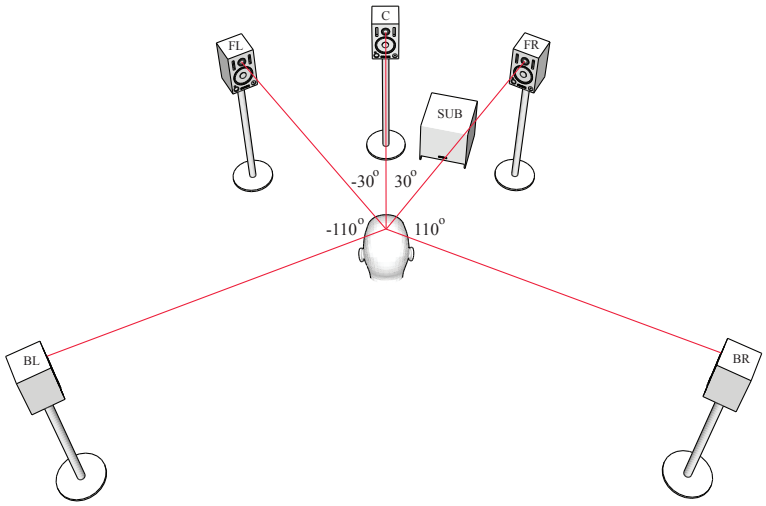


Figure 3.8: *ITU-R BS.775 Recommendation for 5.1 Surround [ITU, 2006].*

Stereo later developed into its discrete counterpart, Dolby Digital, which utilizes 6 discrete channels of audio, commonly referred to as 5.1 surround sound. The ITU recommendation for loudspeaker placement of 5.1 systems is shown in Figure 3.8 [ITU, 2006]. This consists of 3 full bandwidth front channels, 2 full bandwidth surround channels and a low frequency effects channel. Dolby Digital explicitly refers to the AC3 codec used to encode the audio onto a digital track at the side of a 35mm film and is also the mandatory codec on DVD movie discs [Dolby, 1999]. Like Dolby Surround, it is effective in the creation of cinematic type auditory effects for a distributed audience, but it cannot, nor did its designers ever aspire to, create virtual images

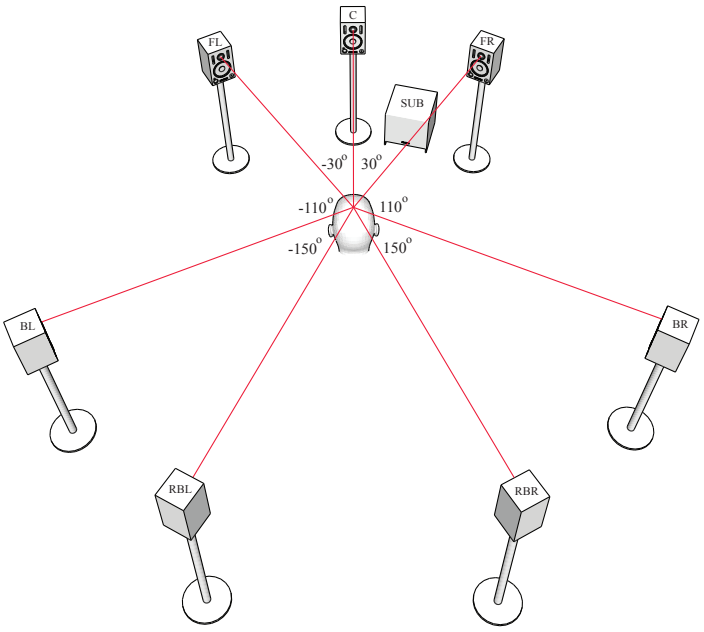


Figure 3.9: *7.1 'Standard' Surround Sound [ITU, 2006].*

all around a distributed audience. Further developments in cinematic soundtrack reproduction have led to surround codecs which support 8 channels of audio, enabling 7.1 reproduction. As a consequence, several 7.1 layouts have been recommended, the most popular of which is the 7.1 surround standard, shown in Figure 3.9 [ITU, 2006]. Layouts which include periphony (height) have also been suggested for domestic reproduction, the most notable being the 'Front-High' recommendation by DTS, which incorporates two loudspeakers above a standard 5.1 setup as shown in Figure 3.10 [DTS, 2007]. Further layouts have been suggested incorporating even higher speaker counts such as the NHK 22.2 system [Hamasaki et al., 2004], designed to accommodate ultra-high definition video.

At the same time as the development of the Dolby systems and cinematic surround, another survivor of the Quadraphonic era developed, called Ambisonics, a unified system for sound reproduction, recording and transmission. It survived the limits imposed by two channel systems since it does such an excellent job of recreating a full horizontal surround sound image over loudspeakers, whilst still retaining good stereo and mono compatibility. As a spatialization technique, it is a logical development of the work of Blumlein, and was developed by Gerzon, Barton and Fellgett [Gerzon, 1973]. As a unified system, it outlined different formats for how soundfield information is stored, processed and reproduced: The A-format is used for microphone pickup, the B-format for studio equipment and processing, the C-format for transmission, and the D-format for decoding, [Gerzon, 1975a], [Gerzon, 1977b], [Gerzon, 1977c], [Fellgett, 1975]. Ambisonics will be discussed in detail later in this chapter, but for the moment it is important to note that the Ambisonic theory originally outlined by Gerzon dictated regular (symmetrical)

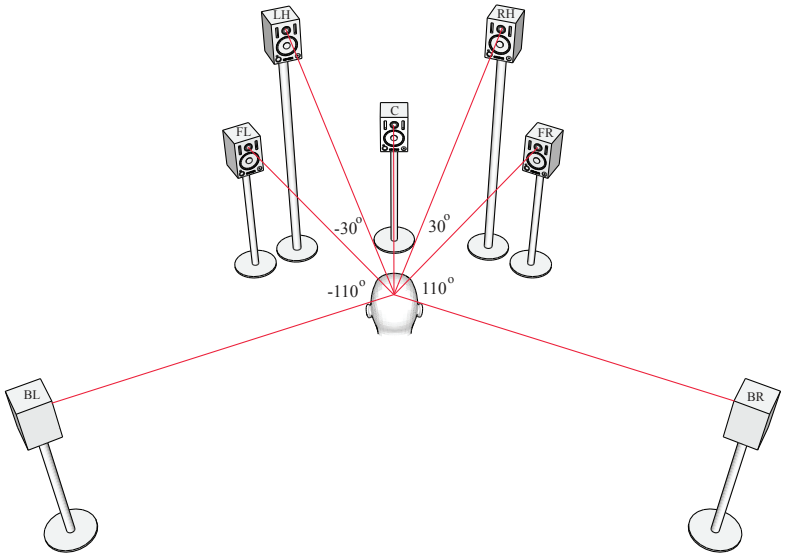


Figure 3.10: *DTS 'Front-High' Recommendation for 7.1 Surround [DTS, 2007].*

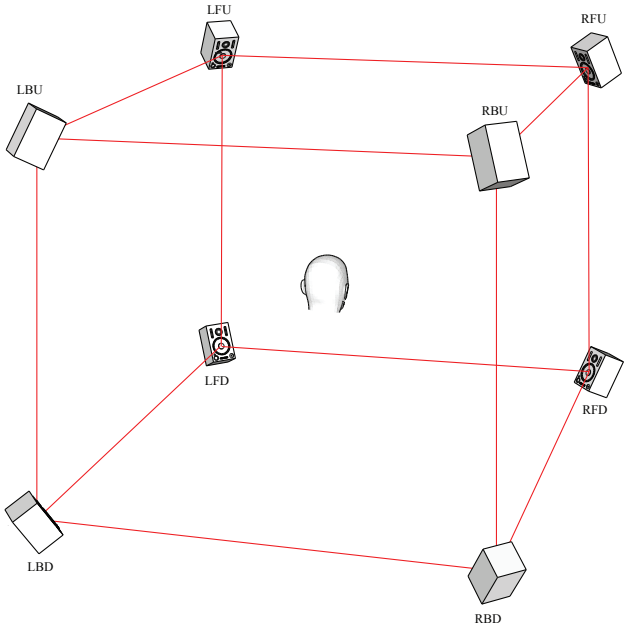


Figure 3.11: *Ambisonic cube for periphonic soundfield reproduction.*

loudspeaker setups, with all loudspeakers equidistant from the listener, such as the cube shown in Figure 3.11. This was necessary since the underlying psychoacoustical theory demanded diametrically opposed loudspeaker pairs [Gerzon, 1980]. However, since the early 1990s, further developments in Ambisonic decoders have allowed reproduction on irregular loudspeaker layouts such as ITU-R BS-775 [Gerzon and Barton, 1992].

We can therefore appreciate that there are a myriad of loudspeaker reproduction formats,

which are applicable for VAE presentations. Given that it is the purpose of spatialization techniques to accurately place the virtual (phantom) sound sources within the confines of the reproduction class, it is therefore important that we focus on loudspeaker layouts which give optimal performance for a chosen spatialization system. With this in mind, we will now outline the advantages and disadvantages of spatialization methods in the context of recording and reproduction for virtual auditory environments. In this chapter, we will focus on the spatialization techniques of VBAP, Ambisonics and Wavefield Synthesis. However, before these systems can be reviewed we must first outline the theory behind phantom image formation in conventional stereophonic recording and reproduction methods.

3.3 Stereophony

Early experiments into stereophonic reproduction were implemented by Harvey Fletcher and his team of researchers at Bell Labs in the 1930s [Lipshitz, 1986]. Fletcher visualized a curtain of microphones in a studio which were connected to an equal number of correspondingly placed loudspeakers in an auditorium. As Figure 3.12(a) shows, the concept was to pick up the sound wavefronts in the studio and recreate them using the loudspeakers in the auditorium, commencing where they 'left off' in the studio. The concept (which could be considered a forerunner to wavefield synthesis) required a large number of microphones and loudspeakers. However, due to obvious practical considerations, two and three speaker-microphone setups were found to be sufficient to give an adequate spatial impression in a limited region around the head of a listener centered on the reproduction axis, as show in Figure 3.12(b). This is known as the 'sweet spot'. Fletcher employed spaced microphone techniques, which meant that any off centre source wavefront would arrive at the microphones at different times. In this way, phantom images were created using inter-channel time differences. If the microphones contained directional characteristics, such as cardioids, then significant inter-channel level differences would also ensue.

As previously mentioned, another approach to phantom source localization was developed by Blumlein, who invoked psychoacoustic criteria in designing his system [Blumlein, 1931]. He realized that sound *intensity* differences at the loudspeakers would create both phase and level differences at the listener's ears since each ear hears *both* loudspeakers and that the ears are laterally separated. By suitably choosing these level differences it is possible to produce stable images between the loudspeakers. Figure 3.13 shows Blumlein's setup for 'intensity stereo' and his patent of 1931 documents how the signals from a coincident or closely spaced pair of microphones can be reproduced on two loudspeakers [Blumlein, 1931].

Based on the pioneering works of Fletcher and Blumlein, it can be concluded that the placement of stereophonic virtual sources is governed according to:

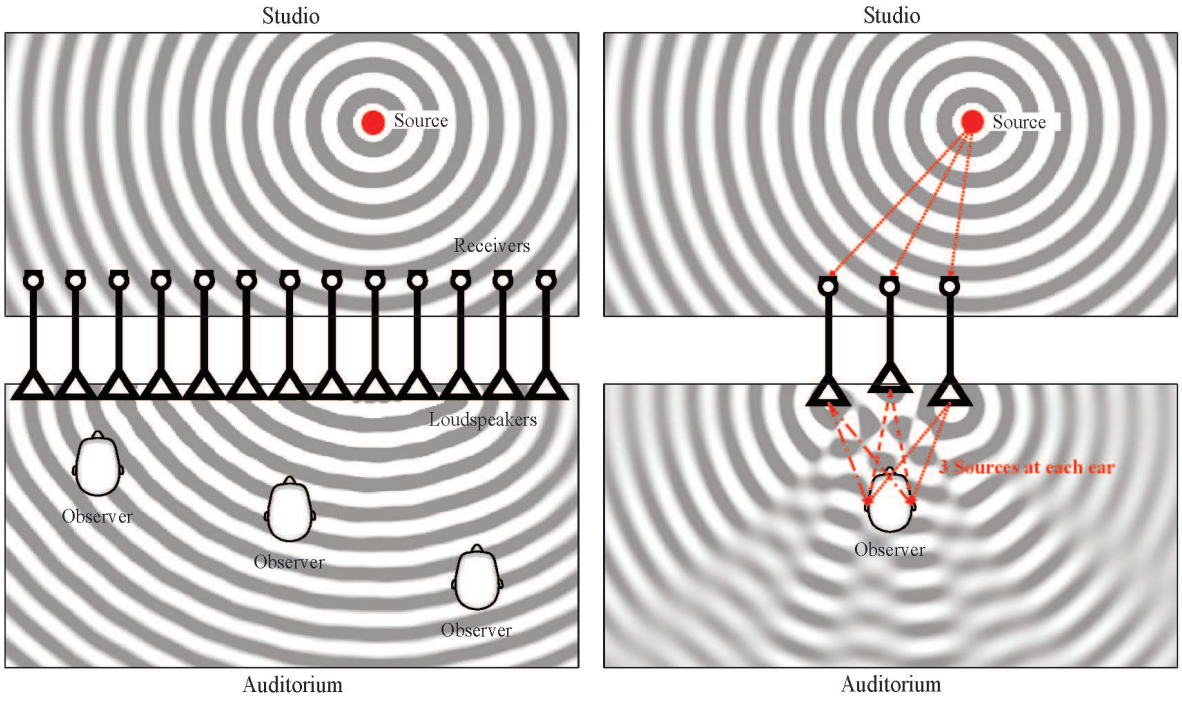


Figure 3.12: *Bell Labs Stereophonic Experiments (a) Curtain of microphones configuration (b) Three channel stereo configuration.*

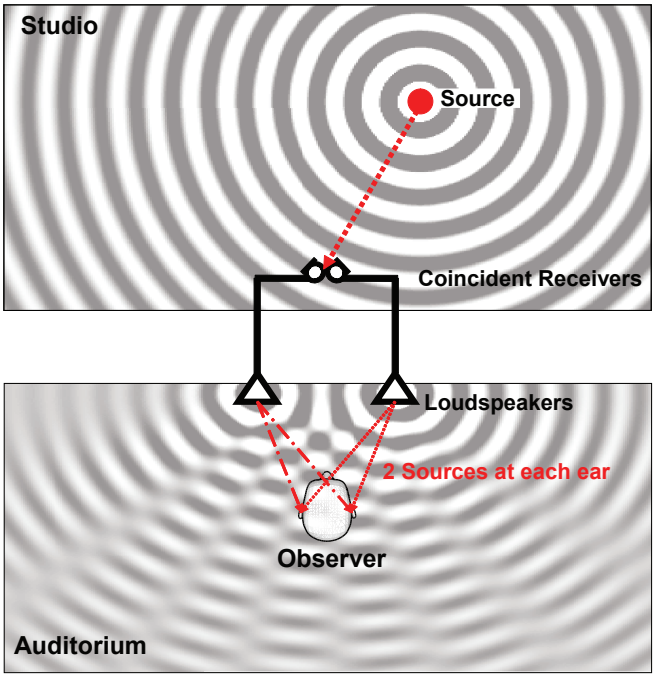
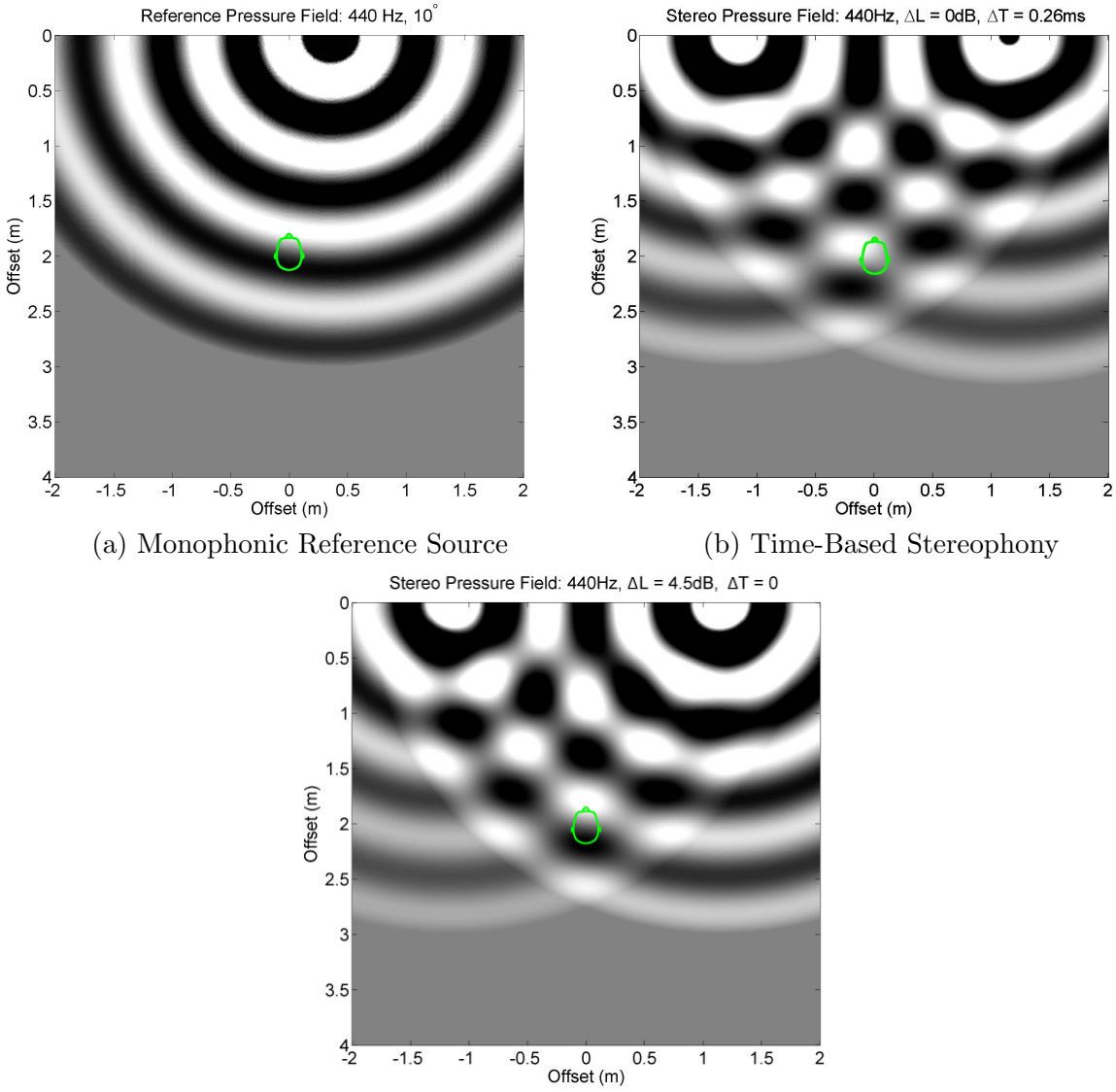


Figure 3.13: *Blumlein's configuration for 2-channel stereo.*



(a) Monophonic Reference Source

(b) Time-Based Stereophony

(c) Intensity-Based Stereophony

Figure 3.14: *Comparison of a reference pressure field created by monophonic 440Hz source at 10° to two channel stereophonic reproduction (Sweet spot is marked in green and pressure maxima in white).*

---

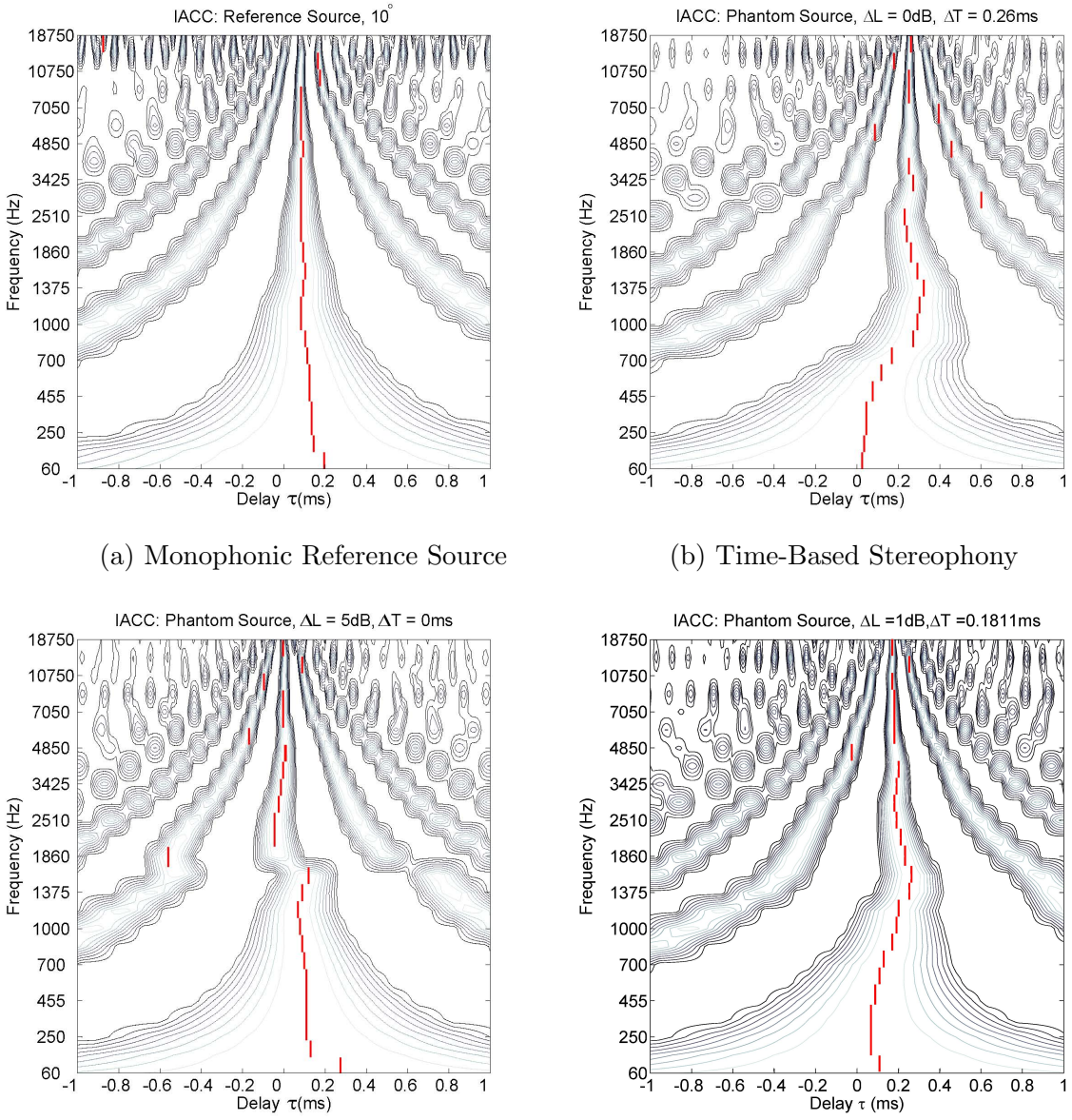
1. The intensity difference ΔL between the coherent signals presented to the transducers.
2. The time-difference ΔT between the coherent signals presented to the transducers.
3. A combination 1 and 2.

3.3.1 Stability of Stereophonic Images

In real stereophonic reproduction, the loudspeakers are at a finite distance to the ears, quite often in the near-field, and spherical waves are produced. In a similar vein to the analysis of Wittek [Wittek, 2007], such a pressure field illustrating the interference of time-based stereo at 440Hz is shown in Figure 3.14. In this simulation, there is no level difference in the loudspeaker signals, but a time-bias of 0.26ms is introduced. The constructive and destructive interference due to the delayed loudspeaker signal results in opposing pressures on either side of the head, seen in the snapshot as a pressure null (in black) on the right side of the head, and a pressure peak (in white) on the left. In contrast a snapshot of the pressure field created by loudspeaker signals with intensity differences only shows a closer relationship in the limited region of the head to the monophonic reference source, than the time-based case.

However examining the pressure at a single frequency, such as in this case, is too trivial. Instead we must look at the ITDs and ILDs due to broadband stereophonic source. The interaural cross-correlation values for a perceived 10° phantom source image generated by finite distance loudspeakers (1m) is documented in Figure 3.15. Here, the anechoic HRTFs of a KEMAR [Algazi et al., 2001] binaural microphone from sources at ±30° are subject to the time and level biased broadband loudspeaker signals and the resultant cross-correlation is calculated in critical bands according to Equation 2.14. The IACC in each band is normalized for presentation. The peak of the correlations (marked in red) are representative of the interaural time difference. We see here how the ITD for time-based stereophony is significantly different in comparison to the ITD of the monophonic reference (also at 10°, 1m). For low frequencies the interchannel time difference has little effect on creating interaural time differences, and it appears as though phantom source formation is at 0° centre. However, as the frequency increases, the ITD causes the image to move from the centre to the right, since P_R (the right ear pressure) $> P_L$ (the left ear pressure) in this frequency range. Conversely, for intensity-based stereophony, the ITDs match the mono source reference well up to approximately 1200 Hz. Above this the head-shadowing inhibits the summing of the loudspeaker signals at the ears, and as a consequence, correct interaural phase is not possible. However, given that ILDs are the more dominant localization cue at higher frequencies, the resultant localization may be unaffected.

ILDs for two-channel stereophony are shown in Figure 3.16. For time-based stereophony, comb-filtering effects are shown, which are particularly prominent in the ipsilateral ear, resulting in large fluctuations in the interaural level difference. For intensity-stereo, the ILD bias is correct, but we note that there are significant spectral deviations in the high frequency ILD, due to comb filtering, which will also lead to perceived coloration of the phantom source. Below 1200Hz we note that the spectrum is highly comparable to the monophonic reference. We can conclude from this analysis that intensity panning results in more accurate and stable localization than time-panning due to the greater spectral coloration and ITD inaccuracy exhibited in the time-panning



(a) Monophonic Reference Source

(b) Time-Based Stereophony

(c) Intensity Based Stereophony

(d) Time-intensity based stereophony

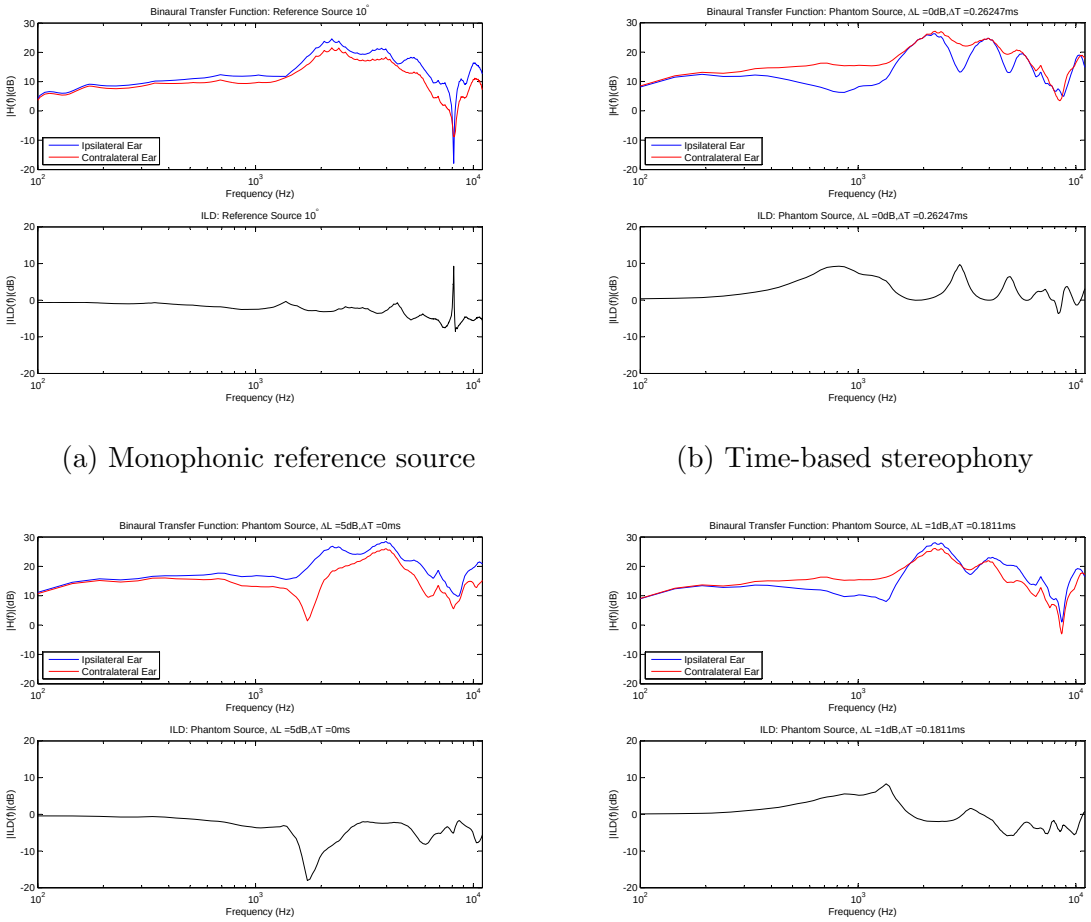
Figure 3.15: *Comparison of IACC created by two channel stereophonic reproduction methods to reference (broadband) monophonic sound at* 10° *(ITD is marked in red).*

case.

A simplified mathematical basis for time and level differences in loudspeaker reproduction has been presented in [Lipshitz and Vanderkooy, 1987], where it is shown how the phantom source image position due to the level of the loudspeaker signals is governed by the equation

$$\sin\theta = \frac{L-R}{L+R}\sin\theta_0 \tag{3.1}$$

where L and R are the left and right loudspeaker signals and θ_0 is the desired source angle. This



(a) Monophonic reference source

(b) Time-based stereophony

(c) Intensity Based Stereophony

(d) Time-intensity based stereophony

Figure 3.16: *Comparison of ILD between two channel stereophonic reproduction methods to reference (broadband) monophonic sound (at* $10°$*).*

is known as the 'stereophonic law of sines', first postulated in [Clark et al., 1958].

3.3.2 Recording for Intensity Stereo

Coincident microphone techniques were first proposed by Blumlein in support of his intensity stereo theory. The most famous and widely adopted of these are the 'Blumlein technique', and the 'mid-side technique'. For the Blumlein technique, 'figure of eight' [1] polar pattern microphones are used, which have maximum gain for signals in front and behind the microphone, with zero amplitude for signals at the side of the microphone. Signals recorded at the back of the microphone are out of phase with those recorded at the front. Two such microphones are used in the setup, as shown in Figure 3.17. This technique is well suited to natural recording scenarios as it gives a uniform spread of reverberant energy across the stereo image [Gerzon,

[1] Also known as 'Velocity' or 'Pressure gradient' microphones.

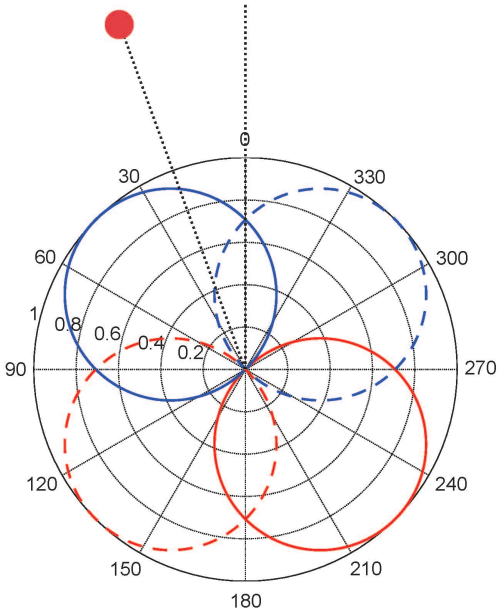


Figure 3.17: *Blumlein recording configuration using two figure of eight microphones. The microphones are angled* 90° *relative to each other.*

1976]. Further information on this technique, as well its mathematical basis satisfying equation 3.1 can be found in [Lipshitz and Vanderkooy, 1987].

Of more significant interest to the work in this thesis is the 'mid-side technique'. Here a microphone with a broad directional characteristic (usually omnidirectional or forward facing cardioid) is used to capture the entire sound pickup and is angled symmetrical to the 0° incident line of the source, whilst a pressure gradient microphone is placed to the side, with its null point facing the direction of the source. This is shown in Figure 3.18(a). The signals from the omnidirectional microphone M and the pressure gradient, S, are then subject to the sum and difference equations below to form the left and right stereophonic channels given as

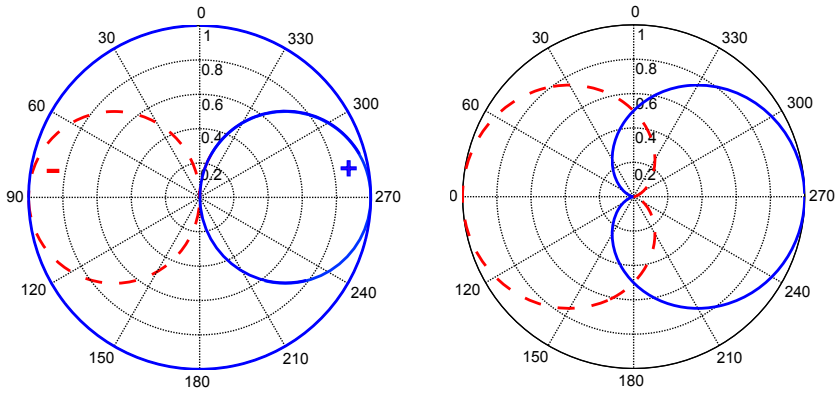
$$\begin{aligned} L &= M + S \\ R &= M - S \end{aligned} \tag{3.2}$$

In practice, the two microphone signals are fed to two individual channels of a mixing console, with a polarity reversed (180° phase shift) copy of the side signal fed to a third channel. The pressure signal is sent to both channels of the stereophonic bus[2] and the side-signal and its inversion to the left and right channels of the stereophonic bus respectively. The result is the same as two coincident *virtual* cardioids, one facing left of the pressure-gradient null-axis and the other facing right as shown in Figure 3.18. The stereophonic width of the mid-side technique can

[2]Stereophonic bus: Two mono channels on a mixing console, usually at the output stage, treated as a stereophonic group, where one channel is the left stereo signal and the other the right stereo signal.

V	k_0	k_1
Omnidirectional	1	0
Sub-cardioid	0.75	0.25
Cardioid	0.5	0.5
Supercardioid	0.333	0.666
Hypercardioid	0.25	0.75
Bidirectional	0	1

Table 3.1: *Weightings for virtual microphone formation using equation 3.3.*



(a) Mid-Side Configuration (b) Resultant Virtual Cardioids

Figure 3.18: *Mid-Side Microphone Configuration.*

be adjusted by changing the ratio of the mid and side signals. The result is that the directivity of the virtual microphone changes. In this way any virtual microphone directionality function V between figure-8 and omnidirectional can be achieved using the weighted sum of a pressure and pressure gradient with

$$V = k_0 + k_1 \cos(\theta) \tag{3.3}$$

where k_0 and k_1 are the weights of the pressure and pressure gradient microphones respectively, and θ is the incident angle to the source. Table 3.1 shows the most common weights for k_0 and k_1 and their resultant polar plots, are shown in Figure 3.19.

3.3.3 VBAP: Stereophony for Arbitrary Numbers of Loudspeakers

In the context of immersive VAEs, we must recognize that if stereophonic reproduction is to be considered as a viable way of forming phantom sources in three dimensions, then it is necessary to extend its principles beyond simple two channel setups. It is a common misconception that stereophonic reproduction refers to a sound system based solely on two loudspeakers. In fact, the word 'Stereophony' refers to the creation of a 'solid' three-dimensional acoustic image for the listener, using *two or more* loudspeakers. Two channel stereo is limited in that the whole of

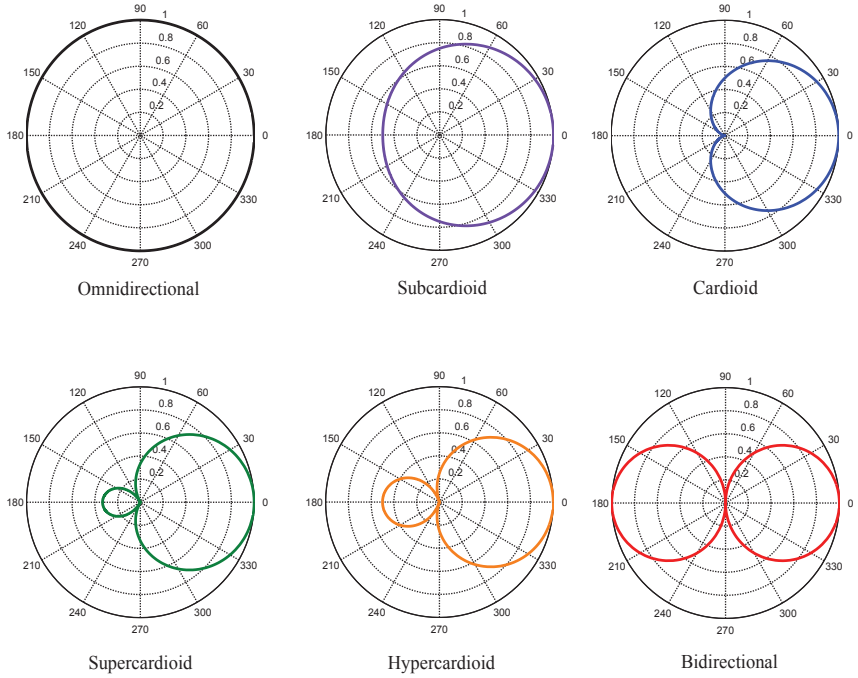


Figure 3.19: *Virtual microphone directionality in mid-side recording.*

the 3-D acoustic environment has to be mapped onto the space between the speakers. The quest for 'realism' in stereophony has led to the design of systems which attempt to produce phantom sources using multichannel setups that completely surround the listener. The limitations of quadraphonic systems, which utilize 4 loudspeakers in laterally positioned pairs, were discussed by Ratliff [Ratliff, 1974]. It was found that quadraphonic systems are unable to create reliable phantom images, in particular for lateral imaging. This problem is attributed to the difficulty in creating interaural differences of any significance from differences between loudspeaker pairs to the same sides of the head [Rumsey, 2001]. Theile et al. [Theile and Plenge, 1977] also carried out a number of experiments to investigate phantom imaging with loudspeaker pairs (using the optimal stereophonic loudspeaker aperture of 60°) at various lateral displacements. The results of these tests confirm Ratliff's findings that phantom images cannot be located at ±90°. This led to the conclusion that a six-speaker array (at ±30°, ±90° and ±150°) was the optimal arrangement for an "all round effect" [Theile and Plenge, 1977], as this would provide a physical loudspeaker at the lateral 90° directions. The results of a listening test using this array with noise impulses and male speech indicate that reasonably good localization was achieved for a centrally positioned listener in an anechoic environment [Theile and Plenge, 1977].

Another further consideration in the amplitude panning case is how to control the loudspeaker gains in order to pan virtual sources over an arbitrary number of loudspeakers. Vector Base Amplitude Panning (VBAP), a vector reformulation of amplitude panning method, is one way to achieve this. Here, the direction of each individual loudspeaker from the perspective of

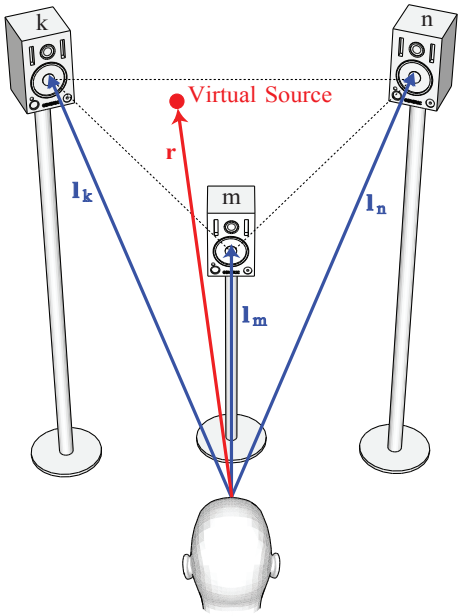


Figure 3.20: *Vector Base Amplitude Panning.*

the listening position is expressed as cartesian unit vector $\mathbf{l}_n = [l_{n1} \quad l_{n2} \quad l_{n3}]^T$. The panning direction in between the three loudspeakers (denoted loudspeakers n, m and k) is then given by another vector $\mathbf{r} = [r_n \quad r_m \quad r_k]^T$. As an example of the method, let us consider a triplet-wise pan using a 3-channel stereophony such as that shown in Figure 3.20. The vectors describing the loudspeakers can be denoted $\mathbf{l}_n$, $\mathbf{l}_m$ and $\mathbf{l}_k$. $\mathbf{r}$ can then be described as a linear combination of each loudspeaker vector by

$$\mathbf{r} = g_n\mathbf{l}_n + g_m\mathbf{l}_m + g_k\mathbf{l}_k \tag{3.4}$$

where g_n, g_m and g_k are the loudspeaker gain factors. If $\mathbf{g} = [g_n \quad g_m \quad g_k]$ and $\mathbf{L}_{nmk} = [\mathbf{l}_n \quad \mathbf{l}_m \quad \mathbf{l}_k]^T$, then we can solve for $\mathbf{g}$ with

$$\mathbf{g} = \mathbf{r}^T\mathbf{L}_{nmk}^{-1} = \begin{pmatrix} r_n & r_m & r_k \end{pmatrix} \begin{pmatrix} l_{n1} & l_{n2} & l_{n3} \\ l_{m1} & l_{m2} & l_{m3} \\ l_{k1} & l_{k2} & l_{k3} \end{pmatrix}^{-1} \tag{3.5}$$

The gain factors can then be normalized such that the loudness of the virtual source is kept constant. Schemes for this normalization are shown in Section 4.2.2.4. The loudspeakers themselves should be arranged equidistant from the listener position. If this is not possible due to logistics in the listening environment, then suitable delay should be applied to the loudspeakers closest to the sweet spot.

VBAP refers only to the reproduction of virtual sources and is not in itself a complete system for recording and reproduction of spatial audio. However, Pulkki has outlined methods for the decomposition of recordings made with soundfield microphones into separate diffuse

and directional components which can be represented on reproduction via VBAP [Merimaa and Pulkki, 2005]. The directional coding technique behind this method is discussed further in Chapter 7. The localization accuracy for VBAP has the potential to be quite high, since in an N-channel reproduction, the localization error will be limited to the space between two loudspeakers. However, for low numbers of loudspeakers, VBAP can also suffer the standard pairwise panning downfalls, such as poor side imaging, or the sensation of phantom sources 'jumping' from loudspeaker to loudspeaker. Source coloration due to comb-filtering will also be apparent in VBAP, in particular with moving sources.

3.3.4 Stereophony Summary

We conclude that stereophonic reproduction of spatial audio is highly applicable to VAE presentations. Spatialization can be achieved in three main ways; time-bias, intensity bias, and time-intensity bias of the loudspeaker signals. Phantom images can be formed by artificially introducing such biases on monophonic signals fed to two or more loudspeakers, or by microphone techniques designed to exploit these interchannel differences. In general, the formation of phantom images over loudspeakers using intensity-based stereophony leads to low frequency ITD information that correlates well with real sources. In contrast, time-based stereophony is unstable for the formation of coherent phantom sources. Combined time and intensity panning (or recording) leads to good perceptual results, but there are still many questions regarding the role of localization cues in using such techniques that cannot be accounted for with summing localization models.

The extension of intensity stereophony to larger numbers of loudspeakers is made possible using VBAP. It follows that for surround sound presentations, VBAP is a viable method of creating virtual sources from monophonic source material. However, there are also possible limitations for VBAP (and stereophony in general) as a reproduction method for VAEs. Firstly, the performance of stereophony under normal listening conditions, and for off-centre listeners needs to be evaluated, in particular in comparison to other conventional spatialization techniques. Furthermore, pre-formed stereophonic presentations require N channels of information, where N is the number of loudspeakers used in the formation of phantom images around the listener(s). For example, stereophonic techniques employed for 7 channel surround sound require 7 discrete stereophonic information channels. However, there exists another spatialization technique, which has its origins in intensity-based stereophony, but which also has a low information channel count and can be adapted to any number of loudspeakers. This is known as Ambisonics.

3.4 Ambisonic Soundfields

In its simplest form, Ambisonics can be viewed as an extension of the mid-side technique shown in Section 3.3.2 to three dimensions. Previously, we encountered the concept of the virtual

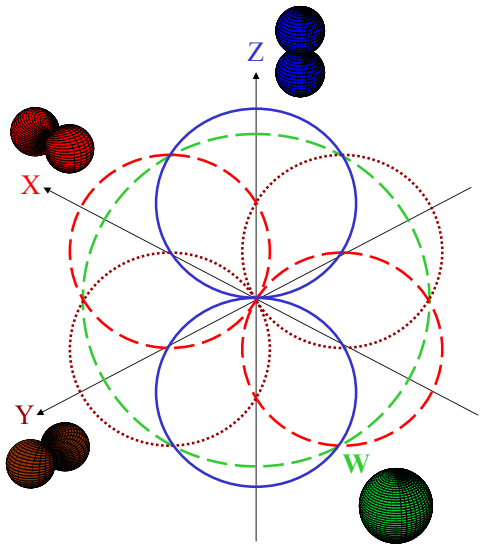


Figure 3.21: *Ambisonic B-Format soundfield representation. X,Y, and Z represent the velocity components in the x, y, and z directions respectively and W represents the omnidirectional pressure component.*

microphone, created from the summation of a pressure and velocity microphone. Now let us look again at the mid-side technique and consider an omnidirectional pressure microphone and a side facing velocity microphone as giving us directional information in the Y-direction. Then, if we add in another velocity receiver with a front-back orientation (X-direction) we can then describe the soundfield in a two-dimensional space. Furthermore, if we add in a third, upwards facing (Z-direction) velocity receiver, the formation of a virtual microphone in any direction becomes possible. This coincident microphone configuration is shown in Figure 3.21.

Thus, for a first-order Ambisonic system, we have four signals which characterize the complete soundfield. What is important to remember is that these four signals are not fed to loudspeakers directly. Instead a linear combination of these signals is used to form virtual microphones whose output is fed to an associated loudspeaker. In essence then, the encoding format is completely independent of the reproduction format. The virtual microphone can be steered in any direction (including height) and have any polar pattern ranging from figure-8 to omnidirectional. This is known as first order 'B-Format' Ambisonics, a legacy term referring to the encoding stage of the Ambisonic reproduction chain originally envisaged by Gerzon and Fellgett [Gerzon, 1975a], [Fellgett, 1975]. The B-Format signals describe the position of the sound source on a unit sphere, and as such, are known as *spherical harmonic components*. They are more than just a coincident microphone technique however, since the position of any source

S can be encoded as B-Format using

$$W = s \times b_w = s \tag{3.6}$$
$$X = s \times b_x = s \times \sqrt{2}\cos\theta\cos\phi \tag{3.7}$$
$$Y = s \times b_y = s \times \sqrt{2}\sin\theta\cos\phi \tag{3.8}$$
$$Z = s \times b_z = s \times \sqrt{2}\sin\phi \tag{3.9}$$

where b_w, b_x, b_y, and b_z are directional cosine weighting factors, describing the direction of the encoded plane wave. If only horizontal information is required then the Z component can be disregarded.

One of the main advantages of B-Format Ambisonic encoding is the transformations that can be applied to the soundfield in post processing. If we wish to rotate the soundfield about the Z-axis, we can transform the W, X, and Y signals such that

$$W' = W \tag{3.10}$$
$$X' = X\cos(\theta) + Y\sin(\theta) \tag{3.11}$$
$$Y' = Y\cos(\theta) - X\sin(\theta) \tag{3.12}$$
$$Z' = Z \tag{3.13}$$

For rotation about X (tilt) we have

$$W' = W \tag{3.14}$$
$$X' = X' \tag{3.15}$$
$$Y' = Y\cos(\theta) - Z\sin(\theta) \tag{3.16}$$
$$Z' = Y\sin(\theta) + Z\cos(\theta) \tag{3.17}$$

And similarly for Y (tumble),

$$W' = W \tag{3.18}$$
$$X' = X\cos(\theta) - Z\sin(\theta) \tag{3.19}$$
$$Y' = Y' \tag{3.20}$$
$$Z' = Z\cos(\theta) + X\sin(\theta) \tag{3.21}$$

Furthermore, it is also possible to zoom-in on part of the soundfield. This is what Gerzon referred to as 'forward dominance' and a control for this can be found on most B-Format encoders. For this, the horizontal only B-Format signals can be reinterpreted as

$$W' = W + \frac{1}{\sqrt{2}} \times d \times X \tag{3.22}$$
$$X' = X + \sqrt{2} \times d \times W \tag{3.23}$$
$$Y' = Y\sqrt{1-d^2} \times Y \tag{3.24}$$
$$Z' = \sqrt{1-d^2} \times Z \tag{3.25}$$

where d is the dominance parameter ranging between -1 and 1. Forward dominance can also be seen as a type of width control, since it changes the directional distribution and azimuths of sounds. For example, if the forward dominance gain d is approximately 0.1, the images 90° at the sides move forward by an angle of 5°.

3.4.1 Recording for Ambisonics

Since a totally coincident microphone array consisting of three velocity microphones and a pressure microphone is physically impossible to achieve, Soundfield recording in 1st order Ambisonics is accomplished using a coincident tetrahedral microphone array [Craven and Gerzon, 1977]. This is commonly referred to as an A-Format Ambisonic microphone. The A-format microphone captures four signals from four sub-cardioid capsules mounted on the left-front (LF), right-front (RF), left-back (LB) and right-back (RB) of the tetrahedron. B-format signals can be derived from an A-format microphone using the sum and difference equations:

$$W = 0.5\ (LF + LB + RF + RB) \tag{3.26}$$

$$X = 0.5\ ((LF - LB) + (RF - RB)) \tag{3.27}$$

$$Y = 0.5\ ((LF - RB) - (RF - LB)) \tag{3.28}$$

$$Z = 0.5\ ((LF - LB) + (RB - RF)) \tag{3.29}$$

These equations are typically implemented in hardware A-Format to B-Format converters such as the Soundfield MKV control unit [Soundfield Ltd., 2009]. An example of a tetrahedral soundfield microphone is shown in Figure 3.22. Typically, the orientation of the soundfield microphone corresponds to the classic X (positive forward), Y (positive due-left), and Z (positive upwards)



Figure 3.22: *Tetrahedral microphone capsule used in soundfield microphone (Courtesy Soundfield Ltd.).*

directions, but the engineer has complete freedom in post production to rotate, tilt and tumble the soundfield using Equations 3.10, 3.14 and 3.18.

3.4.2 Decoding Ambisonics

The encoding stage represents only one part of the Ambisonic process, since Ambisonics also refers to the reproduction of the encoded/recorded audio. We can decode Ambisonics to different loudspeaker configurations, but there are several constraints on the loudspeaker layout which must be imposed depending on the decoding strategy, which will be described in this section. There are two main schools of thought in Ambisonic decoding. These are decoding through *projection* and decoding through *pseudo-inverse*.

3.4.2.1 Decoding Through Projection

A basic Ambisonic decoder works by supplying each loudspeaker a weighted sum of all Ambisonic signals. The value of the weight is dependent on the value of the spherical harmonic representation of the loudspeaker position on the unit sphere. That is

$$v_n = \frac{1}{N}\left[l_{w_n}W + l_{x_n}X + l_{y_n}Y + l_{z_n}Z\right] \tag{3.30}$$

where v_n is the n^{th} loudspeaker feed, N is the number of loudspeakers, and each l is a directional cosine weighting factor, describing the position of the loudspeaker on the unit sphere in the same manner as the source signal encoding, thus

$$l_{w_n} = 1 \tag{3.31}$$

$$l_{x_n} = \sqrt{2}\cos\theta_n\cos\phi_n \tag{3.32}$$

$$l_{y_n} = \sqrt{2}\sin\theta_n\cos\phi_n \tag{3.33}$$

$$l_{z_n} = \sqrt{2}\sin\phi_n \tag{3.34}$$

where θ_n and ϕ_n represent the position of the n^{th} loudspeaker on the unit sphere. One notes that equation 3.30 represents the summation of the pressure component and three pressure-gradient components, resulting in a virtual microphone formation in the same manner as in equation 3.3. This virtual microphone points from the centre of the array to the direction of the loudspeaker and has a supercardioid directivity pattern. If $v_n = sg_n$, where g_n is the loudspeaker gain, and s is the source signal to encode into B-Format, then we can rewrite equation 3.30 purely in terms of the loudspeaker gain as

$$g_n = \frac{1}{N}\left[l_{w_n}b_w + l_{x_n}b_x + l_{y_n}b_y + l_{z_n}b_z\right] \tag{3.35}$$

where b_w, b_x, b_y and b_z represent the source plane wave directional cosine weightings.

The decoding process presented here assumes regular loudspeaker layouts, but this is not always the case. It is therefore termed 'decoding through projection' since the Ambisonic decoder

will always project the decoded signals onto the loudspeakers, even if the array is irregular. In such cases, a non uniform soundfield is to be expected on reproduction and localization errors ensue. However, the ability to change the virtual microphone polar patterns forms the basis for psychoacoustically optimized Ambisonic decoders, which are discussed in Section 3.4.3.

3.4.2.2 Decoding Through Pseudoinverse

It is extremely useful to consider the Ambisonic decoding process in terms of matrix notation. First, let us consider a plane wave source encoded into B-Format using equations 3.7 to 3.9

$$\begin{aligned} \mathbf{B} &= [W, X, Y, Z]^T && (3.36)\\ &= s\ [b_w, b_x, b_y, b_z]^T && (3.37)\\ &= s\ \mathbf{b}^T && (3.38) \end{aligned}$$

where S is the source signal and $\mathbf{b}$ is the directional B-Format vector. In the resultant Ambisonic decode, we wish to obtain a set of loudspeaker gains $\mathbf{g}$ which, when multiplied by a given re-encoding matrix, $\mathbf{L}$, would yield the B-format direction $\mathbf{b}$. That is

$$\mathbf{b} = \mathbf{L}\mathbf{g} \tag{3.39}$$

and therefore

$$\mathbf{g} = \mathbf{L}^{-1}\mathbf{b} \tag{3.40}$$

The matrix $\mathbf{L}$ represents the positions of the loudspeakers given by

$$\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, ..., \mathbf{l}_N]^T \tag{3.41}$$

where each element in the column vectors $\mathbf{l}_n = [l_{w_n}, l_{x_n}, l_{y_n}, l_{z_n}]^T$ is obtained from loudspeaker encoding equations 3.31 to 3.34. $\mathbf{L}^{-1}$ is the inverse of $\mathbf{L}$ and is known as the decoding matrix. However, to invert $\mathbf{L}$ we need the matrix to be a square which is only possible when the number of Ambisonic channels is equal to the number of loudspeakers. When the number of loudspeaker channels is greater than the number of Ambisonic channels, which is usually the case, we then obtain the pseudo-inverse of $\mathbf{L}$ where

$$\mathbf{D} = \text{pinv}(\mathbf{L}) = \mathbf{L}^T(\mathbf{L}\mathbf{L}^T)^{-1} \tag{3.42}$$

That is, for a given loudspeaker layout, there is a specific decode matrix $\mathbf{D}$, and in order to derive this decode matrix we obtain the pseudo-inverse of the loudspeaker matrix $\mathbf{L}$. This method minimizes the least squares norm of $\mathbf{g}$, and as a consequence, the total output power of the loudspeaker array. If the loudspeaker matrix $\mathbf{L}\mathbf{L}^T$ is diagonal due to the array geometry, then a trivial solution exists for the decode matrix $\mathbf{D}$, by spatially sampling the soundfield in the direction of the loudspeakers using the aforementioned virtual microphone technique. This represents the method of decoding through projection.

3.4.3 Psychoacoustic Optimization of Ambisonic Decoders

Gerzon [Gerzon, 1977a] realized (as Blumlein had) that the low frequency localization below 700Hz was dominated by phase differences between the ears and that high frequency localization was dominated by intensity difference between the ears. He related the localization mechanisms to vector theory and he defined the low frequency localization vector, known as the 'Makita' localization vector, as the the direction that the head must be oriented such that interaural phase differences are zero [Gerzon and Barton, 1992]. Likewise, the energy vector points in the direction that the head must orientate such that high frequency interaural amplitude differences are zero. The length of each of these vectors is 1 for natural sources, and lesser values result in reduced localization accuracy. In the discussion that follows, we will limit the study of localization vectors to the horizontal plane.

The velocity vector $\mathbf{V} = [v_x \quad v_y]$ can be calculated from the low frequency acoustical pressure gain P by

$$v_x = \sum_{n=1}^{N} g_n \cos(\theta_n)/P_v \tag{3.43}$$

$$v_y = \sum_{n=1}^{N} g_n \sin(\theta_n)/P_v \tag{3.44}$$

where g_n and θ_n are the real gain and angle respectively of the n^{th} loudspeaker, N is the number of loudspeakers and

$$P_v = \sum_{n=1}^{N} g_n \tag{3.45}$$

Similarly the energy vector $\mathbf{E} = [e_x \quad e_y]$ can be calculated from the square of the high frequency gain of each loudspeaker by

$$e_x = \sum_{n=1}^{N} g_n^2 \cos(\theta_n)/P_e \tag{3.46}$$

$$e_y = \sum_{n=1}^{N} g_n^2 \sin(\theta_n)/P_e \tag{3.47}$$

where

$$P_e = \sum_{n=1}^{N} g_n^2 \tag{3.48}$$

The magnitude of the velocity and energy vectors can then be determined by

$$r_v = \sqrt{v_x^2 + v_y^2} \tag{3.49}$$

$$r_e = \sqrt{e_x^2 + e_y^2} \tag{3.50}$$

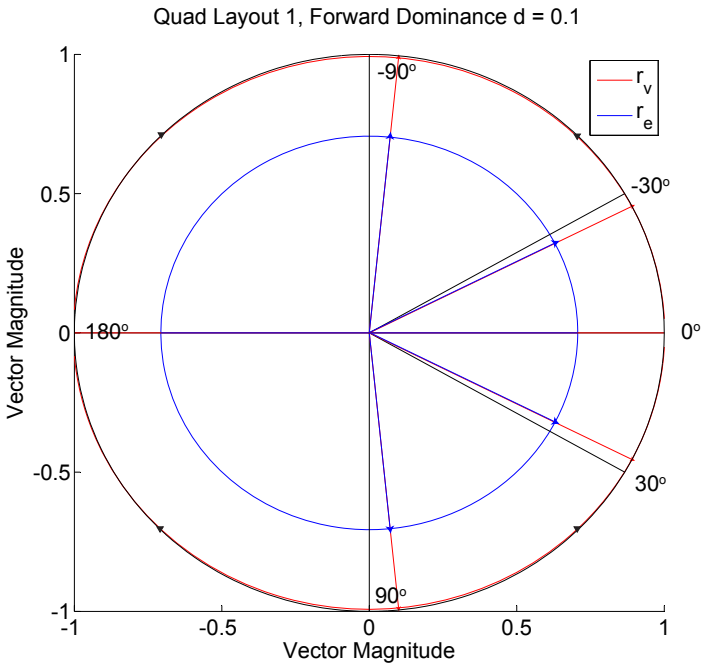


Figure 3.23: *Example velocity and energy vectors for a quadraphonic layout with forward dominance.*

and their corresponding angles of orientation by

$$\theta_v = \tan^{-1}(v_x/v_y) \tag{3.51}$$

$$\theta_e = \tan^{-1}(e_x/e_y) \tag{3.52}$$

An example of the vector magnitudes r_v and r_e for a quad layout are shown in Figure 3.23. The red and blue rings indicate r_v and r_e respectively in the horizontal plane. The vector lines at 0°, ±30°, and ±90° allow us to gauge the angular accuracy of reproduction, as well as see the effect of rotations and forward dominance (which is employed in this example). r_e is broadly indicative of the localization of reproduced sources between 700 and 4000Hz, whereas r_v gives the apparent direction of localization below 700Hz. Since r_e is an average of N vectors with positive coefficients, it will only ever be equal to 1 when the sound comes from a single loudspeaker. For all other directions r_e will always be less than 1, so the aim is to get it as high as possible. For frontal cues, the value of r_e is a strong indicator of image stability whereas r_v is not as critical (providing it lies somewhere between 0.8 and 1.2). Conversely, for images to the side of the head, the value of r_v should be as close to 1 as possible, since low frequency cues for these directions can be made correct for the centrally seated listener.

For regular layouts, Gerzon outlined several criteria to ensure that the velocity and energy vectors give the same localization directions:

1. All loudspeakers are equidistant from the centre of the array

2. All loudspeakers are placed in diametrically opposed pairs

3. The sum of the signals in each diametrically opposed pair is the same for each diametric pair.

This is known as the 'diametric decoder theorem'. The advantages to using this theorem are that the encoded and decoded source angles will always coincide, and that the virtual source energy and pressure do not change for different source angles.

For irregular layouts, such as the ITU-R BS.775, the Ambisonic criteria are generalized as

1. The direction of localization given by θ_v and θ_e should be equal at least up to 4kHz,

2. r_v is as close to 1 as possible for all frequencies below 700Hz,

3. r_e is substantially maximized across the 360° reproduction area as much as possible.

In general, if a decoder does not attempt to meet the aforementioned requirements then it cannot be deemed Ambisonic. In the case of irregular layouts, the directional gain pattern for P can be considered to be non-uniform, and the frontal sound stage can be designed to have better image stability than the rear sound stage. This is achieved using the forward dominance operations of equations 3.22 to 3.25. An exception to the above criteria can be considered in the case of low (first) order off-centre listening situations. In such instances a psychoacoustic decoder design may not work, since the anti-phase components required to maximize the velocity vector result in localization errors for off-centre listeners. Thus, different Ambisonic decoders have emerged based on different localization requirements. These will now be investigated in detail.

3.4.4 Ambisonic Decoder Types

In this section, we will consider different Ambisonic decoder designs in the context of an 8-channel regular array setup as shown in Figure 3.24. In this analysis, we will investigate further the relationships between the ITDs and ILDs, the velocity and energy vectors, and the virtual microphone patterns formed in the Ambisonic decoder designs. There are four main types of 1^{st} Order Ambisonic decoder, and these are listed in Table 3.2. These decoders are formed by changing the ratio of the W component to the X, Y and Z components using the modified form of equation 3.30

$$p_i = \frac{1}{N}\left[k_0 l_w W + k_1 (l_x X + l_y Y + l_z Z)\right] \tag{3.53}$$

where the weights k_0 and k_1 determine the directionality of a virtual microphone and N is the number of loudspeakers. k_0 and k_1 are taken from Table 3.1 and determine the overall localization accuracy in the decoder design according to the original psychoacoustical requirements outlined by Gerzon. The virtual microphones formed from k_0 and k_1 are shown in Figure 3.25.

Name	Frequency	Optimized	k_1/k_0
Velocity Decode	Low	r_v	2
Energy Decode	High	r_e	$\sqrt{2}$
Cardioid Decode	All	In-Phase	1
Shelf Filter Decode	All	r_v and r_e	2 (LF), $\sqrt{2}$ (HF)

Table 3.2: *Decoder weightings for virtual microphone formation using equation 3.53.*

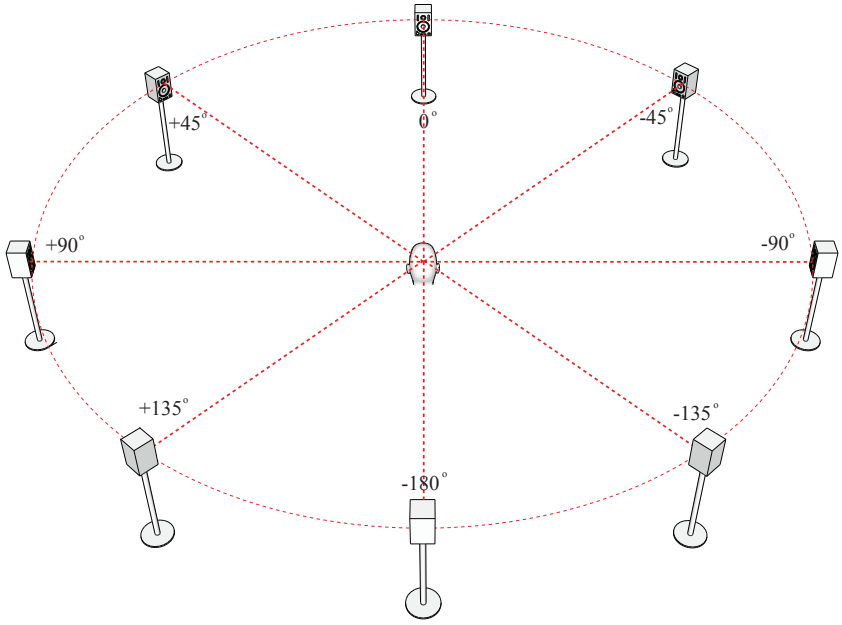


Figure 3.24: *8-Channel regular array setup for Ambisonics investigations.*

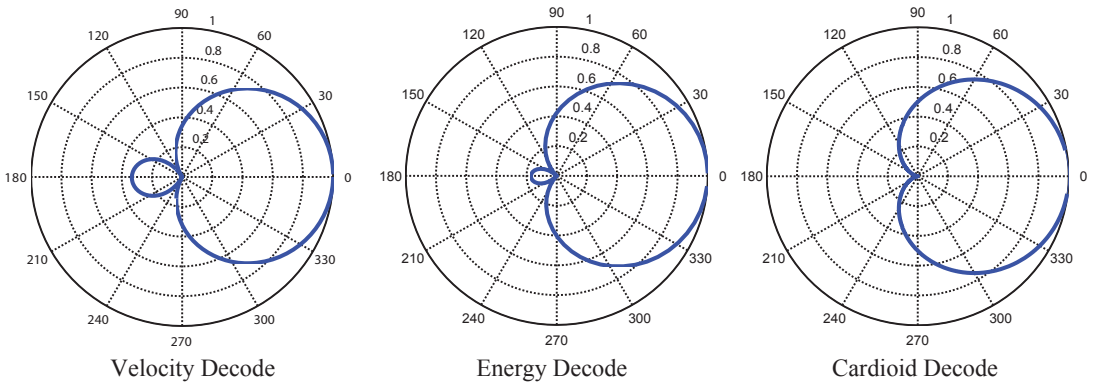


Figure 3.25: *Virtual microphones for different Ambisonic decoding schemes.*

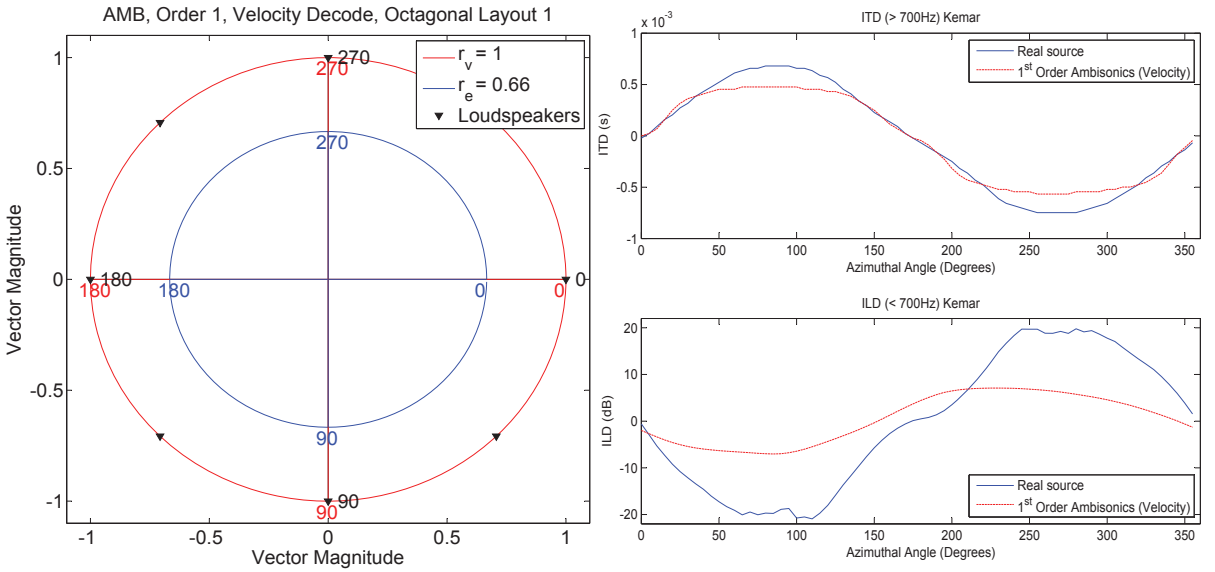


Figure 3.26: *Comparison of velocity and energy vectors for* 1^{st} *Order Ambisonic 'Velocity' decode to an 8-Channel regular array to ITD and ILD.*

The first type of decoder is known as the 'velocity' Ambisonic decoder, which is characterized by supercardioid virtual microphones. This type of decoder maximizes the velocity vector to 1, ensuring good low frequency localization. Figure 3.26 compares the localization vectors of this decoder to the ITDs and ILDs created at reproduction. The ITD and ILD curves are formed by obtaining the loudspeaker feeds for a 1^{st} Order velocity decode and convolving them with the KEMAR (large pinnae) HRTF data pertaining to the loudspeaker directions (i.e. a virtual loudspeaker approach). We note that although a smooth response is obtained in both the ILD and ITD for the Ambisonics decode, the interaural cues for the virtual sources are not equivalent to that of a monophonic source. The deviations in the level and time differences are most noticeable at the ear points, and the ILD is generally low indicating that localization ambiguity in the high frequency range is to be expected.

A comparison of the localization vectors to the true ITD and ILD shows a good correlation: the value of r_v is close to 1, and a good low frequency ITD is achieved, whereas the value of r_e is approximately 0.66, and the ILD localization accuracy is fair in comparison to a true source.

The second decoder type is known as the 'Energy' decoder. As its name suggests, this decoder optimizes the energy vector for better high frequency ILD accuracy. The maximum value of r_e achievable for a first order system is 0.707. This is because a narrower virtual microphone pattern results in a greater angular discrimination with the effect of reduced energy in the direction of the loudspeaker, whereas a wider directional pattern reduces the angular discrimination and hence localization accuracy. A further trade off is that the velocity vector is now also reduced. This is a consequence of the change in the directional microphone pattern reducing the negative loudspeaker gains exhibited in the decode, and hence the pressure differential between

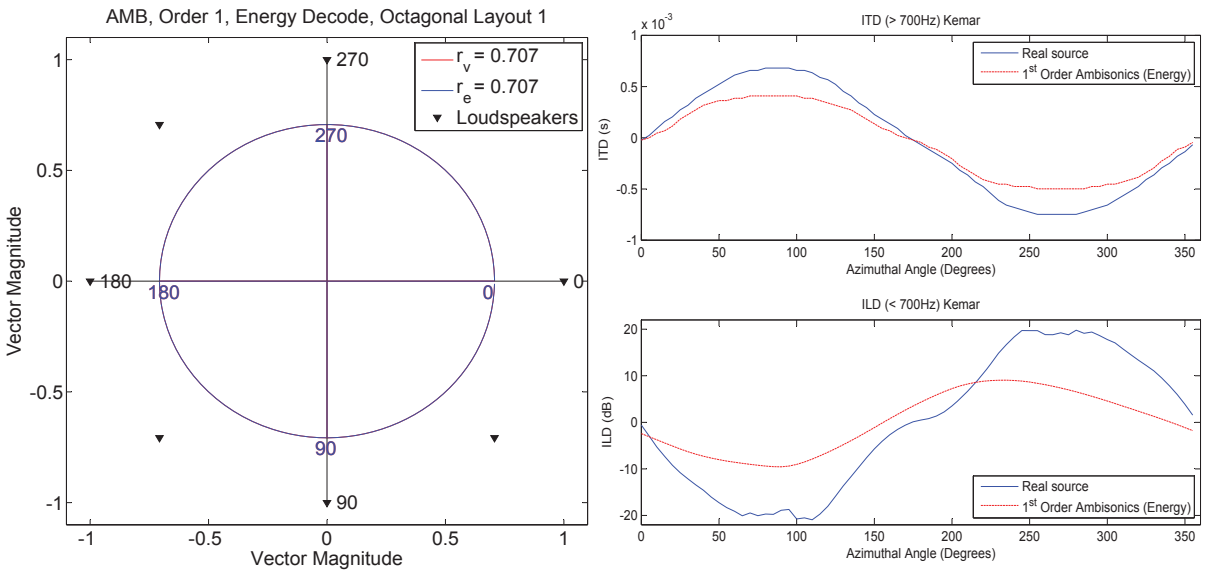


Figure 3.27: *Comparison of velocity and energy vectors for* 1^{st} *Order Ambisonic 'Energy' decode to an 8-Channel regular Array to ITD and ILD. Here* $r_v = r_e$.

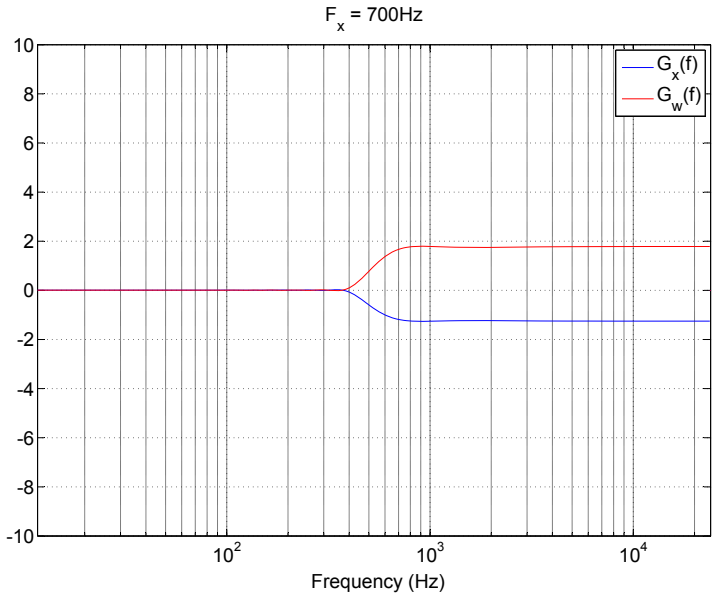


Figure 3.28: *Shelf filters for changing low and high frequency directivity responses.*

diametrically opposed loudspeakers. The resultant ITDs and ILDs confirm what is reported by the localization vectors. The ILD is now closer to that of a real monophonic source, although localization accuracy is still not ideal, and the ITD accuracy is reduced.

We can increase the localization accuracy of the decoder if we change the directivity factor for the low and high frequencies separately, such that the r_v and r_e are as close to 1 as possible. This is achieved by employing 'velocity' decoding below 700Hz, and 'energy' decoding above

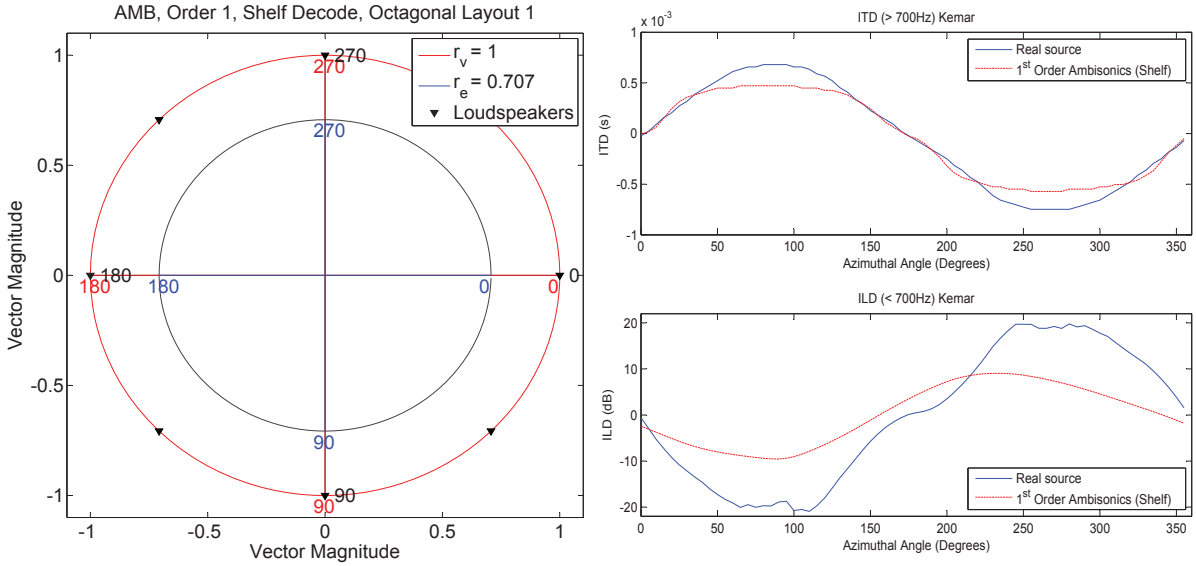


Figure 3.29: *Comparison of velocity and energy vectors for* 1^{st} *Order Ambisonic 'Shelf Filter' decode to an 8-Channel regular Array to ITD and ILD.*

this. In practice, the change between the high and low frequency virtual microphone directivity is achieved by two types of 'shelf filters', one gain compensating the W signal and the other gain compensating the X, Y and Z signals as shown in Figure 3.28 [Gerzon, 1980]. The 3dB difference between the W and XY components at high frequencies ensures the energy decode, whereas the equivalent gain at low frequencies [3] guarantees velocity optimization. The resultant localization vectors are shown in Figure 3.29. The localization accuracy at both low and high frequencies is now maximized within the capabilities of a 1^{st} order Ambisonic system, i.e. $r_v = 1$ and $r_e = 0.707$.

A fundamental issue with the aforementioned decoding schemes is the effect that anti-phase components due to the rear lobes of the virtual microphones has on off-centre listening. Such components form negative loudspeaker gains, which are in the opposite orientation of the microphone, and aid in velocity decoding. Whilst this will improve the localization accuracy at the central seat, it is possible that precedence can occur in the direction of the anti-phase components for off centre listening positions, and that a source can be localized in entirely the wrong direction. A solution to this issue was proposed by Malham [Malham, 1992], and is known as 'cardioid', or 'in-phase decoding'. Here $k_0/k_1 = 1$, resulting in virtual microphones with cardioid polar patterns thereby eliminating negative loudspeaker gains. The consequence is that the level of the velocity vector is lower than that of the energy vector, meaning that ILD plays a strong role in this decode. Since the magnitude of the velocity vector is 0.5, and the energy vector is 0.666 we expect to have reduced localization quality for a centrally seated listener and

[3] There exists a 6dB difference between the pressure and velocity components at low frequencies based on the encoding and decoding equations.

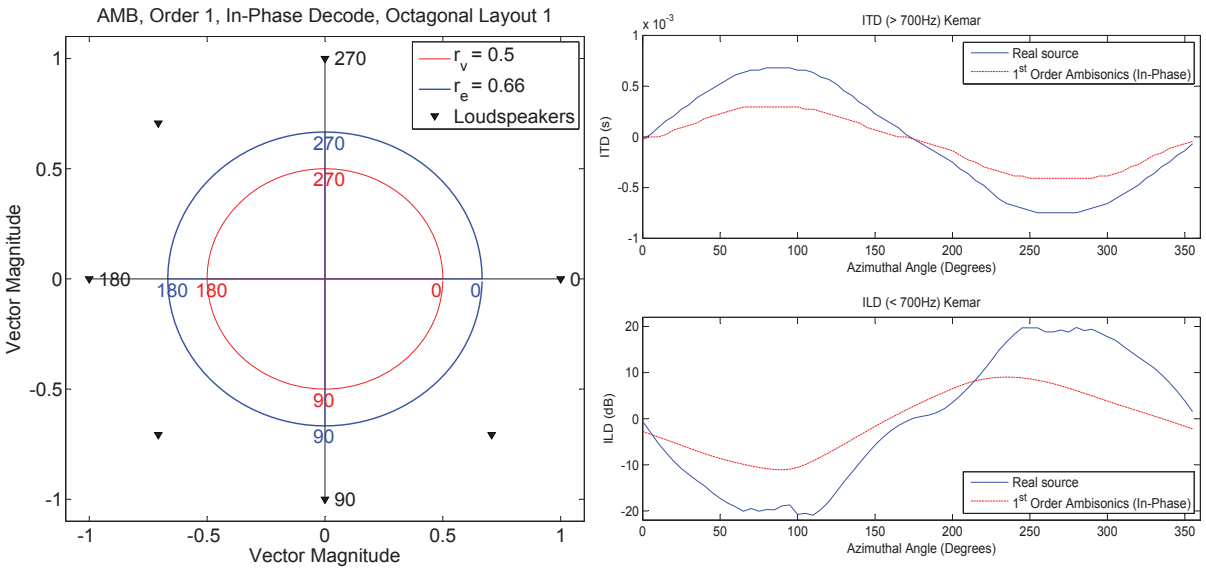


Figure 3.30: *Comparison of velocity and energy vectors for* 1^{st} *Order Ambisonic 'Cardiod' decode to an 8-Channel regular Array to ITD and ILD.*

particularly poor localization quality at the sides. This is confirmed in the ITD and ILD plots of Figure 3.30.

3.4.5 Higher Order Ambisonics

The velocity and pressure components of the Ambisonic system described thus far, are in fact equivalent to first order spherical harmonic basis functions. Thus, so far, we have only considered first order Ambisonic systems. However, higher order spherical harmonic functions can be defined with

$$Y_{mn}^{\sigma}(\theta,\phi) = \sqrt{(2m+1)(2-\delta_{0,n})\frac{(m-n)!}{(m+n)!}}\tilde{P}_{mn}(\sin\phi) \times \begin{cases} \cos n\theta & \text{if } \sigma = +1 \\ \sin n\theta & \text{if } \sigma = -1 \end{cases} \tag{3.54}$$

where m is the order and n is the degree of the spherical harmonic, $\tilde{P}_{mn}(\zeta)$ is the associated Legendre function, and $\delta_{0,n}$ is the Kronecker delta function which is equal to 1 for $n = 0$ and 0 otherwise. Note that the adopted notation Y_{mn} for the spherical harmonics should not be confused with the B-Format channel Y. The spherical harmonic basis functions arising from this equation up to third order are shown in Figure 3.31. For each order m there are $(2m + 1)$ spherical harmonics, including notably, two horizontal components per order (where the degree n = the order m). The first four components can be recognized as the W, X, Y and Z components previously introduced. More formally, these are termed the Y_{00}^{1}, Y_{11}^{1} ,Y_{11}^{-1} and Y_{10}^{1} components respectively. The Ambisonic signals themselves are termed B_{00}^{1}, B_{11}^{1} ,B_{11}^{-1} and B_{10}^{1}. Given these

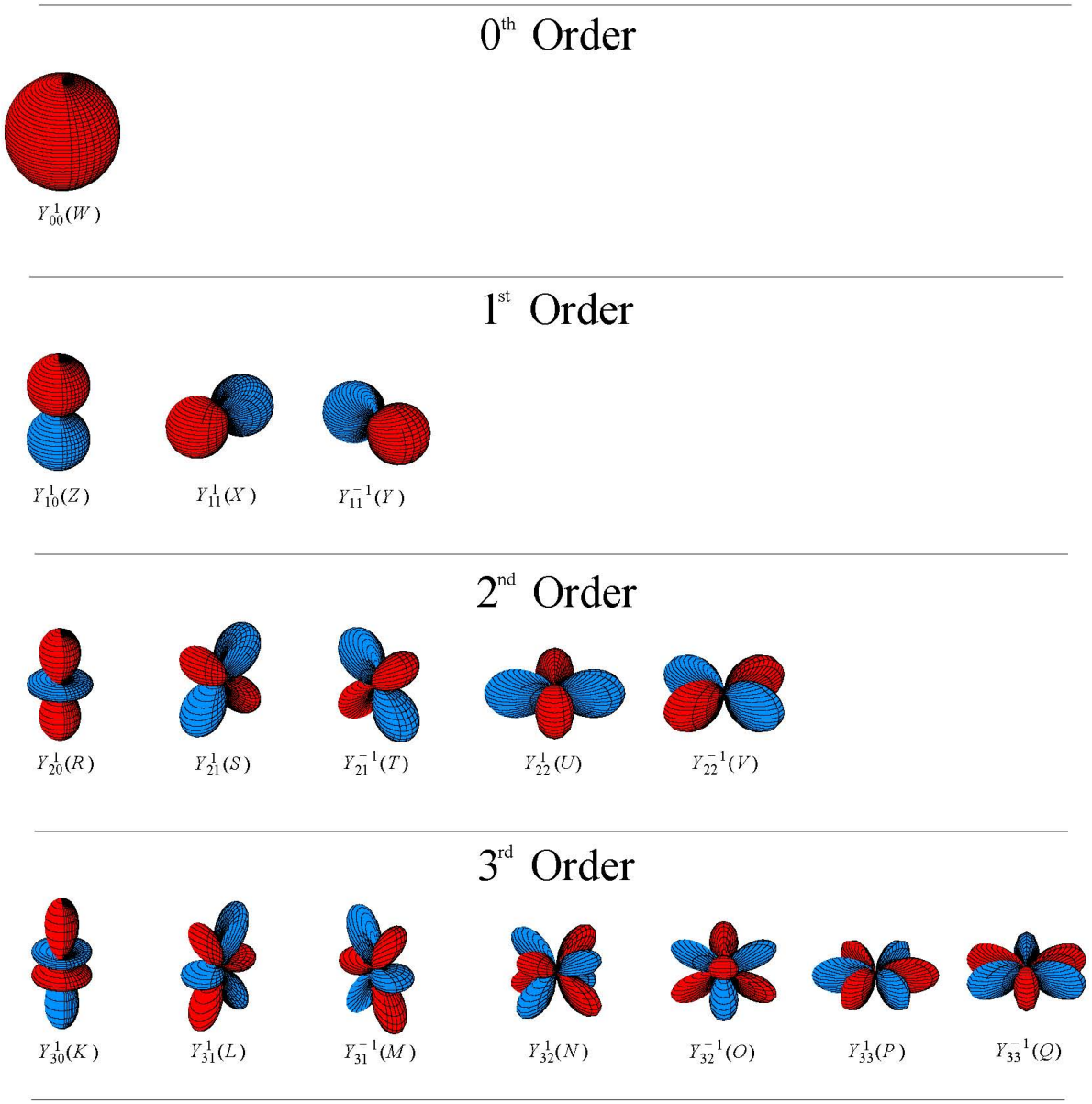


Figure 3.31: *Table of 0^{th}, 1^{st}, 2^{nd} and 3^{rd} Order Spherical Harmonic Basis Functions.*

spherical harmonic components, the in-coming pressure-field can then be written as a Fourier-Bessel series, whose terms are the weighted product of radial functions and spherical harmonic components [Daniel, 2003a]

$$p(\mathbf{r}) = \sum_{m=0}^{\infty} j^m j_m(kr) \sum_{0 \leq n \leq m, \sigma = \pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta, \phi) \tag{3.55}$$

where $k = (2\ \pi\ f)/c$ is the wave number, r is the radius of the spherical coordinate system, $\mathbf{r}$ is a coordinate vector on the pressure field and $j_m(kr)$ are the Bessel functions. The Bessel functions $j_m(kr)$ are shown in Figure 3.32. It is interesting to note that as we add in higher

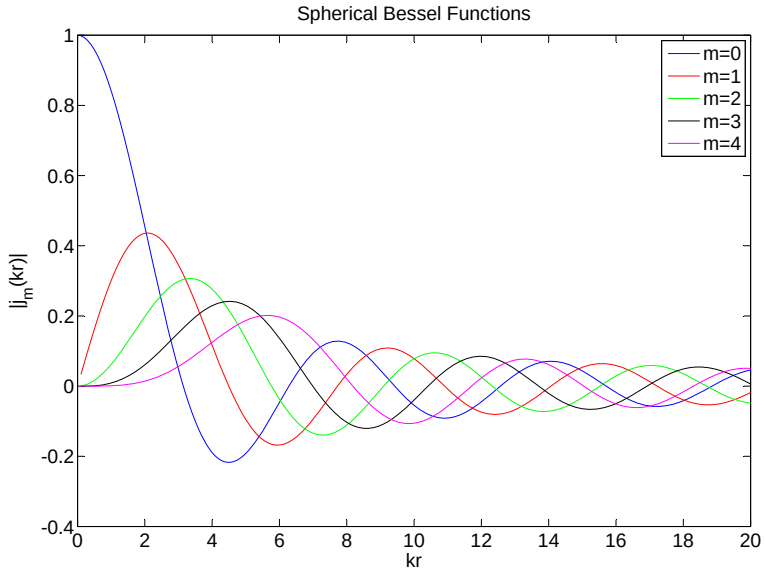


Figure 3.32: *Spherical Bessel Functions.*

order spherical harmonics to the Fourier-Bessel series, the associated Bessel function displays maxima that occur at large distances of kr. Thus, as we move up in spherical harmonic order, we achieve a greater radial expansion. This can be demonstrated if we consider the plane wave case. Figure 3.33 shows the 1st, 3rd and 5th order Ambisonic representations of a 440 Hz plane wave at 10°. We can readily see that as we move up in order, the area of correct soundfield reconstruction becomes greater.

It is important to remember that in order for us to implement an m^{th} order system with periphony, we need to include all the harmonics up to and including that order, yielding $(m+1)^2$ components. Consequently this is also the minimum number of loudspeakers required for the reproduction. If we wish to consider horizontal only reproduction however, the Fourier-Bessel series reduces to

$$p(r,\theta) = B_{00}^{1} J_0(kr) + \sum_{m=1}^{\infty} \left(B_{mn}^{+1} \sqrt{2} \cos m\theta + B_{mn}^{-1} \sqrt{2} \sin m\theta \right) J_m(kr) \tag{3.56}$$

This is referred to as the cylindrical harmonic formalism, and requires $(2m+1)$ harmonics for an m^{th} order encode. To summarize, for horizontal only reproduction we require that the minimum number of loudspeakers N be equal to and

$$N = 2m + 1 \tag{3.57}$$

and for periphonic array reproduction,

$$N = (m+1)^2 \tag{3.58}$$

However, for regular loudspeaker arrays satisfying the diametric decoder theorem, equation

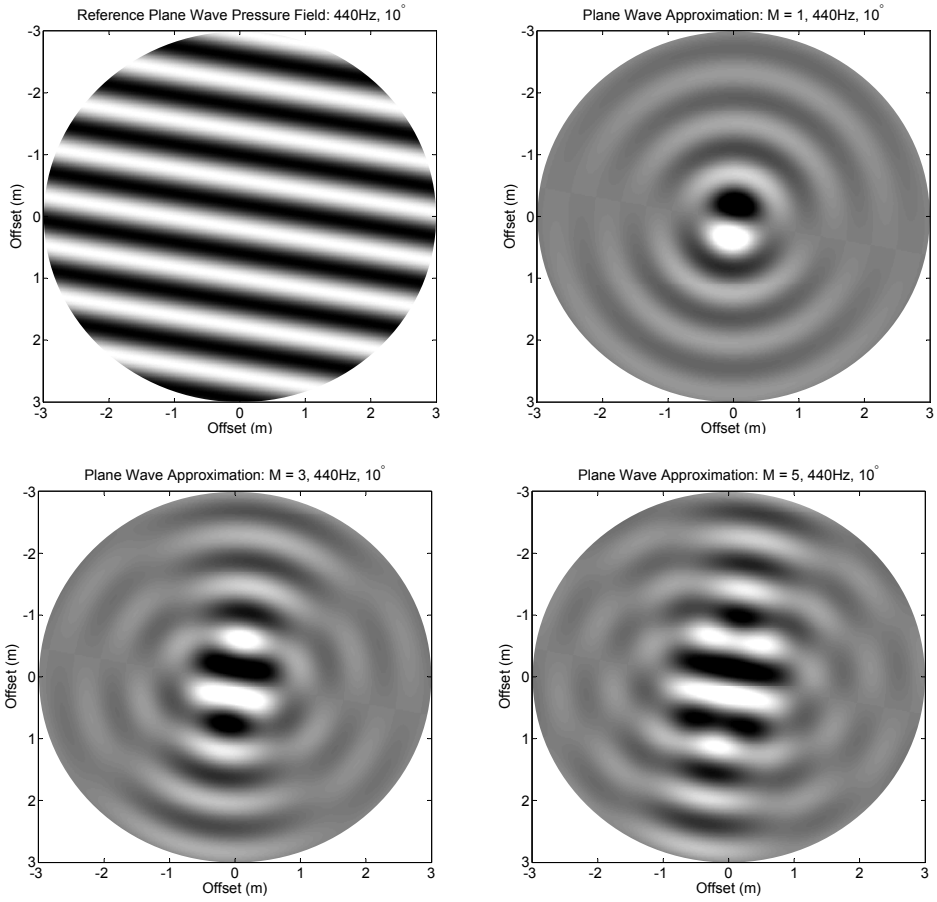


Figure 3.33: 1^{st}, 3^{rd} *and* 5^{th} *Order Ambisonic plane wave approximation.*

3.57 becomes

$$N = 2m + 2 \tag{3.59}$$

For periphonic setups satisfying the diametric decoder theorem, we can use equation 3.57 for odd Ambisonic orders and the following for even Ambisonic orders

$$N = (m + 1)^2 + 1 \tag{3.60}$$

In practice, this leads to an insufficient number of loudspeakers for stable Ambisonic decoding (e.g. only four loudspeakers for full sphere decode). Good decoding with diametrically opposed loudspeakers is guaranteed with

$$N = 2(m + 1)^2 \tag{3.61}$$

One can consider the issue of the minimum number of loudspeakers as being equivalent to the digital sampling problem. Figure 3.34 shows the loudspeaker gains for an infinite number of loudspeaker channels for 1^{st} to 3^{rd} order Ambisonics. We see that although the peak of the loudspeaker gains is at the desired source position (180° in this case), the multiple positive maxima of each order (starting from second order) occur at different positions.

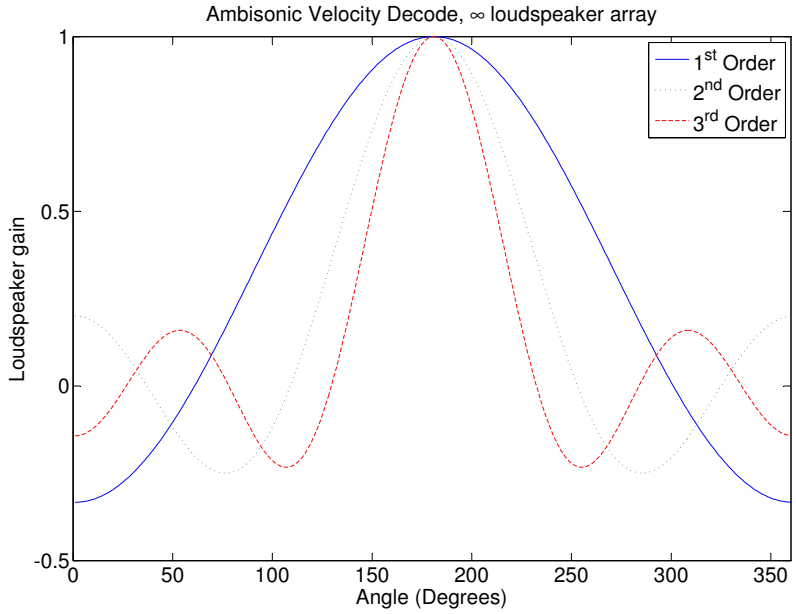


Figure 3.34: *Gains for infinite loudspeaker array for 1^{st}, 2^{nd} and 3^{rd} Order Ambisonic 'velocity' decode.*

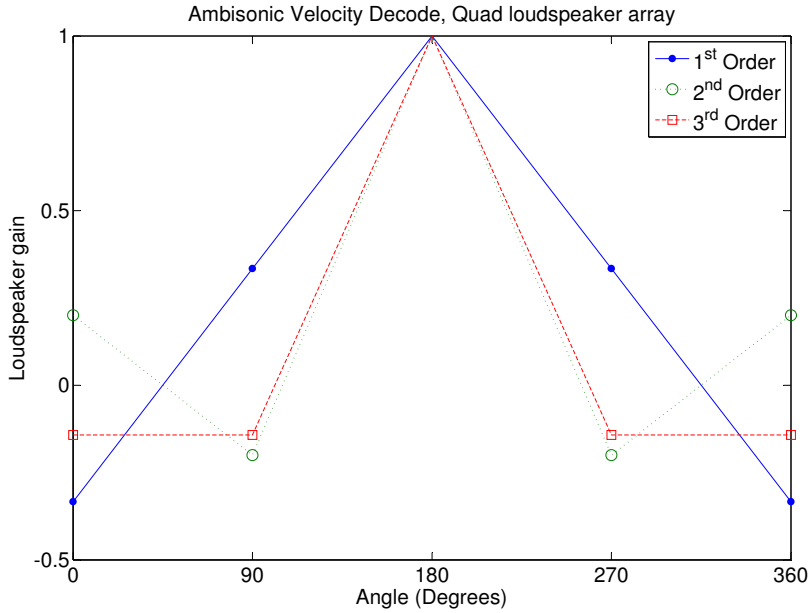


Figure 3.35: *Loudspeaker gains for 1^{st}, 2^{nd} and 3^{rd} Order Ambisonic 'velocity' decode in a Quad system.*

Now we can consider the case where we have an under-sampled system, i.e. the number of loudspeakers is insufficient to represent the decoder order. Figure 3.35 illustrates this case for a quadraphonic setup. We see that whilst the first order decode is represented well, the second and third order decodes are not represented sufficiently. The positive side lobes of the third order decode are completely gone and all other loudspeakers bar the loudspeaker at 180° exhibit

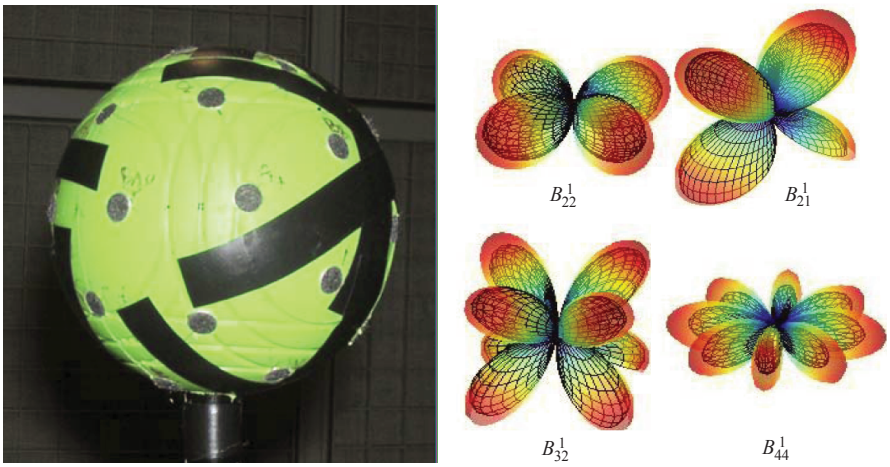


Figure 3.36: *4^{th} Order Spherical Ambisonic microphone and example synthesized spherical harmonics [Moreau et al., 2006].*

negative gains. The result is that the image pulls towards one loudspeaker when the source is in that general direction. Ideally, the minimum number of loudspeakers for the second order decode is 6, and 8 for the third order decode.

3.4.5.1 Recording for Higher Order Ambisonics

First order Ambisonics implementations are the most common due to the fact that the directional responses readily available for spherical harmonic representation are zero and first order (i.e. omnidirectional and figure of 8). Moreover, the creation of directional microphone responses for higher orders is not trivial and no commercial microphones exist to satisfy them. Poletti has shown how circular arrays of microphones can be used to capture higher order spherical harmonics using a two-dimensional Fourier transform [Poletti, 2005]. Moreau et al [Moreau et al., 2006] outline the design of a 4^{st} Order Ambisonic microphone, based on placement of 32 sensors in a pentakis-dodecahedron, as shown in Figure 3.36. Here, the pattern of a higher order spherical harmonic can be synthesized up to 4^{th} Order by filtering the microphone outputs with a set of matched FIR filters formed from large datasets of impulse response measurements around the microphone, and performing a numerical least-squares minimization between the filtered signals and their theoretical counterparts.

However, one should also note that any monophonic recording can be encoded into higher order Ambisonics. For example, a multi-mono microphone approach can be taken for recording a musical ensemble and the auditory scene reconstructed in higher order Ambisonics. In such cases, it is important to bear in mind that the number of microphone channels required equals the number of instruments. This is discussed in further detail in Chapter 5.

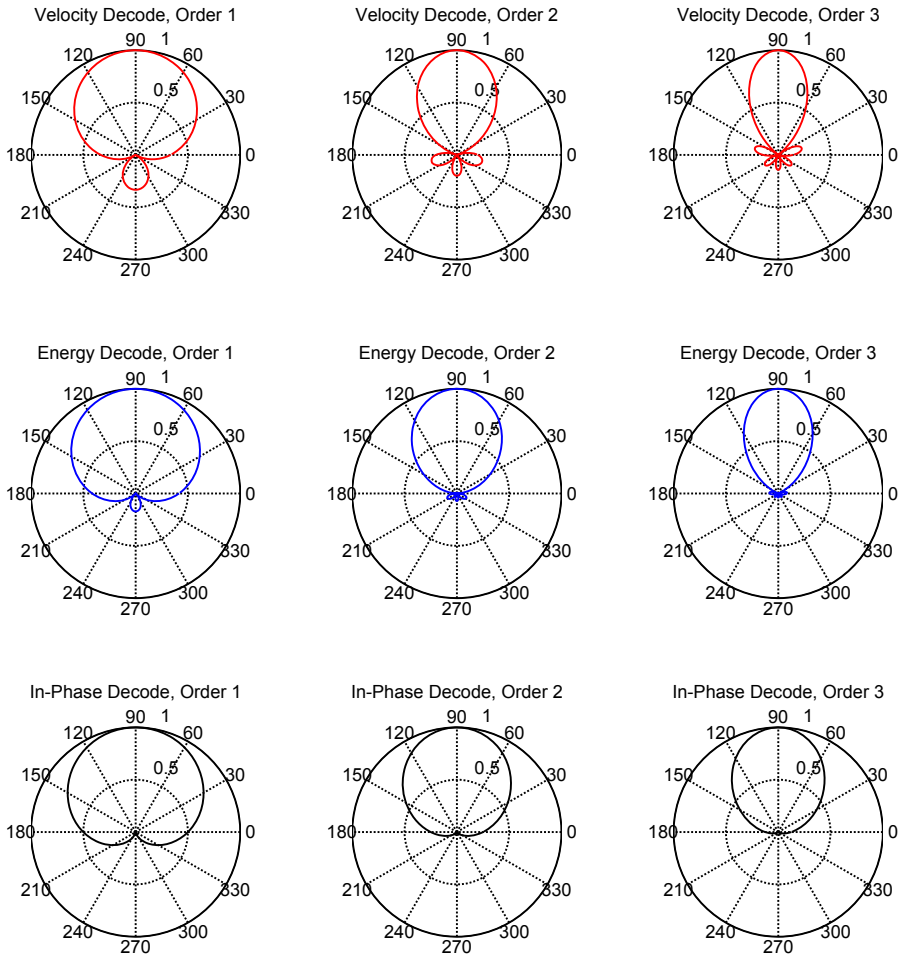


Figure 3.37: *Virtual microphones for* 1^{st}*,* 2^{nd} *and* 3^{rd} *Order Ambisonics decoding.*

3.4.5.2 Decoding for Higher Order Ambisonics

The decoding principles outlined in Section 3.4.2 can easily be expanded to higher order Ambisonics. For horizontal arrays, each element of the matrix describing the loudspeaker array $\mathbf{L}$ can now be rewritten as

$$\mathbf{l}_n = \left[1,...,\sqrt{2}\cos(m\theta_n), \sqrt{2}\sin(m\theta_n)\right]^T \tag{3.62}$$

where θ_n is the azimuth of loudspeaker n and M is the Ambisonic order. The resultant decoders are optimized for low frequency localization, but again, weights can be applied to the B-Format signals such that the virtual microphone patterns can be psychoacoustically optimized.

Just like the 1^{st} Order investigations in Section 3.4.4, there are 4 types of HOA decoder: velocity, energy, in-phase, and 'shelf' decoders. We readily see that as we move up in order, the virtual microphones become narrower, resulting in finer angular directivity. Furthermore, the contribution of the rear lobes is reduced, and we note that a further two rear lobes are

introduced per order. The exception to this is of course, the 'in-phase' decode.

3.4.6 Ambisonics Summary

Ambisonics has been utilized over the last 30 years as an alternative spatialization approach to stereophony. Although its origins can be found in the early stereophonic work of Blumlein, Ambisonics is a soundfield reproduction method and unlike stereophony, uses all loudspeakers to control the soundfield at the centre of the reproduction array. Localization accuracy at the sweet spot using this system is determined by maximizing the velocity vector at low frequencies (<700Hz) and the energy vector at high frequencies (> 700Hz). 1^{st} Order Ambisonics, unlike VBAP, is a complete solution for storage and reproduction of multichannel audio, and using only 4 information channels, full periphonic surround sound using any number of loudspeakers can be accomplished. It is also possible that the sweet-spot can be extended by increasing the order of the spherical harmonics in the soundfield representation, but we note that no commercial microphones exist to accommodate this. Furthermore, practical implementation of HOA depends on the number of reproduction channels available. In terms of storage on commercial high definition media, it would be possible to record up to 8 circular harmonic components yielding a 3^{rd} Order horizontal only system (In fact, only 7 cylindrical harmonics are required for 3^{rd} Order, but 9 are required for 4^{th} Order, making 3^{rd} Order the maximum order Ambisonic reproduction method suitable with currently available high definition storage media). A HOA decoder would then be required on reproduction.

Another important aspect of the Ambisonic system is its smooth panning ability. It is well known that as a source is panned from left to right in a standard two-channel setup, that its timbre changes due to the multiple delayed versions of the source at the ears. In contrast, Ambisonics does not suffer from this coloration to the same extent, since all loudspeakers contribute to formation of the phantom source.

A current limiting factor for the widespread adoption of HOA is the lack of exposure and commercial availability of higher order Ambisonic systems. However, recent work by Craven [Craven, 2003], Wiggins [Wiggins, 2007] and Poletti [Poletti, 2007] has shown how higher order Ambisonic decoding coefficients can be numerically derived for irregular loudspeaker layouts such as the ITU 5.1 layout [ITU, 2006]. Thus, it is conceivable that with the advent of low-cost implementation of HOA microphones, commercial support for HOA will develop significantly within the foreseeable future.

Perhaps one of most favorable benefits of expanding Ambisonics into higher orders is that the sweet spot increases significantly over the listening area. At high orders, the resultant soundfields then approach the performance of wavefield synthesis systems. This comparison will be discussed in detail in Section 3.6. Before this however, we must first investigate the theoretical and practical aspects of Wavefield Synthesis reproduction.

3.5 Wavefield Synthesis

Unlike other spatialization techniques, Wavefield Synthesis (WFS) is a *volume* solution, which attempts to recreate wavefields that would be experienced naturally. The great advantage to WFS is that there is no 'sweet spot', and so it is especially applicable to distributed audience scenarios. In this context, it is generally implemented using line arrays of hundreds of loudspeakers surrounding an audience as shown in Figure 3.38.

First proposed by Berkhout in the 1980's [Berkhout, 1988], WFS is based on the Huygen's principle in optics, which states that the propagation of a sound wave through a medium can be qualitatively described by adding the contributions of all secondary sources positioned along a wave front. This concept is illustrated in Figure 3.39. This implies that the wave field in a source-free volume, V, can be described by a distribution of secondary sources along the boundary surface, S, of volume V [Berkhout, 1998]. Mathematically, this property can be qualitatively described by the Kirchhoff-Helmholtz integral,

$$P(\mathbf{r},\omega) = \frac{1}{4\pi}\int\int_S \left[P(\mathbf{r}_S,\omega)\frac{\partial}{\partial n}\left(\frac{e^{-jk|\mathbf{r}-\mathbf{r}_S|}}{|\mathbf{r}-\mathbf{r}_S|}\right) - \frac{\partial P(\mathbf{r}_S,\omega)}{\partial n}\frac{e^{-jk|\mathbf{r}-\mathbf{r}_S|}}{|\mathbf{r}-\mathbf{r}_S|}\right] dS \tag{3.63}$$

where $P(\mathbf{r},\ \omega)$ is the sound pressure in the Fourier domain, k is the wave number ω/c, 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 expression represents the sound pressure at an arbitrary point within a volume as expressed by the sound pressure $P(\mathbf{r}_S,\omega)$ and its derivative $\partial P(\mathbf{r}_S,\omega)/\partial n$ on the boundary of the volume. In the expression, the first term represents a distribution of dipoles that have a source strength given by the sound pressure of the soundfield at the surface, and the second term represents a distribution of monopoles that have a source strength given by the normal velocity of the soundfield (which is proportional to $\partial P/\partial n$). The only restriction with the Kirchhoff-Helmholtz integral is that there must be no *real* sources within the volume.

If instead of a closed surface we consider a plane surface, the Kirchhoff-Helmholtz equation can be simplified, leading to Rayleigh integrals that describe how the wavefield can be generated from only secondary monopoles or dipoles. For the first case, the Rayleigh I integral describes how the sound field can be reconstructed by a distribution of monopoles in the reproduction plane, if we know the particle velocity in the measurement plane. Following the derivation of Boone [Boone and Verheijen, 1993] for multichannel systems, the Rayleigh I integral is given as

$$\textbf{(Rayleigh I)} \quad P(\mathbf{r},\omega) = \rho c\frac{jk}{2\pi}\int\int_S \left[V_n(\mathbf{r}_S,\omega)\frac{e^{-jk|\mathbf{r}-\mathbf{r}_S|}}{|\mathbf{r}-\mathbf{r}_S|}\right] dxdy \tag{3.64}$$

where $k|\mathbf{r}-\mathbf{r}_s| >> 1$ (i.e. far field approximation). Similarly, the Rayleigh II integral defines the recording of the pressure in the measurement plane for reproduction over dipole loudspeakers.

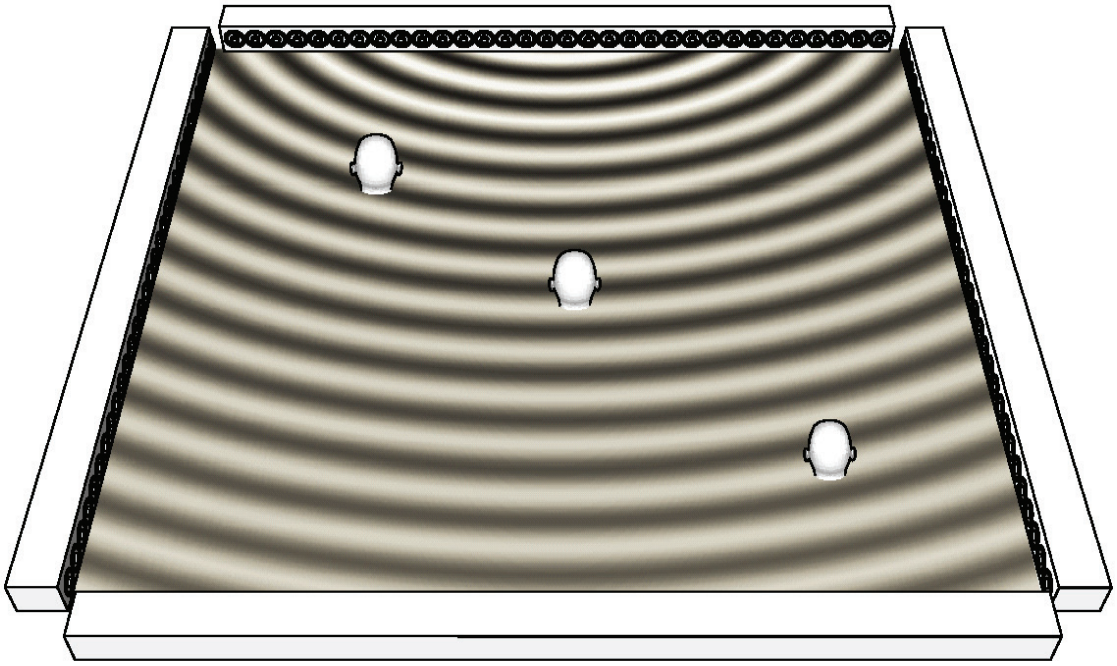


Figure 3.38: *Horizontal only wavefield reproduction.*

It is defined as

$$\textbf{(Rayleigh II)} \quad P(\mathbf{r},\omega) = \frac{jk}{2\pi}\int\int_S \left[P(\mathbf{r}_S,\omega)\frac{1+jk|\mathbf{r}-\mathbf{r}_S|}{jk|\mathbf{r}-\mathbf{r}_S|}\cos\theta\frac{e^{-jk|\mathbf{r}-\mathbf{r}_S|}}{|\mathbf{r}-\mathbf{r}_S|}\right]dxdy \tag{3.65}$$

In practice, because we can only use a finite number of secondary sources, we must use discretized versions of equations 3.64 and 3.65, given as

$$P(\mathbf{r},\omega) = \rho c\frac{jk}{2\pi}\sum_{n=1}^{N}\left[V_n(\mathbf{r_n},\omega)\frac{e^{-jk|\mathbf{r}-\mathbf{r}_n|}}{|\mathbf{r}-\mathbf{r}_n|}\right]\Delta x\Delta y \tag{3.66}$$

and

$$P(\mathbf{r},\omega) = \frac{jk}{2\pi}\sum_{n=1}^{N}\left[P(\mathbf{r}_n,\omega)\frac{1+jk|\mathbf{r}-\mathbf{r}_n|}{jk|\mathbf{r}-\mathbf{r}_n|}\cos\theta_n\frac{e^{-jk|\mathbf{r}-\mathbf{r}_n|}}{|\mathbf{r}-\mathbf{r}_n|}\right]\Delta x\Delta y \tag{3.67}$$

where $P(\mathbf{r}_n)$ and $V(\mathbf{r}_n)$ are the pressure and velocity components of the n^{th} secondary source, and N is the total number of secondary sources. A further simplification is if the reconstruction is restricted to the horizontal plane. The plane of secondary loudspeakers then reduces to an infinite line of secondary loudspeakers, which simplifies things significantly. This simplification is justified by the fact that many common sources such as speech or acoustic musical instruments are projected in the horizontal plane. Furthermore lateral reflections are sufficient for creating a sense of source width and spaciousness. In this case, the wavefield that was propagating spherically, now becomes a cylindrical wavefield. A consequence of this is that the reconstructed soundfield will not be correct for all listener positions.

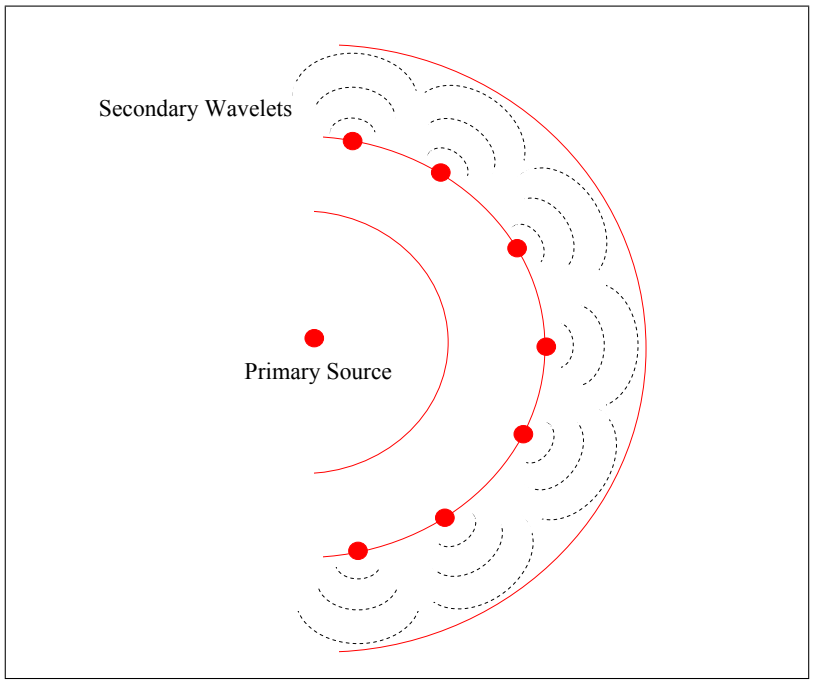


Figure 3.39: *Huygen's Principle.*

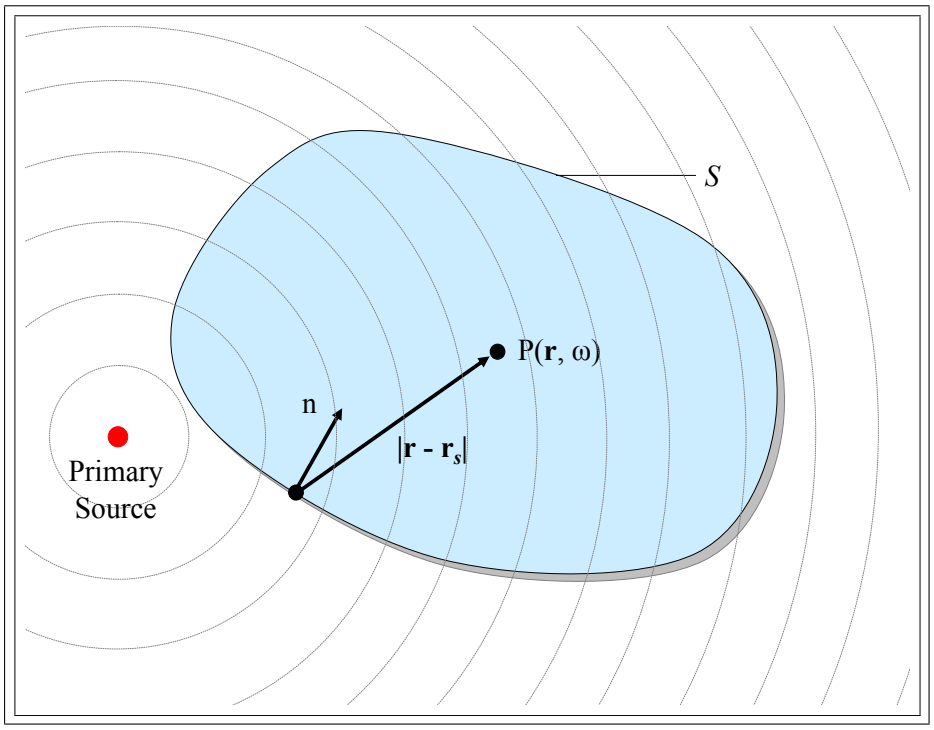


Figure 3.40: *Geometry of the Kirchhoff Helmholtz integral .*

Boone [Boone and Verheijen, 1993] has outlined a methodology for practical implementation of WFS using multichannel systems. It more typical that the secondary sources are monopole, since it is easier to construct the desired radiation patterns, and so the Rayleigh I integral is

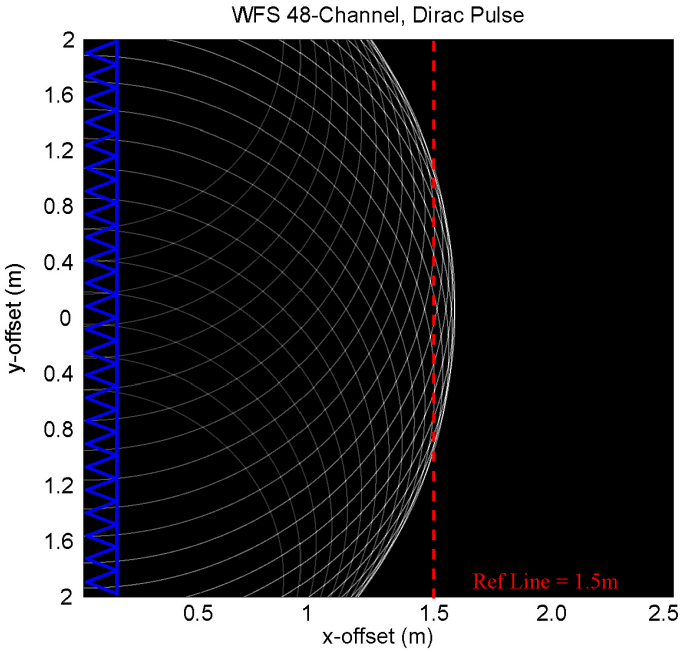


Figure 3.41: *Example of secondary source convergence at reference listening line.*

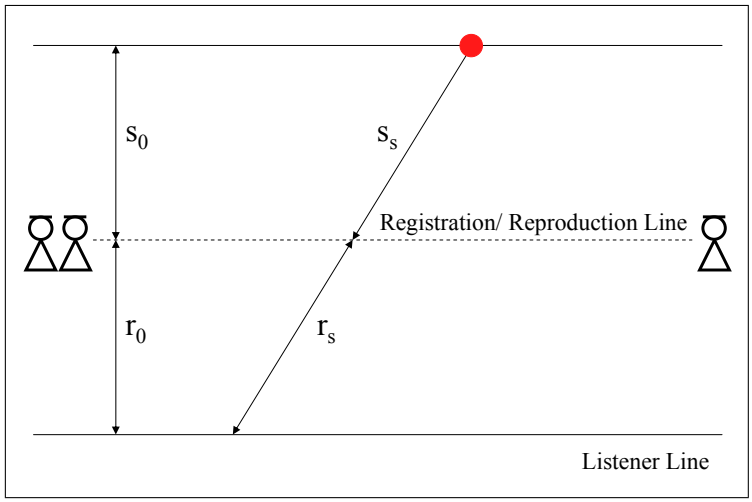


Figure 3.42: *Geometry between source position, listener line and registration/reproduction line.*

---

most commonly used. For this 1-D line array case, equation 3.66 then becomes,

$$P(\mathbf{r},\omega) = \sum_{n=1}^{N} G_n(\omega) V_n(\mathbf{r}_n,\omega) \frac{e^{-jk|\mathbf{r}-\mathbf{r}_n|}}{|\mathbf{r}-\mathbf{r}_n|} \tag{3.68}$$

where $G_n(\omega)$ is a Green's function, such that the sound pressure P due to the monopole secondary sources is equivalent to that of a primary source (created at the virtual source position)

at a listening line reference $\mathbf{r}_m$ so that

$$P(\mathbf{r},\omega) = S_m(\omega)\frac{e^{-jk|\mathbf{r}-\mathbf{r}_m|}}{|\mathbf{r}-\mathbf{r}_m|} \tag{3.69}$$

where $S_m(\omega)$ is the pressure due to a monopole source. Figure 3.41 shows the convergence of the sound pressures due to a line array of secondary sources to form the wavefield of a single source at the reference line.

If we use the far field assumption, then $G_n(\omega)$ can be considered independent of n yielding

$$G_n(\omega) = Z_0 B_v \sqrt{jk} \tag{3.70}$$

where

$$B_v = \sqrt{\frac{\mathbf{s}_s\mathbf{r}_0}{2\pi(\mathbf{s}_0+\mathbf{r}_0)}} \tag{3.71}$$

and $Z_0 = \rho_0 c$ is the impedance of air and ρ_0 and c are the mean density and the speed of sound respectively. The geometry for this setup is shown in Figure 3.42.

The driving signals for the loudspeakers are

$$Q_n(\omega) = \frac{4\pi}{\rho K_m} Z_0 B_v \sqrt{jk} V_n(\mathbf{r}_n,\omega)\Delta x \tag{3.72}$$

The $\sqrt{jk}$ term in this expression equates to a +3dB per octave boost of the driving signals.

3.5.1 Reproduction using Discrete Linear Arrays

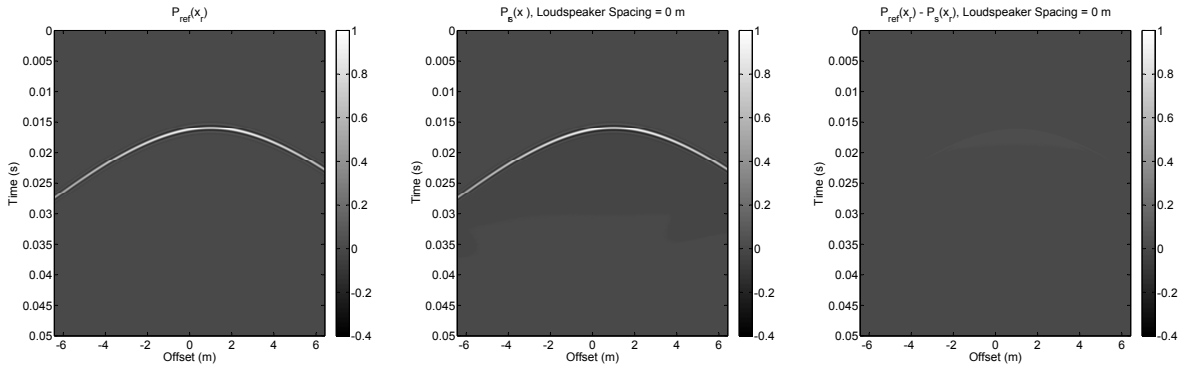
Before, we further discuss the implications of using line arrays with a discrete number of elements, let us consider the best case scenario for line array reproduction. Figure 3.43 shows the example case for reproduction of a band-limited ($<$ 1700Hz) impulse using a *continuous* loudspeaker array infinitely long. We can readily see that there is little difference between the reference pulse and the reproduced wavefield. However, some discrepancies can be seen if we look at the resultant wavefield in the Fourier domain. First, we will consider the spectral error E of the reproduced wavefield, given from [Start, 1997] as

$$E(\mathbf{r},\omega) = P_{pr}(\mathbf{r},\omega) - P_{sec}(\mathbf{r},\omega) \tag{3.73}$$

where $P_{pr}(\mathbf{r},\omega)$ and $P_{sec}(\mathbf{r},\omega)$ are the primary and secondary pressures at a registration point. Since both $P_{pr}(\mathbf{r},\omega)$ and $P_{sec}(\mathbf{r},\omega)$ are complex spectra, it is convenient to look at the spectral error in terms of a magnitude error L_E, given by

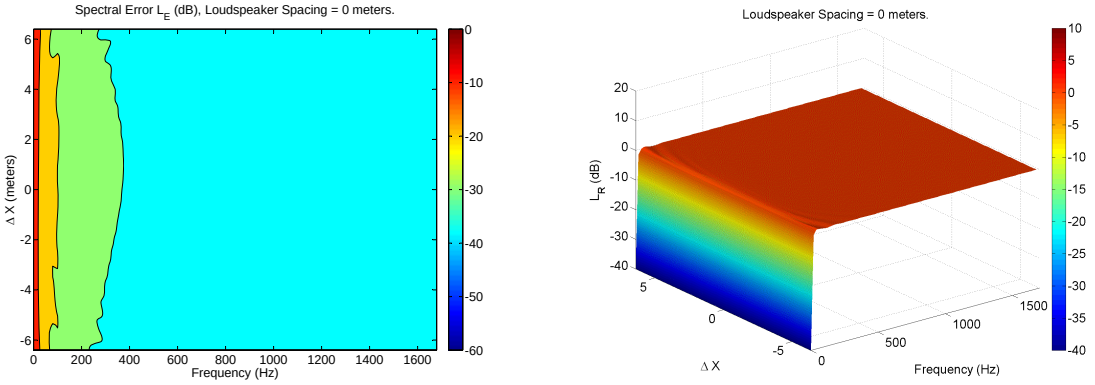
$$L_E(\mathbf{r},\omega) = 10\log\left(\frac{|E(\mathbf{r},\omega)|^2}{|P_{pr}(\mathbf{r},\omega)|^2}\right) \qquad \text{dB} \tag{3.74}$$

where $|P_{pr}(\mathbf{r},\omega)| \neq 0$. This shows us the relation between the sound pressure level of the error wavefield to that of the primary wavefield, and is shown for the continuous infinite array case in



(a) (b) (c)

Figure 3.43: *Wavefield reconstruction of band-limited delta function with reference to a true source. (a) Source wavefield, (b) WFS wavefield, (c) Difference wavefield.*



(a) (b)

Figure 3.44: *Error analysis for infinite loudspeaker line array (a) Spectral Error (b) Pressure ratio.*

Figure 3.44(a). In the ideal case $L_E = -\infty$. However, here we see that errors are exhibited for low frequencies due to far field approximations, although above approximately 100Hz, the error level is quite low, remaining below 20dB.

A further error measure, defined by Start [Start, 1997] is that of the synthesized to primary pressure ratio R_p given by

$$R_p(\mathbf{r}, \omega) = \frac{P_{sec}(\mathbf{r}, \omega)}{P_{pr}(\mathbf{r}, \omega)} \tag{3.75}$$

where $|P_{pr}(\mathbf{r}, \omega)| \neq 0$. $R_p(\mathbf{r}, \omega)$ gives the complex ratio between the reproduced wavefield and the primary wavefield. If we are only interested in the amplitude of the pressure ratio then we can use

$$L_R(\mathbf{r}, \omega) = 10 \log \left(R_p(\mathbf{r}, \omega) \right) \tag{3.76}$$

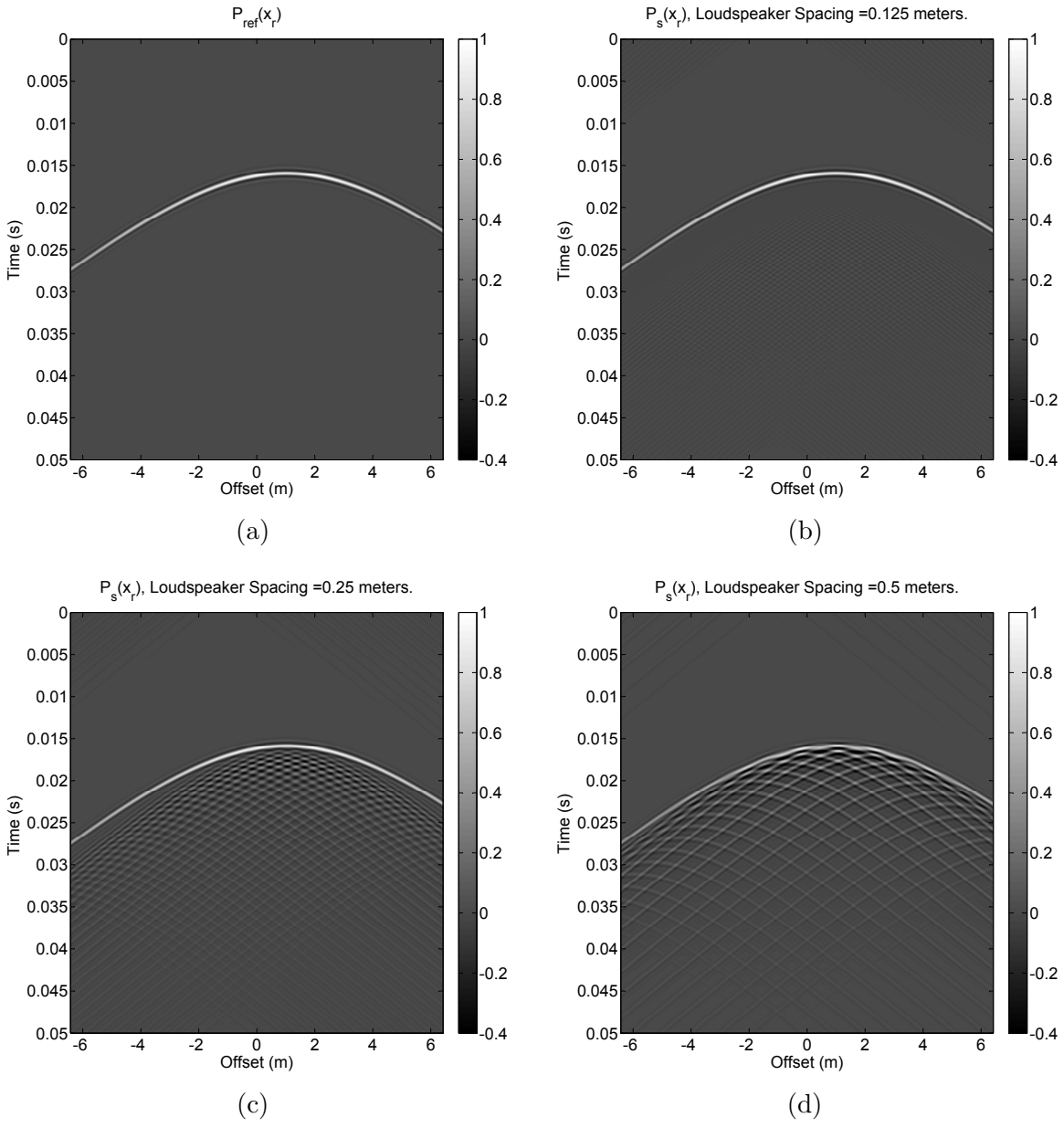


Figure 3.45: *Reproduced wavefields for band-limited impulse at* 10° *for infinite line arrays consisting of (a)* $\Delta x = 0$*, (b)* $\Delta x = 0.125$*, (c)* $\Delta x = 0.25$ *and (d)* $\Delta x = 0.5$ *loudspeaker spacing.*

This gives the level sound pressure error of the reconstructed wavefield relative to the primary wavefield. In the ideal case, $L_R = 0$, meaning that the amplitude spectrum of the reproduced wavefield matches the primary wavefield exactly. This is shown to be the case in Fig 3.44(b) for the continuous infinite linear array.

Given this best case analysis, we can now look at reproduction using finite element arrays. Let us consider the case where we again have an infinitely long array, but with the loudspeaker

elements spatially separated by Δx. Aliasing will not occur as long as the spatial bandwidth does not exceed

$$f_{max} \leq f_{al} = \frac{c}{2 * \Delta x \sin(\theta_{max})} \tag{3.77}$$

where f_{al} is the spatial aliasing frequency, dependent on the maximum angle θ_{max}. θ_{max} is defined by the difference between the angle of the virtual source θ_v and the maximum listening angle on the reproduction side θ_{sec}, yielding for equation 3.77

$$f_{max} \leq f_{al} = \frac{c}{\Delta x |\sin(\theta_{sec}) - \sin(\theta_v)|)} \tag{3.78}$$

The effect of increasing the distance between array elements is shown in the wavefields of Figure 3.45. Here an impulse of 1.7kHz bandwidth is reproduced over arrays with loudspeaker spacings 0, 0.125, 0.25 and 0.5m spacing. The impulse angle is at 10° and the registration of the wavefield is over -6.4m to 6.4m. We can readily see the aliasing artifacts appear, which are particularly prominent in the array with 0.5m spacing. This will cause significant localization errors in the reproduced wavefield.

Consequently, the spectral errors in the under-sampled wavefields are also significant. Figure 3.46 shows the L_E for each case. For a spacing of 0.125m we see that the error is acceptable up to approximately 1400Hz. At this spacing, the ITD cues would be expected to be preserved correctly. Above this frequency range, the error immediately becomes unacceptable.

As we increase the loudspeaker spacing, we directly see the effect predicted with equation 3.78. For the 0.25m loudspeaker spacing f_{al} becomes approximately 700Hz. At this spacing, ITD cues can still be preserved for frontal sources, but if the source/listener angle changes significantly, such cues will be destroyed. For 0.5m spacing the error is completely unacceptable and wavefield synthesis for off-centre listening cannot be accurately achieved. The pressure ratio error, shown in Figure 3.47 tells a similar story for each case, with marginally more optimistic results for the 0.5m spacing line array. Here the magnitude spectrum is still preserved well under approximately 300Hz. However, this is inconsequential for source localization since no interaural level differences are expected in this frequency range, and the complex spectrum is not well preserved. Also, only the timbre of extremely low frequency sources can be preserved.

There has been several attempts made to overcome the effects of spatial aliasing. Both Start and De Vries et al suggest that the minimum source angle should be restricted at higher frequencies, ensuring a high value of f_{al} [Start, 1997], [de Vries et al., 1994]. However, this results in a loss of spatial information. Start has also suggested a 'time-domain randomization' approach, where the driving signals are subject to small randomized time delays, resulting in a diffuse spectrum above the spatial aliasing frequency. This method has serious trade offs, since any correct wavefronts contained above the minimum spatial aliasing frequency would be destroyed. Corteel et al also employ diffusion filters, and demonstrate that the reduced comb-filtering effects lead to a reduction in coloration of the sound [Corteel et al., 2007]. One promising

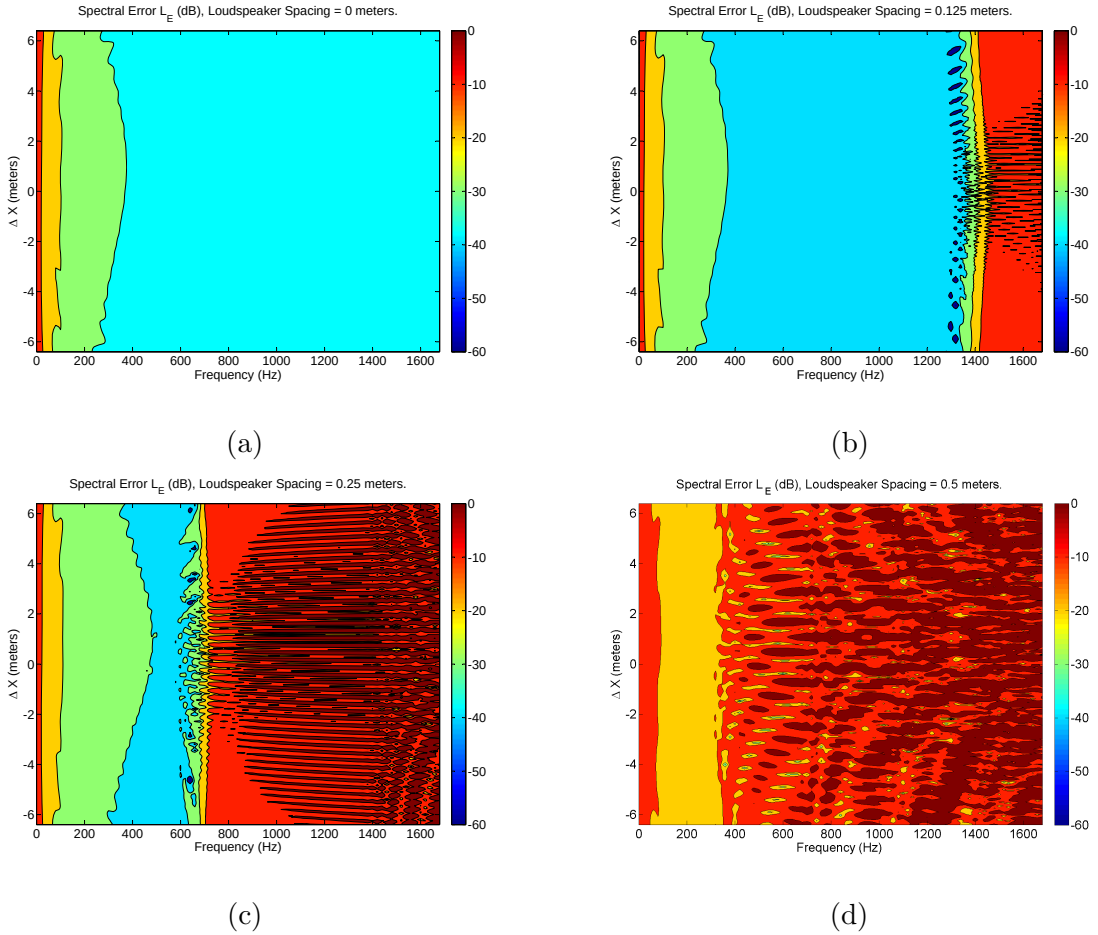


Figure 3.46: *Spectral error in wavefields for 8m line arrays consisting of (a) infinite, (b) 256, (c) 64 and (d) 32 loudspeakers.*

method for reduction of spatial aliasing has recently been proposed by Wittek [Wittek, 2007]. Here a hybrid WFS/Stereo system is proposed, employing Wavefield Reproduction below the spatial aliasing frequency and VBAP above it. The method, termed 'Optimized Phantom Source Imaging' (OPSI) has been shown to perceptually reduce coloration effects and Wittek outlines the bounds of localization accuracy across the listener area using the method.

There is also a further consideration for finite loudspeaker reproduction in WFS. Whenever we have a finite length array diffraction effects occur and phantom images are formed originating from the sides of the array. This is demonstrated in the example shown in Figure 3.48. Here the 32 channel line array is truncated to 3m in length. The diffraction effects are clearly shown and result in amplitude errors in the primary wavefront. In the frequency domain, this interference manifests as ripples in the pressure ratio, leading to timbral change. Even in the region where the primary wavelet is not so affected (approx -2.5m to 2m), the spectral error L_E is high, as shown in Figure 3.49 (b). An effective method to reduce dispersion effects due to truncation is to apply a tapering window to the driving signals in the array. By reducing the overall gain

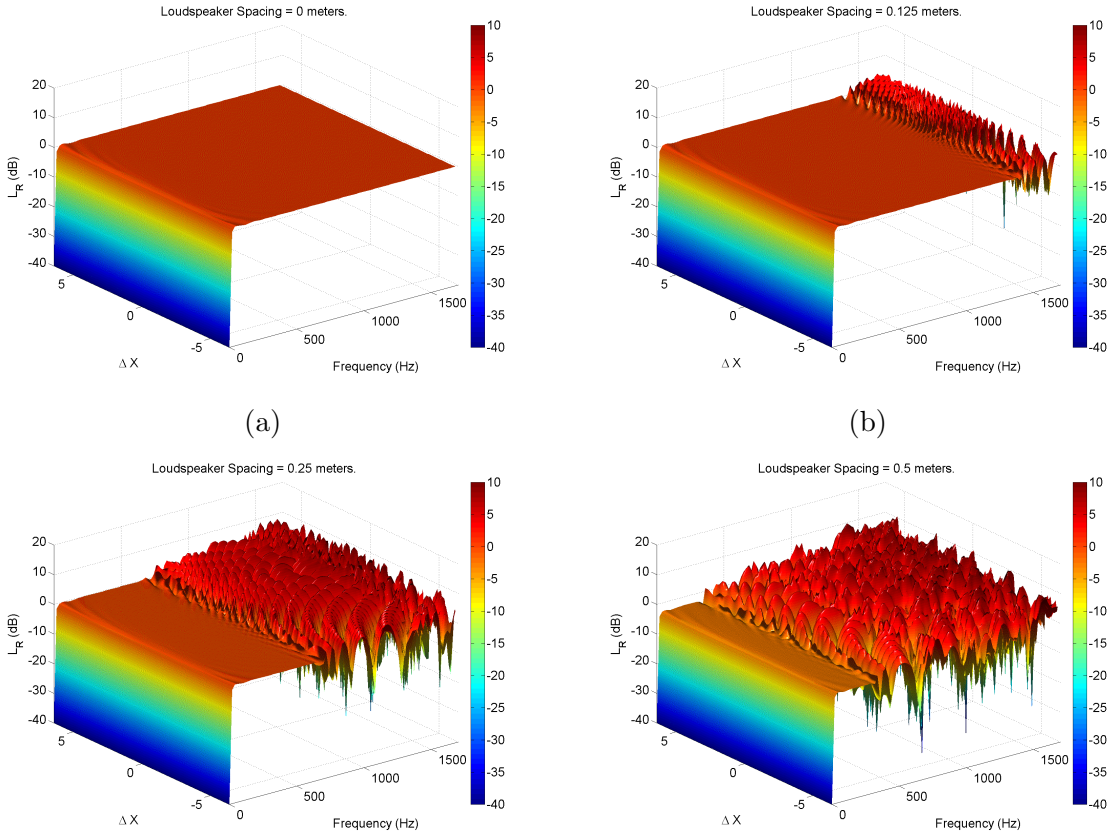


Figure 3.47: *Pressure ratio in wavefields for 8m line arrays consisting of (a) infinite, (b) 256, (c) 64 and (d) 32 loudspeakers.*

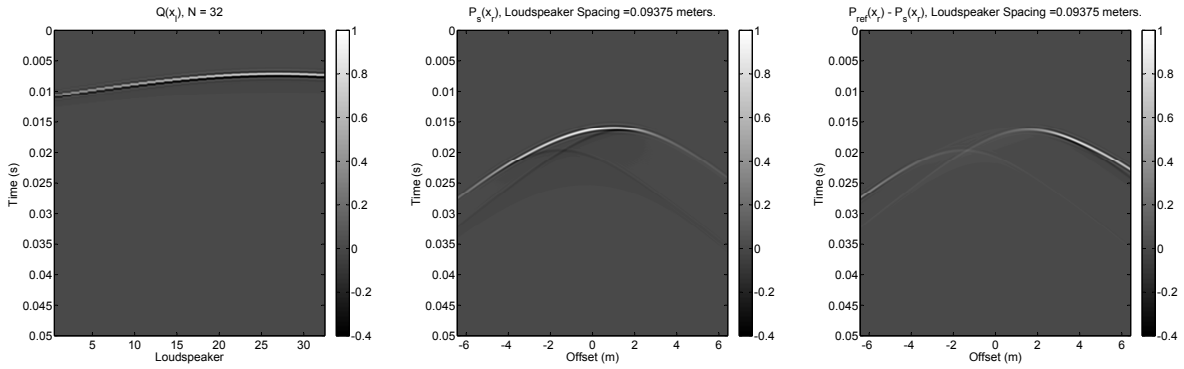
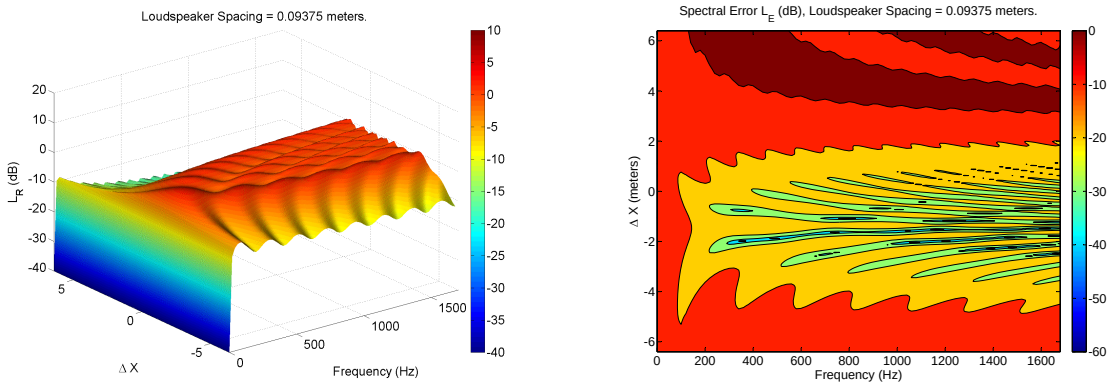


Figure 3.48: *Wavefield Synthesis exhibiting diffraction effects (a) Driving Signals, (b) Reproduced Wavefield, (c) Difference between reference and reproduced wavefield.*



(a) (b)

Figure 3.49: *Error analysis for finite loudspeaker line array (3m) (a) Pressure Ratio (b) Spectral Error.*

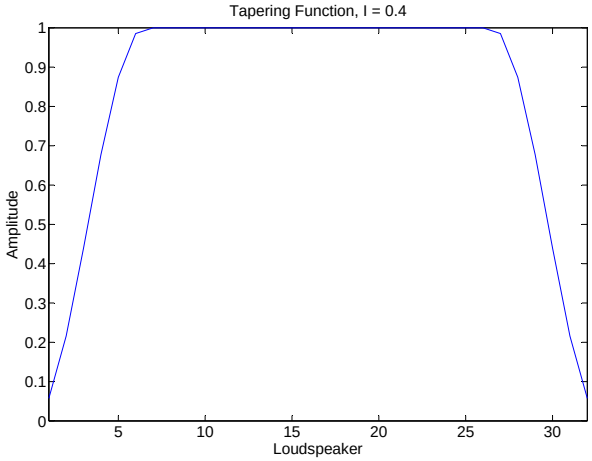
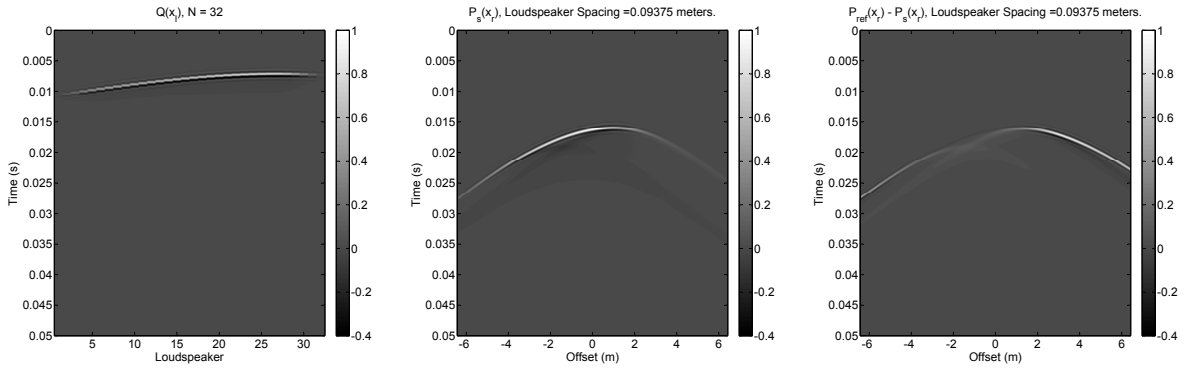


Figure 3.50: *Amplitude tapering function for 32 channel wavefield driving signals.*

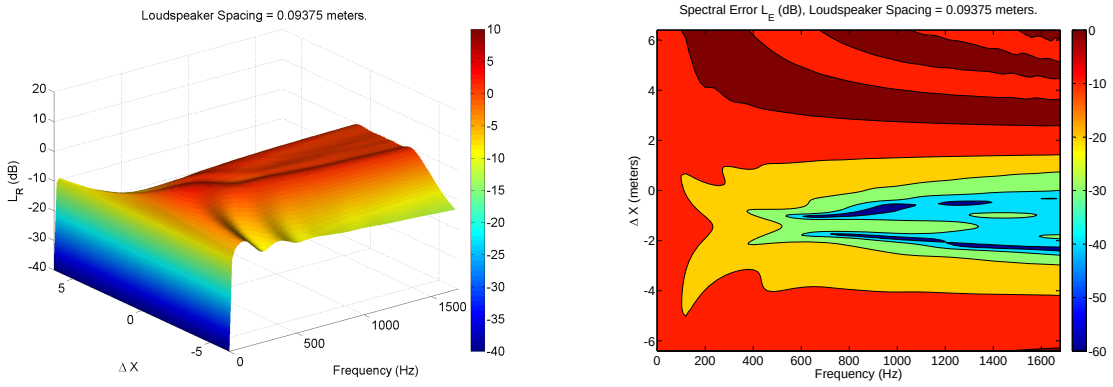
of the signals at the end of the array, the diffraction effects can be reduced significantly. An example of such a tapering function which employs a half-hamming window on the start and end 20% of the array is shown in Figure 3.50.

The tapered driving signals are shown in Figure 3.51(a). The resultant wavefield now has reduced diffraction artifacts and the primary wavelet is still correct. The synthesis error is also substantially decreased as shown in 3.51, and the pressure ratio now has reduced rippling effects. Tapering is a very effective method for reduction of such artifacts, and the amount of tapering employed is a trade-off between the size of reproduction area/aperture and the diffraction effects. A mathematical description of these artifacts can be found in [Start, 1997]. However, it is important to note that it is not possible to completely avoid truncation artifacts for multiple sources at different incident angles and different listener positions in line-array reproduction.



(a) (b) (c)

Figure 3.51: *Wavefield Synthesis with reduced diffraction effects (a) Driving Signals, (b) Reproduced Wavefield, (c) Difference between reference and reproduced wavefield.*



(a) (b)

Figure 3.52: *Error analysis for finite loudspeaker line array with cosine taper (3m) (a) Pressure Ratio (b) Spectral Error.*

3.5.2 Recording For Wavefield Synthesis

There are two main methods for recording for wavefield synthesis. The first method involves the recording of acoustic events with an array of pressure/velocity receivers for reproduction over dipole/monopole arrays. This is a true wavefield recording method but it is costly, since it requires a large number of microphones. The second method involves the close microphone recording of sound sources and then their subsequent convolution with measured room impulse responses. However, in order to capture the correct spatial resolution, a large amount of impulse measurements are required. Hulsebos et. al have critiqued several microphone array configurations for this purpose [Hulsebos et al., 2001]. They investigate the use of linear, cross, and circular arrays for wavefield capture.

In wavefield reproduction, if the reproduction array dimensions does not match the microphone array geometry then wavefield extrapolation procedures must be employed. Hulsebos et. al show that in the line array case, hypercardioid microphones allow for discrimination between forward and inverse wavefields for the purpose of wavefield extrapolation. However, the reconstruction area is limited by the aperture of the array, due to microphone directivity and array configuration, and so linear arrays are not deemed suitable for wavefield extrapolation. Cross hypercardioid arrays, which consist of two linear hypercardioid arrays show a better performance, but the reconstruction is still not perfect. Circular arrays, using both the pressure and velocity components (and thus the Kirchhoff-Helmholtz integral) give the best performance. No diffraction effects are exhibited, but since the Kirchhoff Helmholtz integral is used the wavefield is only reproduced within the circle, and outside the pressure field is zero. The array does not have to be a circle, but can be any shape as long as it is closed, although the circular form simplifies the equations. The corresponding plane wave decomposition then involves decomposing this wavefield into cylindrical harmonic components, where the monopole and dipole sources can be viewed as first and second order cylindrical harmonics. In this way, we begin to identify similarities between WFS and HOA. These will be addressed in Section 3.6.

3.5.3 Wave Field Synthesis Summary

Wave Field Synthesis is a spatialization solution that aims to deliver the wavefronts necessary for correct localization across a listening area. The theory is based on an infinite array of secondary sources, but in reality a finite array of monopole or dipole radiators is required. The system has several limitations, the main one being coloration and localization blur due to spatial aliasing. OPSI is a method which can be used to reduce this coloration, by employing stereophonic techniques above the spatial aliasing frequency. For linear arrays, dispersion effects can be reduced by employing tapering to the array signals. However, such effects are not seen in circular arrays. Recording for wavefield synthesis is best accomplished using circular microphone arrays, but the storage and reproduction of pre-rendered/recorded wave field synthesis signals is non-trivial. Recent developments within the framework of the European CARROUSO[4] Project have attempted to overcome such limitations by merging WFS technology with the MPEG-4 standard for teleconferencing applications. Here, mono source signals and their directional information are transmitted via digital video broadcasting (DVB) streams. Reproduction is accomplished by convolving the monophonic audio with a data-set of measured or modeled impulse responses on playback.

[4]CARROUSO "Creating Assessing and Rendering in Real Time of High Quality Audio Visual Environments in MPEG-4 Context"

3.6 Wave Field Synthesis and Higher Order Ambisonics

It is interesting to note that although WFS and HOA stem from different theoretical backgrounds (the Kirchhoff-Helmholtz Integral for WFS and spherical harmonic expansion for HOA), under certain conditions, they can be seen as equivalent. It has been shown by Daniel [Daniel, 2003b] that for the case of circular arrays, a special case of the cylindrical harmonic decomposition of the soundfield meets similar limitations as a reproduced wavefield. The key element is the introduction of near field compensation filters which cater for the fact that since the loudspeakers are at a finite radius, there is a low frequency boost due to near-field spherical wave propagation. This is shown in Figure 3.53 for up to 8^{th} Order Ambisonics. Consequently near-field compensation filters are required to cater not only for sources outside the array, but also focused sources inside the array. These distance coding filters have a finite low frequency amplification, and example compensation filters for sources at 1.5m and 10m in a 3m radius array are shown in Figure 3.54. The result for outside sources is that correct plane wave approximation is achieved. This is shown in the simulation in Figure 3.55. We clearly see in Figure 3.55(b) that the near field compensation results in a closer approximation to a plane wave across the listening area. Furthermore, Daniel also shows how HOA is a more robust soundfield strategy for simulated virtual environments, since it is not affected at higher frequencies by spatial aliasing at the centre position. As well as this, off centre positions retain consistent spatial information. However, in recording scenarios, both systems are found to meet similar limitations and compromises due to microphone spacing and spatial aliasing. For a WFS array of a given radius, a reduction in the number of transducer elements will result in a lower spatial aliasing frequency. For HOA, a reduction in Ambisonic order will most likely result in localization blur and a reduced sweet-spot. However, the use of psychoacoustically optimized decoders can help to achieve greater localization accuracy. Thus, Ambisonic systems are more advantageous since they are completely scalable to the loudspeaker configuration, and accommodate large audiences and individual listeners in a reasonable manner.

3.7 Conclusions

In this chapter, an in-depth comparison of the spatialization techniques applicable to the spatial audio reproduction classes was undertaken. The analysis was implemented by confirming previous work in spatialization techniques through extensive review and simulations. Two main classes of spatial reproduction were outlined: binaural and loudspeaker reproduction. The binaural virtual loudspeaker approach was introduced as a potential method that will allow for the same spatialization technique to be utilized for headphone and binaural reproduction. To this end, three potential methods for spatialization of audio in a VAE were presented: VBAP,

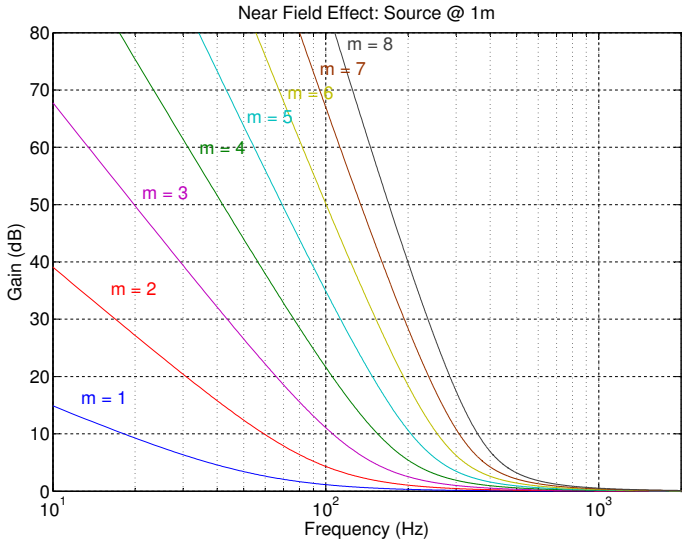
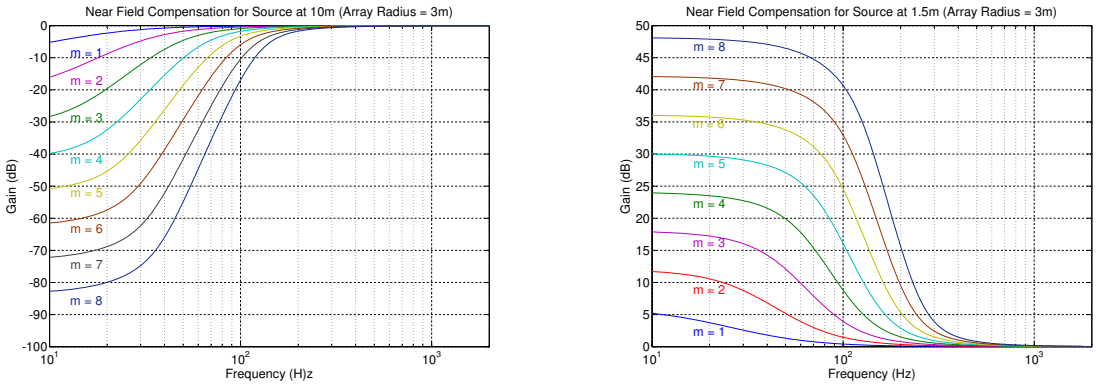
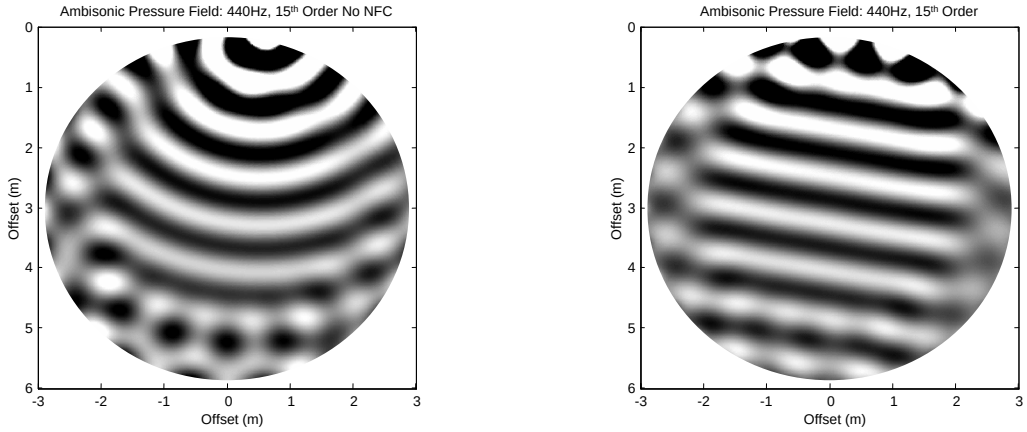


Figure 3.53: *Low frequency boost due to near-field effect.*



(a) (b)

Figure 3.54: *Near field compensation for source (a) outside array and (b) inside array.*



(a) (b)

Figure 3.55: 15^{th} *Order Ambisonic Reproduction (a) without near-field compensation and (b) with near-field compensation.*

Ambisonics and Wave Field Synthesis. Of the systems presented Ambisonics represents the most practical solution, since the information channel count can be lower than the reproduction channel count and that decoding over different loudspeaker layouts is possible. Conversely VBAP and Wave Field Synthesis offer the potential for high localization accuracy over an audience area. The higher order spherical harmonic decomposition of the soundfield was also introduced. It was found to be a promising spatial audio rendering method for use with VAEs, since as the order increases, the area of correct soundfield reconstruction becomes greater. At high orders Ambisonic systems begin to meet the limitations of Wave field Synthesis, and an analysis of these limitations was presented. For circular or line array WFS systems, spatial aliasing is the major limiting factor, since the loudspeaker spacing determines the upper frequency limit at which the wavefields can be correctly reconstructed. For line arrays, diffraction effects must also be considered, and tapering of the array signals should be introduced. For HOA reproduction, near-field compensation filters are also required, which optimize the low frequency contributions of loudspeakers at a finite distance from the array.

For practical application to VAEs, wavefield synthesis still remains logistically difficult to implement. However, it is extremely useful in understanding the fundamental limitations of spatial audio systems, and in particular HOA. For this reason, WFS will be used as an analysis tool later in this thesis. HOA on the other hand is a scalable, practical method. However, while the localization performance of Ambisonic systems has been evaluated in a number of experiments [Jot et al., 1999], [Benjamin et al., 2006], there is a distinct lack of experimental data on its performance under non-anechoic conditions and for a distributed audience. Higher order Ambisonic systems will theoretically recreate the soundfield over a wider listening area but again, additional testing will be required to verify this claim. Such an analysis is undertaken in the following Chapter.

4

A Comparative Study of Spatialization Techniques Under Real Listening Conditions

4.1 Introduction

Over the last thirty years a significant body of work has been conducted into the theoretical evaluation of Ambisonic and stereophonic systems, most notably in [Gerzon, 1980], [Bamford and Vanderkooy, 1995], [Daniel et al., 1998] and [Poletti, 2007]. However, much of the existing research into the *practical* performance of such systems has been carried out under ideal listening conditions for a single listener, for example in [Daniel et al., 1998], [Guastavino et al., 2007], [Wiggins, 2004], [Pulkki, 2001a], [Pulkki, 2001b] to name but a few. Whilst this approach is suitable for assessing the optimal localization performance of a particular system, it does not define the capability of such systems in real listening scenarios. For example, it has been shown in Chapter 2 how the existence of early reflections can significantly affect localization accuracy and perceived source width of monophonic sources. It therefore follows that room acoustics will also have a significant effect on the formation of phantom sources. A further consideration is that even sound pressure levels are generally required over large audience areas and as a consequence, reproduction layouts may be dictated not only by the spatialization technique but also by sound reinforcement considerations. To this end, we will now investigate and critically compare VBAP and 1st Order Ambisonics for off-centre listeners under real listening conditions.

4.2 VBAP and 1st Order Ambisonics in Real Listening Conditions

There are numerous challenges associated with the localization of sources for distributed audiences. In particular, for systems based on conventional stereophonic or Ambisonic principles, it is extremely difficult to present accurate wavefronts for correct localization cues at off-centre listening positions. Even though these systems have been used extensively in theater and music concerts, there has been very little published on their actual performance under such conditions.

In [Pulkki, 2001b] and [Pulkki, 2001c], the localization of amplitude panned phantom sources using VBAP was investigated. In this work, subjects adjusted the direction of a virtual sound source to match that of a real reference source. The test stimuli consisted of broadband pink noise, filtered (one-third octave and two-third octave) pink noise and impulse trains. It was found that the directions implied from the ITD and ILD are not only frequency dependent, but that they deviate from one another considerably at most frequencies. Furthermore, whilst the ILD cues at high frequencies indicate similar directions as the ITD cues at low frequencies, differences exist amongst individuals. In the frequency range 1100 to 2600 Hz, the ITD and ILD cues do not coincide, although this only has an effect on the presentation of narrow-band signals in this region. Moreover, the coincidence of high frequency ILD cues and low frequency ITD cues implies that the sine and tangent panning laws that were derived for low frequencies also hold for high frequencies. However, virtual source spreading can occur due to localization cue deviations between 1kHz and 2kHz.

The localization performance of Ambisonic systems has been evaluated in a number of experiments under different listening conditions. Benjamin et al. [Benjamin et al., 2006] carried out a series of listening tests to investigate the different Ambisonic decoder types. The tests compared a number of different speaker arrays and decoder designs, with the main variables being the number and arrangement of the loudspeakers, and the psychoacoustic parameters of the decoder. The tests were designed to evaluate these models, and the choice of crossover frequency. The four decoders introduced in Section 3.4.4 were investigated over square, rectangular and hexagonal arrays. The source signals used in the test consisted of continuous bandpass filtered noise, voice recordings, various music recordings, applause and fireworks. The listeners were free to switch between the arrays and sources, move their heads and were asked to judge a number of attributes such as the directional accuracy of localization, tonal balance and image stability. The results of the tests indicated that the hexagonal and octagonal arrays were preferred by all listeners. The rectangular and square arrays were judged to exhibit poor lateral imaging, although the rectangular array was comparable to the hexagonal array when the material was limited to a frontal source with ambience. Of the four decoder types tested, the shelf filter decoder was preferred as it was reported to produce the most focused sources with the least artifacts. The energy decoder also produced good results. One interesting conclusion drawn from this test was that changes in layout make significantly more difference than changes in

decoder. It is also noted that the choice of preferred decoder was strongly dependent on the program material and the size of the intended listening area. Benjamin et al. also investigated the use of Ambisonic decoders on an ITU 5.1 array, and found that this layout is sub-optimal in comparison to regular layouts, even when the velocity and energy vectors are maximized for optimal localization accuracy [Benjamin et al., 2006]. Finally, it is be noted that these tests were conducted in an ordinary room without any acoustic treatment. However, no experimental data, acoustical analysis or statistical analysis was presented to support the listening tests.

The tests of Benjamin et al. suggest that for optimal localization to occur for a centrally seated listener, the velocity and energy vector responses should be maximized. However, this can cause significant localization issues for off-centre listeners, since the precedence effect can cause localization in the direction of anti-phase components from diametrically opposed loudspeakers. Malham has worked on a number of large-area Ambisonics systems and has published one of the few papers on this topic [Malham, 1992]. The paper informally covers the experiences of the author in implementing large scale Ambisonics systems in a number of different theaters. The main conclusions of the paper are as follows:

- Informal tests demonstrated that Ambisonics worked effectively with a hexagonal array of diameter 14.5m.
- Localization inaccuracies were experienced at non-central listener positions.
- It is important to distinguish between imaging problems caused by system faults and those resulting from systematic errors caused by the acoustics of the projection space or the nature of the sound being projected.
- Decoding based on the diametric decoder theorem performs poorly for large arrays and should only be used for small listening areas.
- Fast moving sounds were more easily localized.
- The system can work well even for listeners placed outside the array, but not for listeners seated on the surface of the notional sphere of the loudspeaker array.

A comparison of Ambisonic rendered soundfields to VBAP based soundfields has also been implemented in [Pulkki and Merimaa, 2006] in both an anechoic environment and an ITU listening room. Here Pulkki et al. compare first order Ambisonics to the Spatial Impulse Response Rendering (SIRR) method. This method, which is fully described in Section 7.4 decomposes measured first order Ambisonic impulse responses into constituent directional and diffuse field components. The directional components are then panned using VBAP and the diffuse field portion of the W Ambisonics channel is decorrelated and sent to each loudspeaker in the reproduction array. For these tests, impulse responses extracted from a virtual acoustics simulation were utilized. The listeners were asked to grade the difference between the reference virtual

reality simulation and the reproductions. SIRR was found to yield the closest results to the reference presentations. Real B-Format impulse response measurements were also tested, and the listeners were asked to compare the naturalness of the reproduced soundfields. Again SIRR gave the best results. It was found that the Ambisonics system produced a more blurred perception of localization than with the SIRR system as well as some source coloration. Furthermore, for Ambisonics, precedence effects were reported at off-centre listening positions, where subjects localized the sources to the nearest loudspeaker. This is not surprising, since the Ambisonics decode utilized in these tests was of hypercardioid directionality, resulting in strong anti-phase components from opposite loudspeakers to the source. Furthermore, for the centrally seated listening position, the tests did not consider psychoacoustic optimization of the Ambisonic decode. Also, for SIRR, the reproduction quality increased when additional loudspeakers were introduced to the reproduction system.

During the course of the work presented in this thesis, the author has also presented several papers on the localization accuracy of spatialization systems in non-ideal conditions [Bates et al., 2007,Kearney et al., 2007]. The major results of this work will be shown here. Recently another study into the localization accuracy of HOA over a 12 channel loudspeaker array for off-centre listening has also been implemented by Frank et. al [Frank et al., 2008], which corroborates the results presented in this thesis. A comparison between the work presented here and that of Frank et. al is given in Section 4.3.

4.2.1 Localization Study Objectives

Despite the previous localization studies documented, many questions remain pertaining to the localization accuracy of commercial spatialization techniques in real listening environments and for distributed audiences. Based on the work presented in Section 4.2, we will now devise a formal subjective and objective comparison of 1st Order Ambisonics and VBAP, the purpose of which is to prove or disprove the following hypothesis

1. That the localization accuracy of amplitude panned virtual sources is equivalent to that of monophonic source listening in a real (reverberant) listening environment.

2. That the localization accuracy of virtual sources formed using 1st Order Ambisonics is equivalent to that of monophonic source listening in a real listening environment.

3. That the localization accuracy of 1st Order Ambisonic virtual sources is greater than that of VBAP virtual sources at the sweet spot, when compared to real sources in a real listening environment.

4. That the localization accuracy of 1st Order Ambisonic virtual sources is greater than that of VBAP virtual sources for off-centre listening positions in a real listening environment.

These hypotheses are first evaluated in terms of objective localization accuracy over an entire listening area. For this, we utilize the localization vector theory introduced in Section 3.4.3.

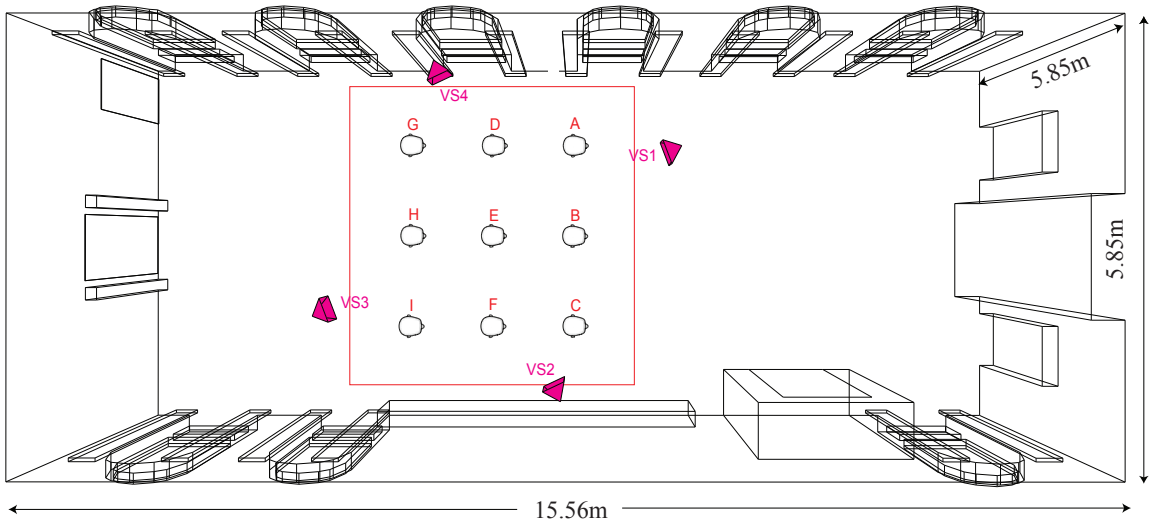


Figure 4.1: *9 listener audience in shoebox-style room acoustic.*

Subjective localization performance using perceptual listening tests is then conducted in a real hall. The results of these tests are then investigated through empirical binaural measurements at each listener position, as well as simulation in an equivalent acoustically modeled environment.

4.2.2 Objective Analysis for Single and Multiple Listeners

In this objective analysis, we will begin by investigating loudspeaker configurations best suited for subjective assessment between VBAP and Ambisonics for a distributed audience. Through this, we will also gain insight into the localization accuracy for a single listener at the sweet spot using these systems under different loudspeaker configurations. In general, for distributed audience scenarios, the choice of loudspeaker array is determined not only by the spatialization scheme adopted but also by the size of the reproduction environment as well as the audience area. For example, in cinematic type presentations, it is common to have three frontal loudspeakers and rear left and rear right loudspeaker arrays fed by the LS and RS tracks in a standard 5.1 reproduction scenario. However, this setup is sub-standard for our requirements, since we wish to reproduce good localization of virtual sources for all listeners and from all directions. When the frontal stage is fixed however, 5.1 and 7.1 setups work well for a single listener provided the information in the rear sound channels has a low correlation with the front channels (such as a diffuse room acoustic).

In this analysis, let us consider a small 9-listener audience in Test Environment 1, as shown in Figure 4.1. This is the same configuration as the monophonic source tests in Chapter 2. The audience area is approximately 4m wide. We will consider localization accuracy in full horizontal 360° plane, but we will pay particular attention to the formation of front, front-lateral, rear and rear lateral virtual sources located at $-22.5°$, $67.5°$, $157.5°$ and $247.5°$ with respect to listener position E. We will name these virtual source positions VS1, VS2, VS3, and VS4 respectively.

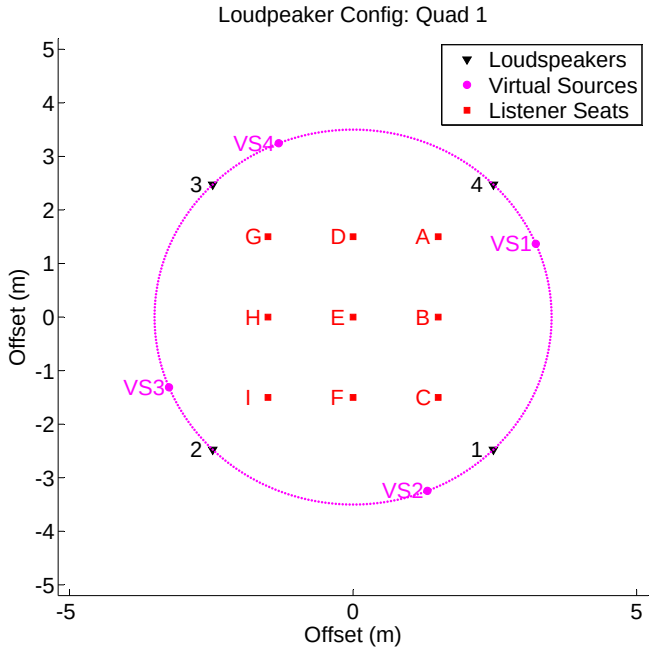
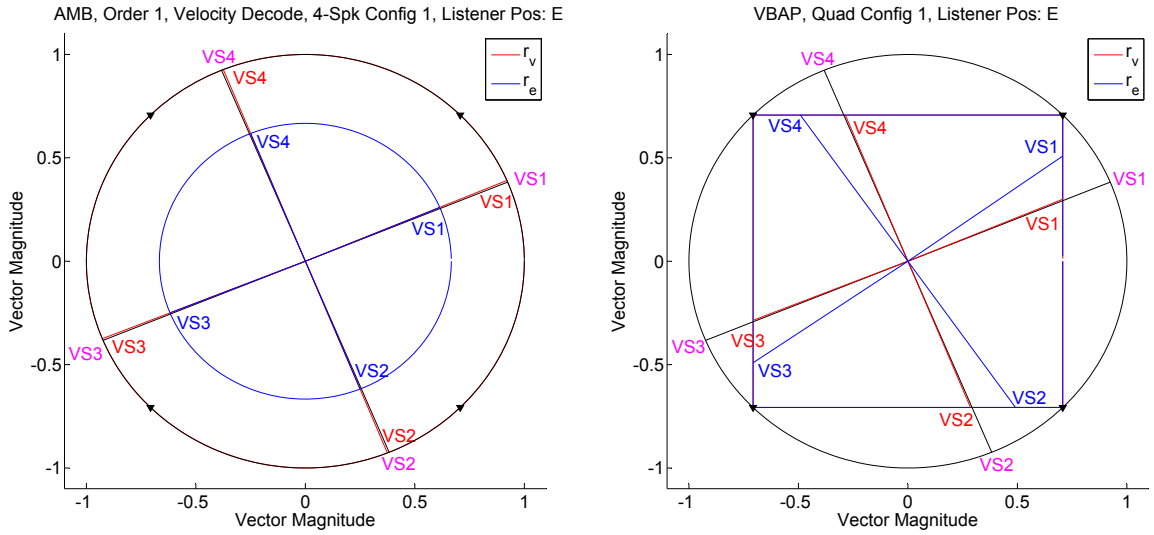


Figure 4.2: *Quadraphonic square layout,* $\theta_l = \pm 45°, \pm 135°$.



(a) 1st Order Ambisonics Velocity Decode

(b) VBAP

Figure 4.3: *Ambisonic and VBAP localization for Quad square layout,* $\theta_l = \pm 45°, \pm 135°$.

4.2.2.1 Sweet-Spot Localization

Typical arrangements found in small-sized auditoria to accommodate such listener groups can be as simple as the square quadraphonic setup shown in Figure 4.2. Here the loudspeakers are located at $\theta_l = \pm 45°, \pm 135°$ with respect to listener position E.

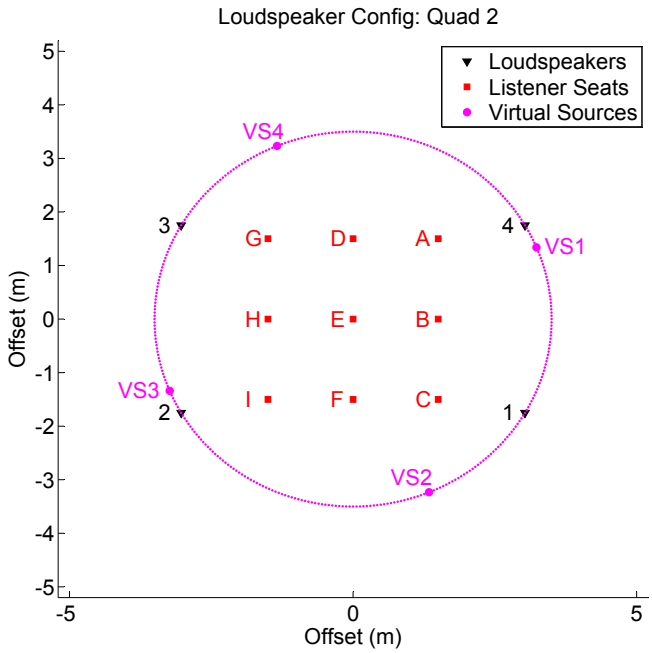
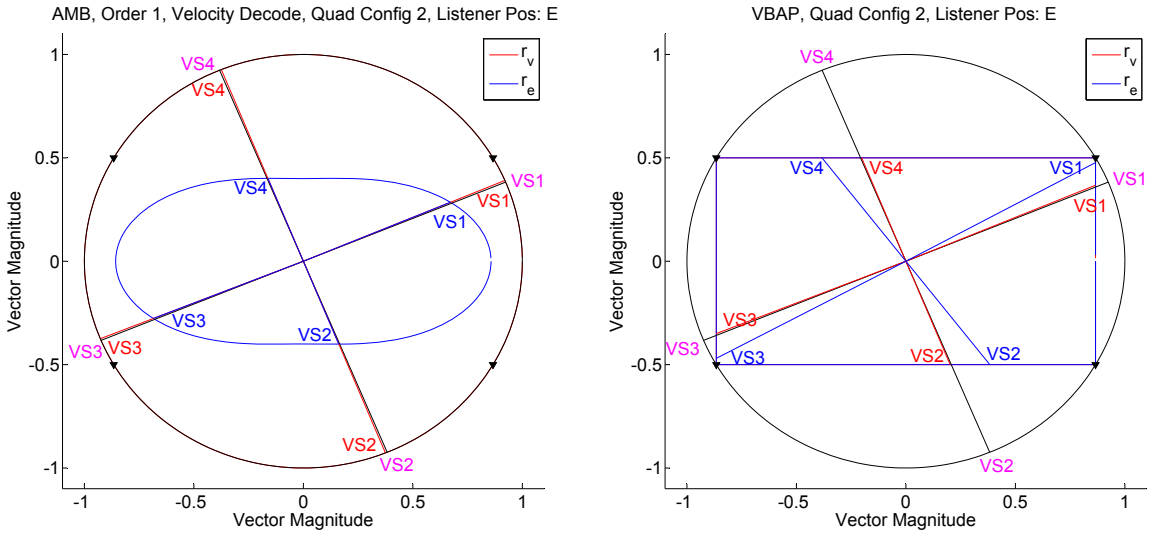


Figure 4.4: *Quadraphonic rectangular layout,* $\theta_l = \pm 30°, \pm 150°$.



(a) 1st Order Ambisonics Velocity Decode

(b) VBAP

Figure 4.5: *Ambisonic and VBAP localization for Quad rectangular layout,* $\theta_l = \pm 45°, \pm 135°$.

The Ambisonic velocity decode for the centre listening position (seat E) is shown in Figure 4.3. We see that since this layout is regular, the velocity vector is maximized, and a smooth albeit poor localization accuracy is indicated by the energy vector. This is a classic Ambisonic decoder design using Equations 3.38 through 3.42. However, the localization accuracy using VBAP over such a layout is poor. This is also shown in Figure 4.3. In VBAP, since all the

loudspeaker gains are positive, the value of the velocity vector will always be less than 1, except when the virtual source is panned to a loudspeaker. At this position, the source is in fact no longer virtual, and both r_v and r_e are 1. Furthermore, r_v and r_e are incongruent for this case.

The maximized velocity vector is one of the main benefits of Ambisonic systems over stereophonic systems for centrally seated listeners. We can therefore see the advantage to using Ambisonics over stereophonic panning on quad layouts, one of the original motivations for the design of Ambisonic systems [Gerzon, 1974].

One of the (many) problems with this loudspeaker arrangement is that symmetrical loudspeaker setups of this type can be difficult to realize, in particular if the audience area is non-symmetric, the room dimensions do not allow it, or the speakers are located too close to the audience. In such situations, the loudspeakers are moved to more convenient locations such as shown in the rectangular Quadraphonic layout of Figure 4.4.

We see that the velocity vector is still maximized indicating good low frequency localization, but the energy vector is biased in the directions of the loudspeakers. Whilst it is noteworthy that both r_v and r_e indicate the same direction, the localization accuracy is poor for high frequencies for virtual sources at the side of the listener, since the magnitude of r_e is low in these regions. This makes the formation of stable virtual sources even more difficult at the centre position. The localization accuracy of VBAP is again worse, suffering from greater deviations from the intended source directions as well as low values of r_v and r_e at the sides. Here the loudspeakers are placed at $\pm 30^\circ$ and $\pm 150^\circ$ with respect to listener position E. The result on the Ambisonic decode is shown in Figure 4.5.

An increase in the accuracy of rectangular or close to rectangular layouts in Ambisonic decoding may be achieved by introducing two side loudspeakers to increase the energy in this direction. This also benefits VBAP, since the side images become more stable, due to the fact that panning in the side directions results in (close to) mono reproduction. Such a hexagonal layout is shown in Figure 4.6. One should note here that not all the loudspeakers in this layout are equal distances from the sweet spot. The lateral loudspeakers are actually only 2.7m away from listener position E, whereas all other loudspeakers are at a 3.5m radius.

The resultant localization vectors are plotted in Figure 4.7. The energy vector for Ambisonics again shows a smooth response, but the localization accuracy is as before with the square Quadraphonic layout. This demonstrates that localization accuracy at the sweet spot in a 1st Order Ambisonic reproduction does not change with an increasing number of loudspeakers, as long as the layout satisfies the diametric decoder theorem. Conversely the accuracy of VBAP increases monotonically with number of loudspeakers. We see that for the hexagonal case there is still a deviation between the localization accuracy at low and high frequencies. If we increase the number of loudspeakers further to 8, then such discrepancies in VBAP begin to disappear. This can be demonstrated for the 8 loudspeaker setup of Figure 4.8. The VBAP velocity and

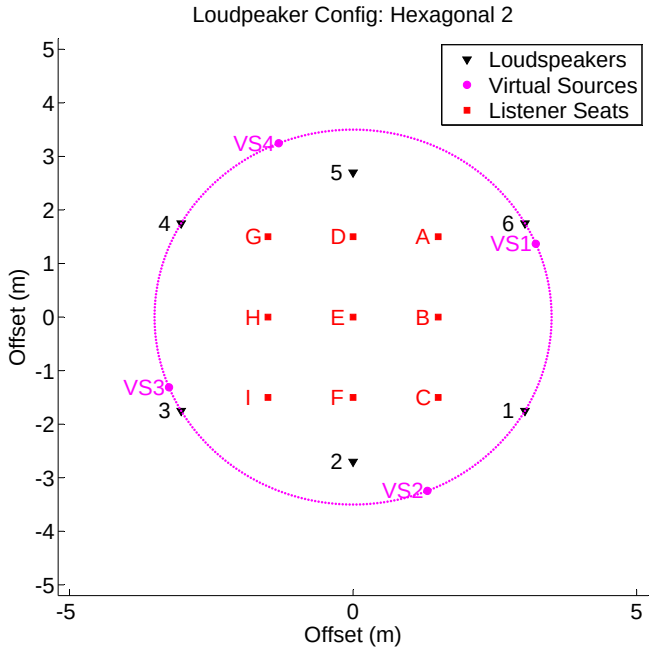
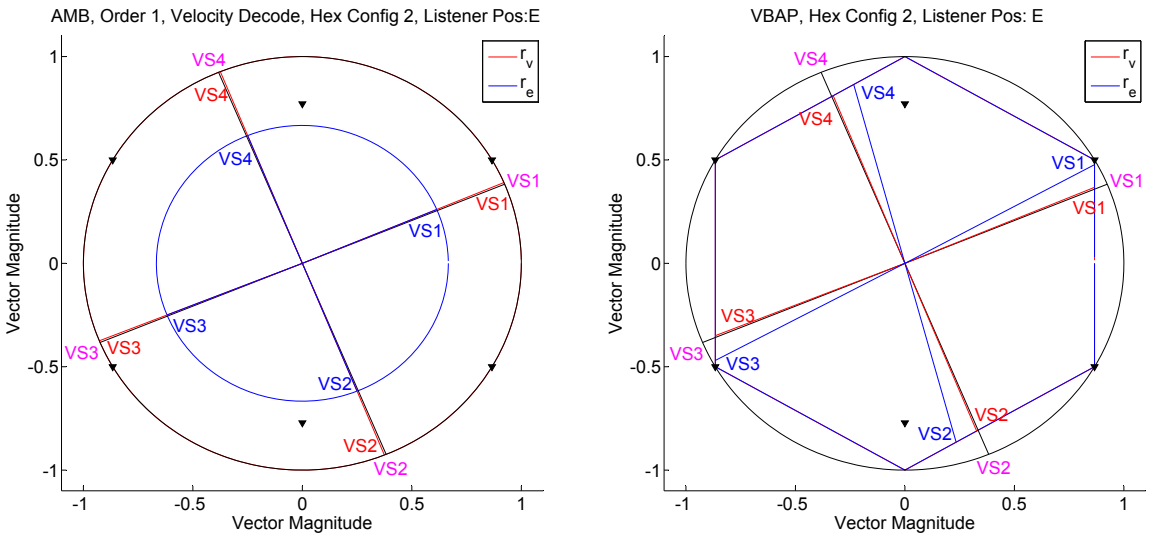


Figure 4.6: *Six loudspeaker hexagonal layout,* $\theta_l = \pm 45°, \pm 150°$.



(a) 1st Order Ambisonics Velocity Decode

(b) VBAP

Figure 4.7: *Ambisonic and VBAP localization for hexagonal layout.*

energy vectors are plotted in Figure 4.9. We see that the r_v and r_e match well for the different directions. The Ambisonic localization accuracy is the same as the square-quad and hexagon layouts, as expected.

Given that commercial high definition audio formats currently only support up to 8 channels of audio, an 8 channel delivery system represents a practical limit for pre-defined spatial audio

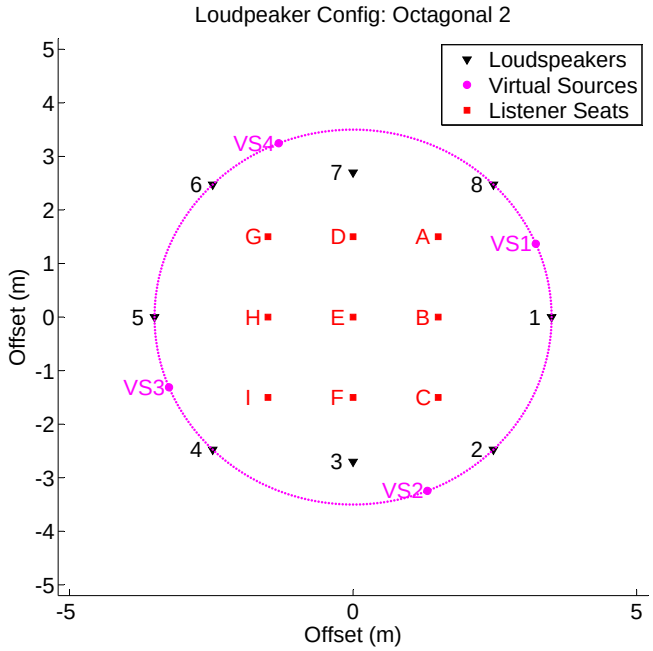
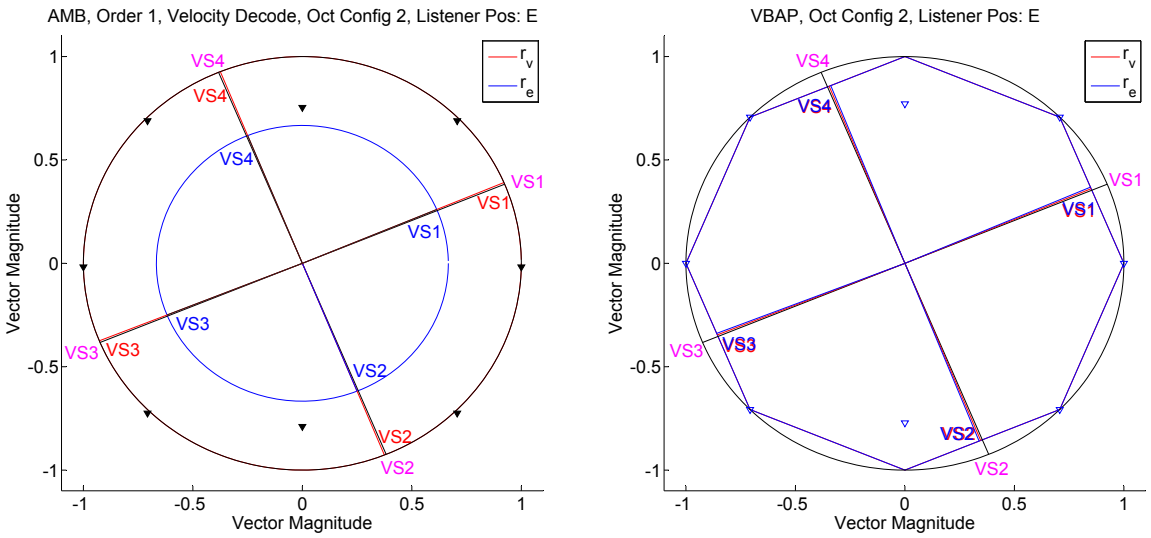


Figure 4.8: *Eight loudspeaker layout.*



(a) 1st Order Ambisonics Velocity Decode

(b) VBAP

Figure 4.9: *Ambisonic and VBAP localization for octagonal layout.*

presentations using VBAP. Whilst the number of loudspeakers *per channel* can be increased, there is no significant benefit other than better sound reinforcement over the listening area, i.e. increasing the loudspeaker count whilst maintaining the same information channel count will not increase the theoretical localization accuracy, but may increase the overall intelligibility in the presence of strong room acoustics. Likewise, in the 1st Order Ambisonics case, increasing

the number of loudspeakers will not lead to better localization accuracy at the sweet spot.

One may conclude from this analysis that for lower loudspeaker counts Ambisonic reproduction will give better localization accuracy whereas for higher channel counts VBAP can deliver better localization performance. Whilst this may be true at the sweet spot, the consequences of using lower loudspeaker counts in both Ambisonic and VBAP systems are completely underestimated in the context of distributed audience scenarios. Let us consider, once again each of the four reproduction layouts now in the context of multiple listener positions.

4.2.2.2 Off-Centre Localization Estimators

The energy vectors for off-centre listeners can be calculated by compensating the reproduction gains for the distance to the listener position. Equations 3.46 and 3.47 then become

$$e_x = \frac{\sum_{n=1}^{N} (g_n\ \alpha_n)^2 \cos(\theta_{l_n})}{P_e} \tag{4.1}$$

$$e_y = \frac{\sum_{n=1}^{N} (g_n\ \alpha_n)^2 \sin(\theta_{l_n})}{P_e} \tag{4.2}$$

and

$$P_e = \sum_{n=1}^{N} (g_n\ \alpha_n)^2 \tag{4.3}$$

These equations have the same form as before, with the exception of an amplitude scaling factor $\alpha_n = l_n/p$ where l_n is the length of the n^{th} loudspeaker vector $\mathbf{l}_n$ from the sweet spot and p is the distance of the listener to the loudspeaker. In this way, we consider the real near-field amplitude effects for off-centre listening. For off-centre listening, this simulation is therefore equivalent to a centrally seated listener listening to an imbalanced loudspeaker array.

However, the velocity vector analysis becomes more complicated, since the loudspeaker signals do not arrive at the off-centre listening positions at the same time. Thus, we consider the case where the phase at the off-centre listener position is completely random and we must employ a statistical analysis to determine the magnitude of the velocity vector. As is the case with the energy vector, we no longer assume that the contributing waves from each loudspeaker are planar, but instead are at a finite distance, and thus we employ the amplitude scaling factor α_n. For $k = (2\pi f)/c$ with random frequency f, and the phase vector $\varphi_n = k\mathbf{r}^T\mathbf{l}_n$, with random measurement point $\mathbf{r}$ in the audience area, we therefore obtain $V_\varphi = [v_x,\ v_y]$ with

$$v_x = \frac{\sum_{n=1}^{N} \alpha_n\ g_n e^{j\varphi_n} \cos(\theta_{l_n})}{\sum_{n=1}^{N} \alpha_n\ g_n e^{j\varphi_n}} \tag{4.4}$$

and

$$v_y = \frac{\sum_{n=1}^{N} \alpha_n\ g_n e^{j\varphi_n} \sin(\theta_{l_n})}{\sum_{n=1}^{N} \alpha_n\ g_n e^{j\varphi_n}} \tag{4.5}$$

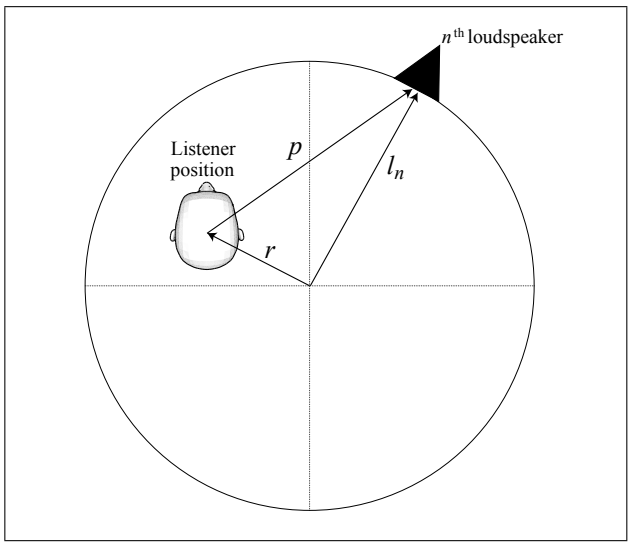


Figure 4.10: *Geometry of equations 4.4 and 4.5.*

In practice, we will use randomized frequencies and listener positions within the audience area in the computation of V_φ. We can then evaluate the density function using a kernel density estimator (KDE) with Gaussian kernels [Silverman, 1986]. Kernel density estimation allows us to extrapolate information about the entire sample population over the listening area given a smaller amount of data. It is a non-parametric method of estimating the probability density function of a random variable. The KDE differs from histograms in that it does not group observations into bins, but instead estimates the distributions at each observation point (determined by the filter kernel, which is Gaussian in this case) and sums these distributions to give the overall estimate of the probability density function. A much smoother result is then obtained than with histogram density estimates.

For this analysis we will generate 10000 random phase vectors, φ, representing random frequencies and listening positions. This analysis is similar to that presented by Daniel [Daniel et al., 1998]. However, here we consider that the loudspeakers are at a finite distance from the listener and so near-field amplitude differences are presented.

4.2.2.3 Off-Centre Localization Analysis

The energy vectors for off centre position F for VBAP in the 4, 6 and 8 channel setups are shown in Figure 4.11. For a full analysis of the VBAP energy vectors at all listening positions, refer to Appendix C. The loudspeaker positions are also overlayed in the vector plots to aid identification of high error regions and one should note that the coordinate system is relative to the listener position. The corresponding localization blur plots are also shown for each setup in Figure 4.12.

It can be seen that as the number of loudspeakers increases the localization blur decreases significantly at off centre listening position F. In the quad-rectangle case, large energy vector

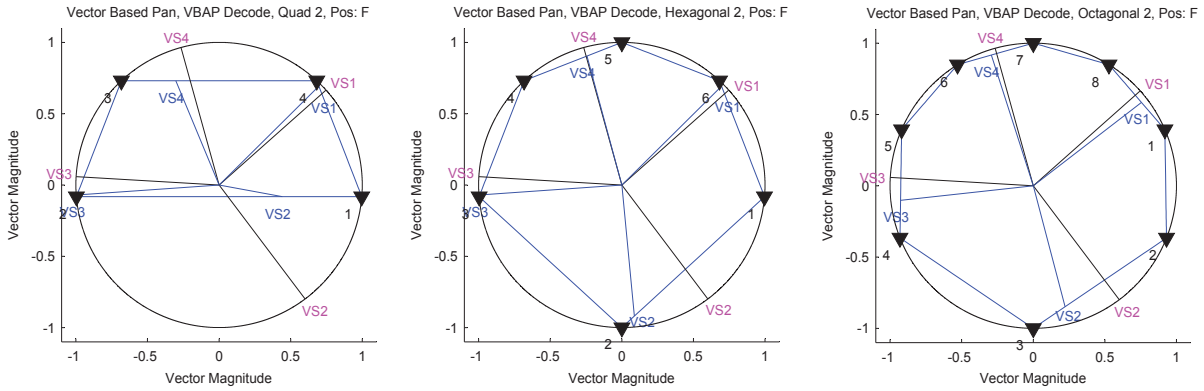


Figure 4.11: *VBAP energy vector for Quad rectangle, Hexagonal and Octagonal layouts for listener position F (Velocity decode).*

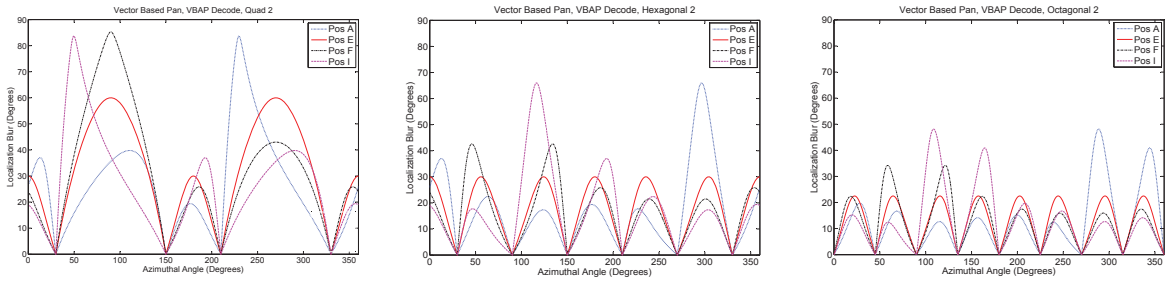


Figure 4.12: *VBAP localization blur for Quad rectangle, Hexagonal and Octagonal layouts for listener positions A, E, F and I (Velocity decode).*

errors are exhibited between loudspeakers 1 and 2. This shows, as we might expect, the region of largest localization instability for this listener position. This is not surprising since, if we refer back to Figure 4.2, we see that this position lies equally close to these loudspeakers. The localization blur for this position reaches 86°. Similar results are found for off-centre positions A and I, and significant localization errors are also exhibited at the sweet spot. These errors tend to zero when the source is panned to a single loudspeaker. The error is significantly reduced for the 6 loudspeaker case, with maximum localization blur of 66° at positions A and I. We note that the localization blur at the sweet spot is below 30°. As we might expect, further improvements are made in the 8 loudspeaker case, with maximum localization blur of around 45°. In all three cases, the energy vectors tell us that source localization is drawn towards the nearest loudspeaker in each pairwise pan. This is observed by viewing the deviations between the correct and localized virtual source angles.

The velocity vector maps, calculated from the kernel density estimation of equations 4.4 and 4.5 are shown in Figure 4.13 for the 4, 6 and 8 loudspeaker setups as well as a monophonic reference. The maps are calculated for the unfavorable source angle of 66.5° (VS2). The maps display the distribution of the magnitude of the velocity vector calculated for 10000 random

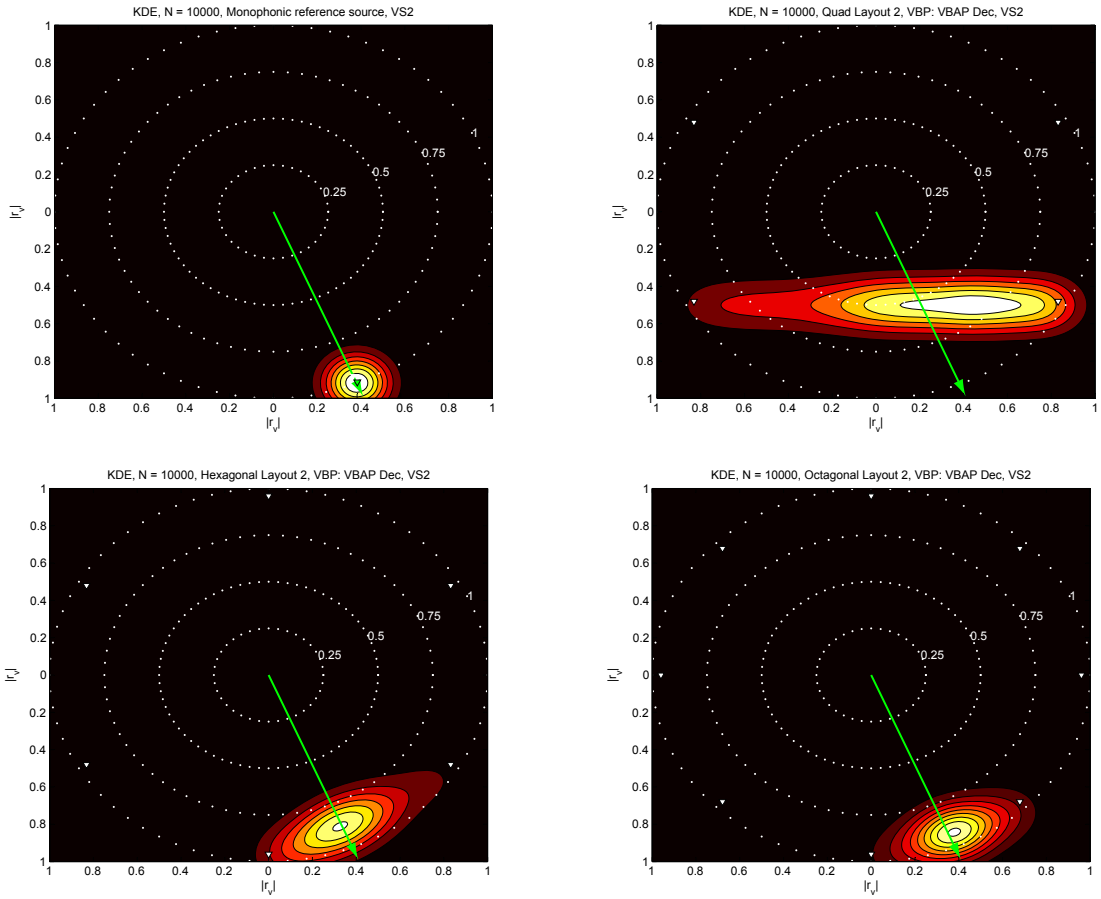


Figure 4.13: *Kernel density estimate of the probability density function for r_v with VBAP: Monophonic reference, quad rectangle, hexagonal and octagonal layouts. The green arrow represents the intended source direction.*

---

phase vectors. Regions of high density indicate the expected localization accuracy for any given position within the array. These velocity vector plots corroborate the results of the energy vector analysis. For the four channel case, the distribution is located entirely within loudspeakers 1 and 2, with a strong pull to loudspeaker 1, which has greater positive gain in this case. The accuracy is improved significantly, in the six and 8 channel cases, but only because of the smaller angle between the loudspeakers. The localization accuracy does not reach that of the monophonic reference, but the fact that localization is guaranteed to be between a given loudspeaker pair is certainly a major advantage of using VBAP.

In the Ambisonics case, we will first look at distributed audience localization using a classic Ambisonic decode i.e. 1st Order Ambisonics with a velocity decode. The energy vectors, again shown for off-centre position F are given in Figure 4.14. Two very important differences between the energy vectors for VBAP and for 1st Order Ambisonics can be seen here. Firstly, the magnitude of the energy vector is always less than 1 in the Ambisonics case, as was previously

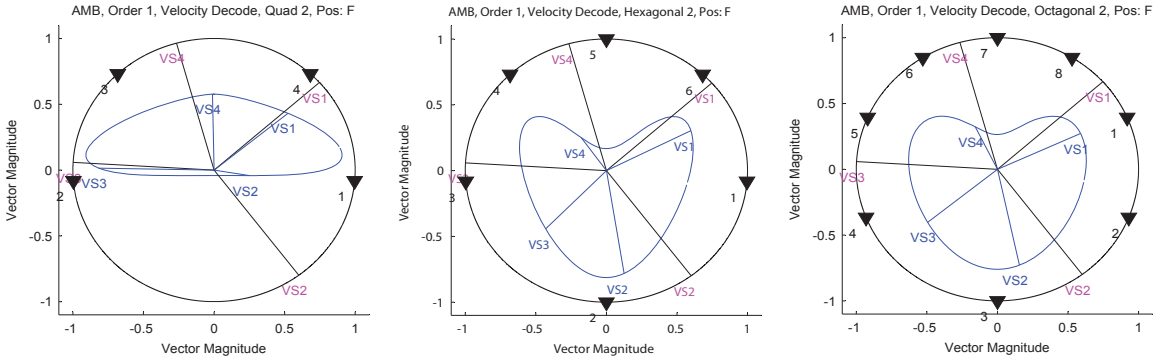


Figure 4.14: *Ambisonic energy vector for Quad rectangle, hexagonal and octagonal layouts for listener position F (Velocity decode).*

seen with the sweet spot examples. Conversely, since the off centre positions are nearer to the loudspeakers, the magnitude of the energy vector can be greater than in the sweet spot case when the source is in the direction of the near loudspeakers. This near-speaker bias also manifests as localization error in the direction of the nearest loudspeaker. Secondly, we note that the energy vector changes more smoothly than in the VBAP case, but unlike the sweet spot localization in layouts that satisfy the diametric decoder theorem, it is not consistent over all angles. The localization blur, shown in Figure 4.15 is very large for off-centre listening positions. The Quad rectangle layout displays the largest fluctuations in localization blur, peaking at 90°. The localization blur for the hexagonal and octagonal layouts is comparable and both give the exact localization blur of approx 49° at the sweet spot as expected. However, some differences can be seen at the off-centre listening positions with somewhat smoother and reduced blur for the 8-channel case.

One may be quick to conclude that the localization accuracy for off-centre positions in Ambisonic systems increases the higher the number of loudspeakers. However, this is not the general case as can be seen from Figure 4.16. There is, in fact, little difference in the localization accuracy exhibited by an 8-channel array and a infinite loudspeaker array for a first order Ambisonic decode. In conclusion, the greatest changes in localization accuracy for a distributed audience occur with lower loudspeaker counts. On average, large high frequency localization errors are exhibited over the listening area with layouts of 4 and 6 loudspeakers. For 8 loudspeakers and above the average values of the energy vectors stabilize, with an average localization blur between 48° to 76° for a velocity decode.

The velocity vector maps for 1st Order Ambisonics over the different loudspeaker layouts are shown in 4.17. Again we review the source at VS2 (67.5°). Beginning with the Quad decode, we see that localization accuracy is extremely poor for low frequencies. The highest density region is close to a vector magnitude of 1, but in the direction of forward loudspeaker 1. Other

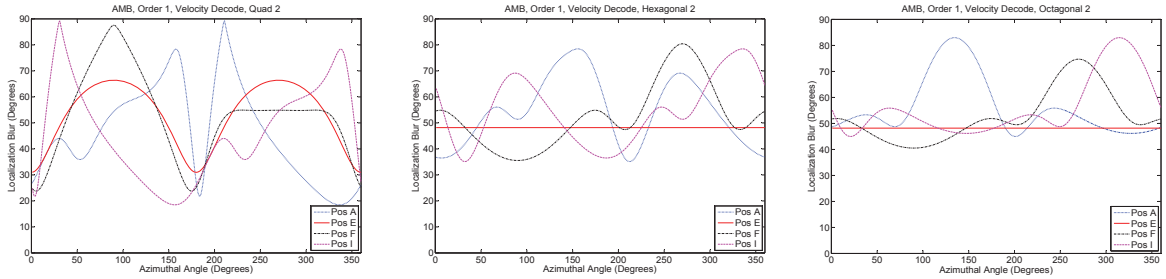


Figure 4.15: *Ambisonic localization blur for Quad rectangle, hexagonal and octagonal layouts for listener positions A, E, F and I (Velocity decode).*

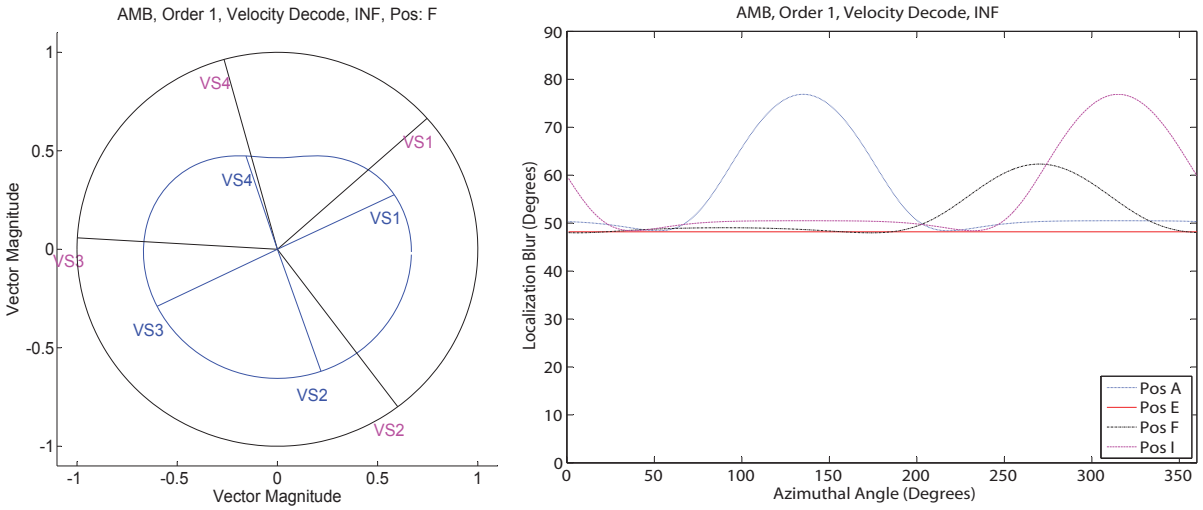


Figure 4.16: *Ambisonic localization vectors and error for infinite loudspeaker reproduction at listener position F (Velocity decode).*

---

high density areas can be seen at loudspeaker 2 and in the direction of loudspeaker 3. The results are also poor in the six and eight channel cases. It is noteworthy that the value of the velocity vector in the high density regions is greater than reported in other studies such as in [Daniel et al., 1998]. This does not indicate better localization accuracy however, since the high density regions are located in the direction of the loudspeakers. This is because the model presented here also considers the finite distance of the loudspeakers, and as such localization vectors are pulled towards the loudspeakers near the listener position, with a high value of r_v. This effect is clearly seen with all loudspeakers in the six channel layout, and loudspeakers 1, 2, 3 and 5 in the 8 channel layout. It is clear from the analysis in this section that VBAP displays better localization accuracy than 1st Order Ambisonics (with velocity decoding), and is best over the 8-channel layout. The energy vector analysis also showed that the octagonal configuration gave the best results. We can therefore conclude that an 8 channel reproduction setup is a reasonable medium for further comparison of Ambisonics and VBAP under real

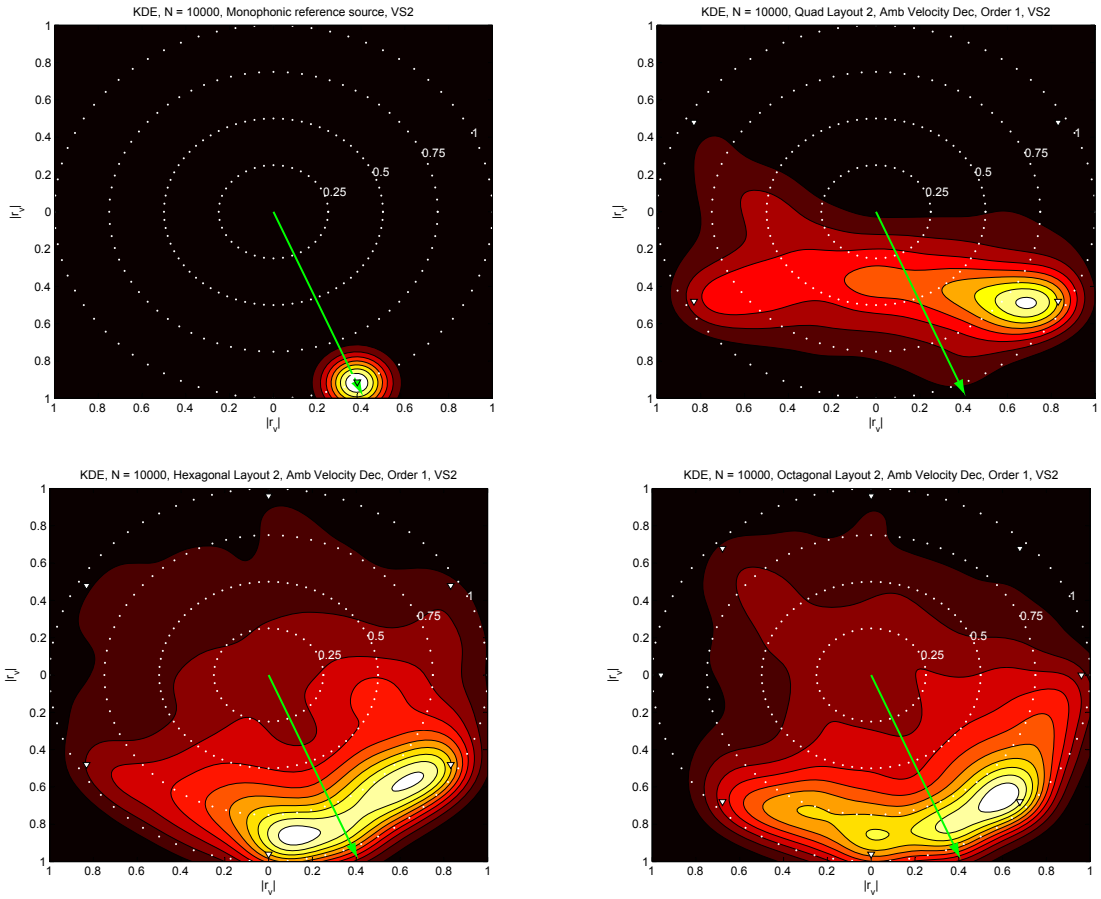


Figure 4.17: *Kernel density estimate of the probability density function for r_v with Ambisonics (1^{st} Order Velocity Decode): Monophonic reference, quad rectangle, hexagonal and octagonal layouts. The green arrow represents the intended source direction.*

---

listening conditions. However, before we continue, we must also consider how the localization accuracy of VBAP and 1st Order Ambisonics changes with the decoder type over the listening area.

4.2.2.4 Decoder Optimization

Firstly, it is possible to make the VBAP energy vector match the VBAP velocity vector by employing psychoacoustic optimization on the loudspeaker gains. This is known simply as the Vector Base Panning approach, which employs vector based amplitude panning at low frequencies and vector based intensity panning (VBIP) at high frequencies (see [Jot et al., 1999]). For example, in a two channel pan, the low frequency normalization of VBP loudspeaker gains (VBAP decode) is given according to:

$$g_n = \frac{a_n}{a_1 + a_2} \tag{4.6}$$

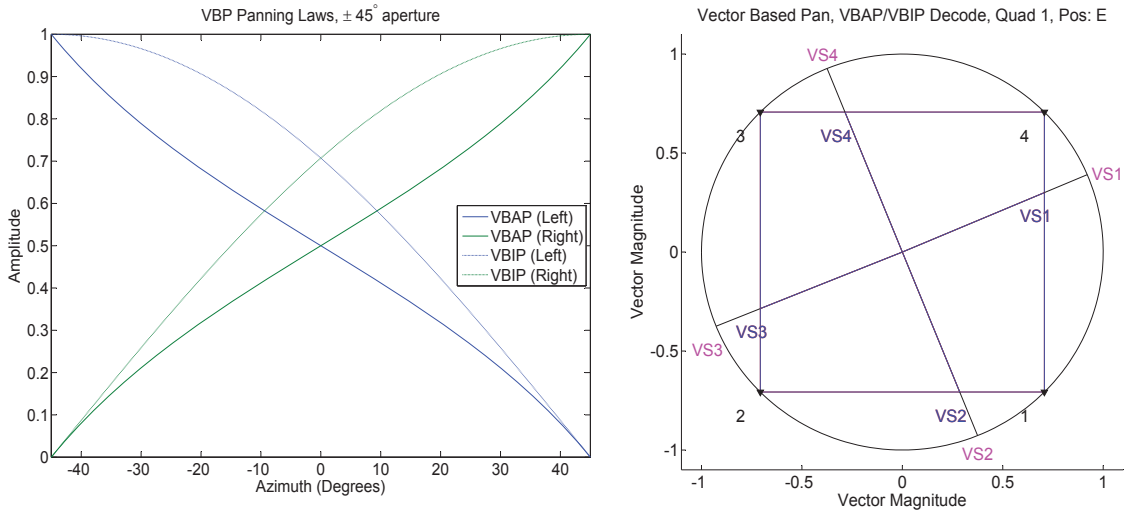


Figure 4.18: *Vector Base Panning, VBAP/VBIP decode, Quad layout 1: Panning law and localization vectors.*

---

where a_1 and a_2 are the non-normalized gains. The high frequency normalization (VBIP decode) is given by the power normalization

$$g_n = \frac{a_n}{\sqrt{(a_1^2 + a_2^2)}} \tag{4.7}$$

In essence, we now have two panning laws for low and high frequencies. An example of VBAP/VBIP panning laws for $\pm 45°$ is given in Figure 4.18. These specific panning laws could be employed with the quadraphonic layout of Figure 4.2. Also shown in Figure 4.18 is the resultant velocity and energy vectors for the Quad layout. We see that the localization vectors are now in good agreement.

However as we have seen, the difference between the velocity and energy vectors when only using VBAP becomes smaller the greater the number of loudspeakers. A comparison between an 8 channel VBAP decode and an optimized VBAP/VBIP decode is shown in Figure 4.19. We see that for the 8 channel case there is in fact little difference between psychoacoustically optimized vector based decoding and just VBAP.

For the Ambisonics case, the 'cardioid' or 'in-phase' decode is generally recommended for off-centre listening since out of phase components are not reproduced [Malham, 1992]. However, the localization accuracy is far from ideal. An example of localization accuracy for off-centre listening position F is shown in Figure 4.20. As we expect, the value of the energy vector is low, indicating poor high frequency localization. The optimal choice lies with the 'energy' decode where the energy vector is greater than than in the 'in-phase' decode case. However, we note that the localization accuracy is once again marred by nearby loudspeakers (speaker 3 in this case). This is the case with all listener positions, as can be seen in Appendix D. The

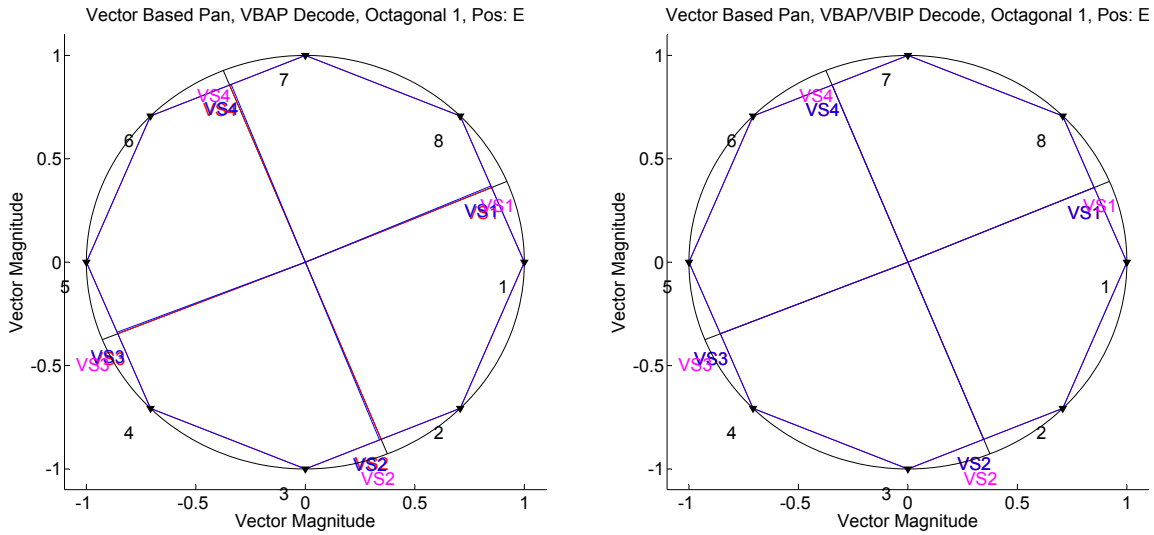


Figure 4.19: *Comparison of localization vectors for an 8 channel VBAP decode and an 8 channel VBAP/VBIP decode.*

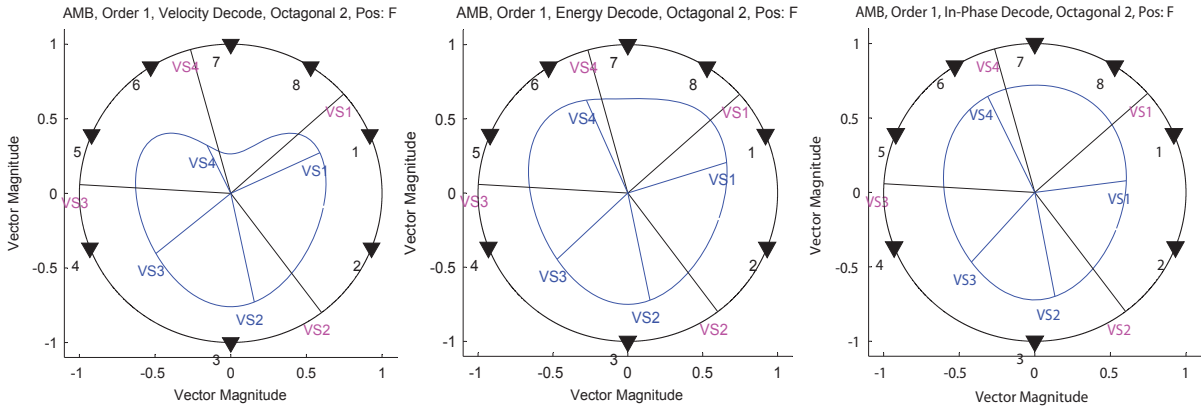


Figure 4.20: *Localization vectors for Velocity, Energy and In-Phase Ambisonic decoders at off-centre listening position F.*

localization blur for the different decoders for off-centre listeners is shown in Figure 4.21. We see that localization improvements over the velocity decode are made with the energy and in-phase decodes. This is a result of the reduced energy from the anti-phase components in their decodes. For the centrally seated listener, the localization accuracy is best using the energy decode, and is comparable to the in-phase decode for off-centre positions. We note that the localization blur in the energy decode fluctuates more than in the in-phase decode, indicating that panning may smoother with the latter.

The KDE velocity vector maps for the different decoders are shown in Figure 4.22. Once again, the in-phase and energy decodes display significant improvements over the velocity decode.

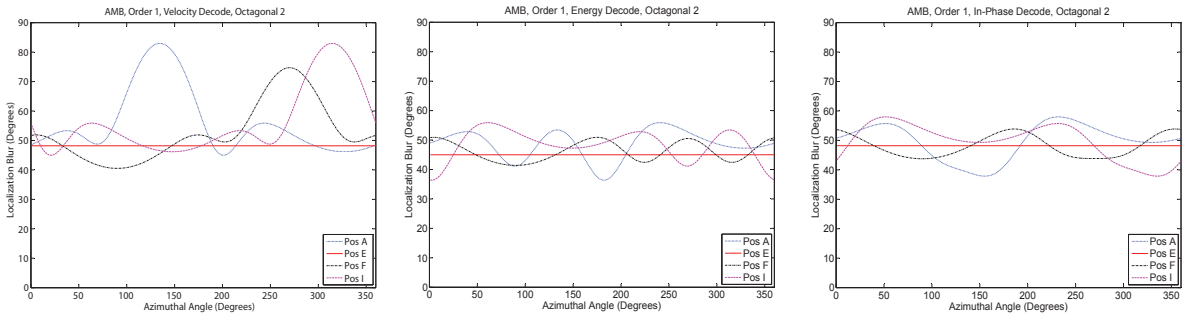


Figure 4.21: *Ambisonic localization blur for 1st Order velocity, energy and in-phase decoders for listener positions A, E, F and I (Velocity decode).*

The velocity decode (and therefore a shelf filter decode) create a significant increase in r_v in the direction of the opposing loudspeakers. Such phenomena are reduced in the energy and in-phase decoders. However, the spread of the velocity vector indicates that low frequency localization is best with the energy vector. This is further reinforced by the fact that the highest density region is in the direction of the virtual source with $|r_v|$ approximately equal to 0.7. One should note however that the spread of the velocity vector distribution for the energy decode case still demonstrates a significant low frequency localization blur.

4.2.3 Comments on Objective Analysis

From the analysis presented we can conclude the following for both VBAP and 1st Order Ambisonic systems:

1. The localization accuracy of 1st Order Ambisonic decoders is inferior to that of VBAP under ideal listening conditions. However, energy vector optimization of Ambisonic decoders leads to improved localization accuracy. The subjective performance of both VBAP and energy optimized Ambisonics under real listening conditions remains to be seen.

2. Localization accuracy with both systems is optimal with regular layouts. However, a good performance can also be achieved using the pseudo-rectangular 8-channel layout presented, which trades off totally symmetrical layout with maximum audience area.

3. Both systems do not work well for a distributed audience with lower channel counts. For VBAP, the difference between the velocity and energy vectors for an 8 channel layout is negligible and is almost equivalent to VBAP/VBIP panning on the same layout. For Ambisonics, a higher channel count increases localization accuracy for off centre listeners, but only up to 8 channels. After this, higher channel counts do not improve localization.

Based on these conclusions, we can now establish a formal listening test strategy to investigate fully the localization performance of the spatialization schemes for a distributed audience in a reverberant environment.

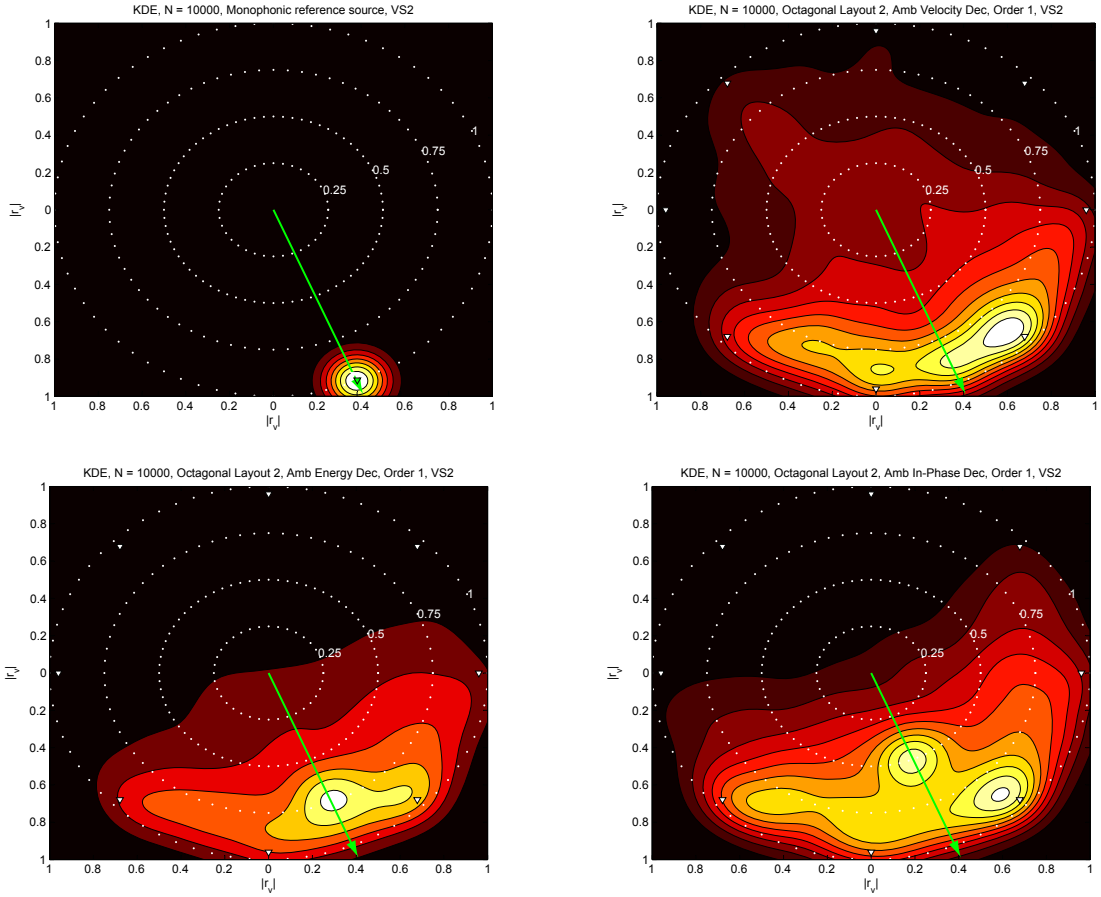


Figure 4.22: *Kernel density estimate of the probability density function for r_v with different Ambisonic decoders for an 8-channel array: Monophonic reference, velcity decode, energy decode, in-phase decode. The green arrow represents the intended source direction.*

4.2.4 Subjective Listening Experiment II.A: Localization of VBAP and 1st Order Ambisonics Under Real Listening Conditions

Based on the previous analysis, a series of experiments were set up in a small reverberant hall in Trinity College Dublin with the objective of comparing localization using VBAP and 1st Order Ambisonics using the 'energy decode' to that of monophonic sources. The configuration and test conditions were exactly the same as the monophonic listening tests conducted in Chapter 2, where a loudspeaker array consisting of 16 Genelec 1029A loudspeakers was arranged around the 9-listener audience area as shown in Figures 4.23 and 4.24.

4.2.4.1 Experimental Procedure

Subjects were presented with virtual and monophonic sources from pseudo-random positions located about the speaker array and were then asked to identify the location of the sources via



Figure 4.23: *Loudspeaker array and audience area for subjective listening tests.*

a questionnaire running concurrently with the tests. In these tests, monophonic sources were presented using the 8 black loudspeakers shown in Figure 4.24 while the 8 gray loudspeakers were used by the various spatialization techniques to generate virtual sources at the same positions. This method allows for a direct comparison between the localization accuracy for a real source and a virtual source positioned at the same location. Unfiltered recordings of male speech, female speech, Gaussian white noise and music with fast transients were used throughout in order to assess the effect of different spectral and temporal stimuli on localization accuracy. These characteristics are shown in Figure 4.25. The loudspeakers were calibrated to 80dBA at the centre position of the array using pink noise (band-limited from 500 to 2000Hz), and their axis lines were coincident with the centre listener position. Each of the 27 subjects, which consisted primarily of music technology students under 35 years of age, were screened before the tests for potential hearing impairments.

Of particular interest were the phantom and real sources presented from the directions outlined in Figure 4.1 i.e., the same front, front-lateral, rear and rear-lateral directions analyzed in the previous objective study. Randomization of the test sample playback was used to negate any order effects during the tests and each sample was presented twice. Listeners were asked to keep their heads in the forward direction throughout the test, but this was not strictly enforced. The test was repeated such that every subject performed the test in each of the nine listening positions. Each of the listeners' answers were weighted, depending on the confidence level of the listener with their choice, with weightings of $1/n$, where n is the number (or range) of speakers that a listener felt the sound originated from.

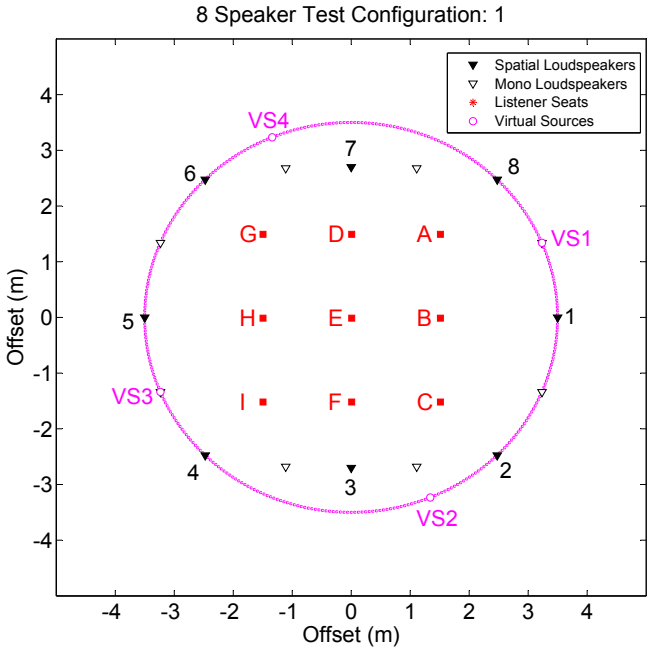


Figure 4.24: *16-loudspeaker configuration for listening tests.*

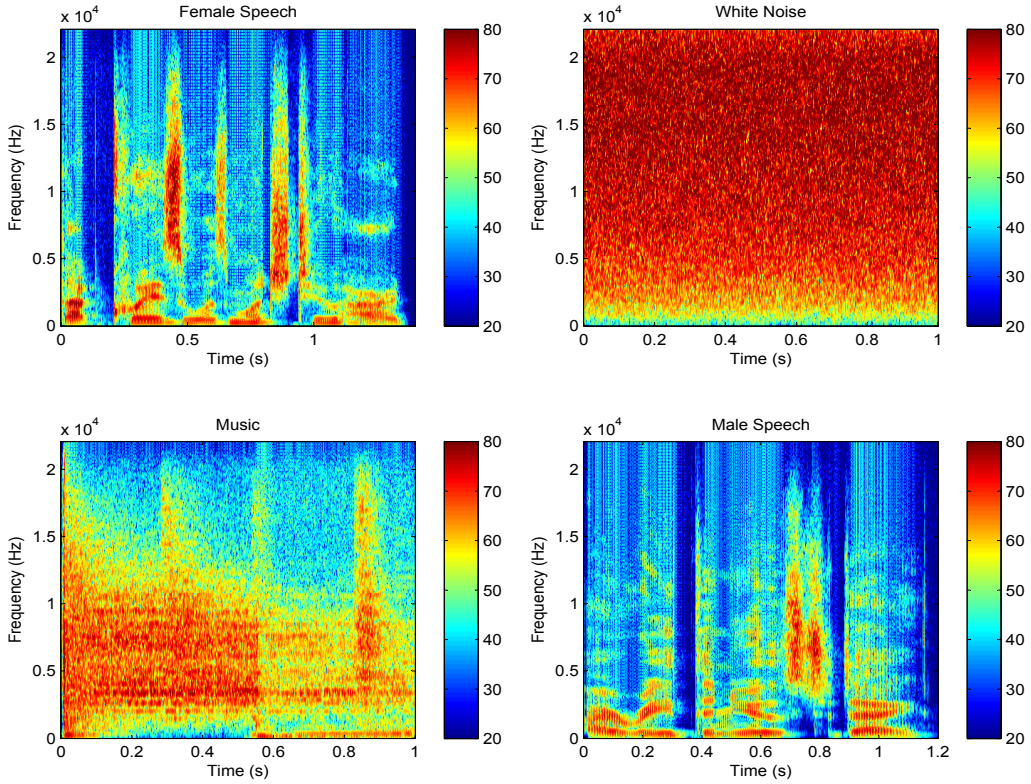


Figure 4.25: *Spectral and temporal characteristics of source samples.*

4.2.4.2 Results

In the analysis of the collected data, we define the following variables:

- v: Spatialization system = Monophonic , 1st Order Ambisonics (Energy Decode), VBAP,
- w: Stimuli = Female, male, music, noise,
- x: Monophonic source position = VS1 ($-22.5°$), VS2 ($67.5°$), VS3 ($157.5°$), VS4 ($247.5°$),
- y: Listener seats =A, B, C, D, E, F, G, H, I,
- z: Chosen loudspeaker position=1:16.

For the first analysis, we will compute several normalized histograms from the data collected, as estimates of the probability density functions of z, knowing x and w only, i.e. the listener seat y is not taken into account in this first experiment. For instance, we will compute the histograms $p_1(z) = p(z|x = VS2, w = \text{female speech}, v = \text{mono})$ and $p_2(z) = p(z|x = VS2, w = \text{female speech}, v = \text{VBAP})$ and we compute a Euclidean distance d between these two along with its standard error se. This is calculated by

$$d = \left\{ \sum_{z=1}^{16} \left(p_1(z) - p_2(z) \right)^2 \right\}^{\frac{1}{2}} \tag{4.8}$$

and

$$se(d) = \left(\sum_{z=1}^{16} \frac{p_1(z)(1 - p_1(z)) + p_2(z)(1 - p_2(z))}{N} \right)^{\frac{1}{2}} \tag{4.9}$$

for N listener answers. The results are depicted in Figure 4.26. The axis labels can be read as:

- FS = Female Speech
- MS = Male Speech
- MU = Music
- WN = White Noise
- VS1 = Phantom Source at loudspeaker position 2
- VS2 = Phantom Source at loudspeaker position 6
- VS3 = Phantom Source at loudspeaker position 10
- VS4 = Phantom Source at loudspeaker position 14

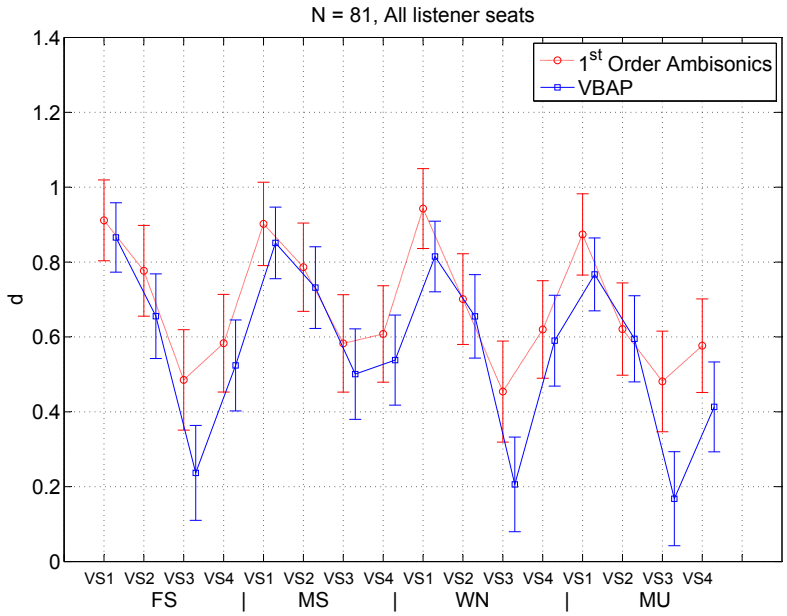


Figure 4.26: *Comparison of 1st Order Ambisonic and VBAP localization accuracy to real source presentations of female speech, male speech, music, and white noise over all 81 listeners for all seats.*

This distance is the first point in Figure 4.26 on the blue curve. Likewise the first point on the red curve represents the euclidean distance between $p(z|x = 2, w = \text{female speech}, v = \text{mono})$ and $p(z|x = 2, w = \text{female speech}, v = \text{AMBI})$ and the standard error. Since we are evaluating 9 listener positions, with 9 listeners answers taken at each position, $N = 81$ listener answers are used to fill these 16 histogram bins.

It can be seen that the distance between localization using VBAP and localization of real sources is smaller than the distance between 1st Order Ambisonics and mono. Thus, spatialization using VBAP behaves closer to mono than Ambisonics does in this case. However the distance is significantly away from zero which shows that neither VBAP or 1st Order Ambisonics achieve a localization accuracy equivalent to that of real sources. For VBAP, the best localization is achieved using broadband sources such as white noise, followed by music. The localization is poorest for the male speech sample. For the Ambisonics case, the localization accuracy amongst sources is relatively consistent, indicating that source material, and therefore broadband spectral content is not as significant a factor in the poor off centre localization of this system as other potential causes such as loudspeaker proximity. It is interesting to note that both VBAP and Ambisonics obtain a local minimum for a virtual source VS3 in all stimuli cases. It is also noteworthy, from the results presented in Section 2.4.3 that the real source which is located at this position is also difficult to localize. In such a case, the localization accuracy of both VBAP and Ambisonics is more similar to real source localization only because of increased lateralization blur.

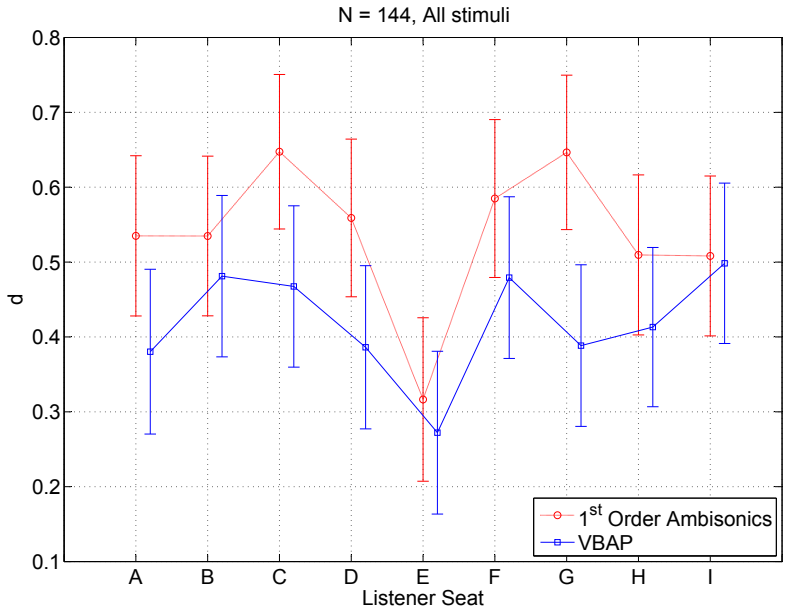


Figure 4.27: *Average localization accuracy in comparison to monophonic sources over all 81 listeners for all stimuli.*

In our second analysis of the captured data, we will focus on localization accuracy over the listening area. This time we will not concern ourselves with the source stimuli or its position. From the data we compute the two-dimensional histogram $p(z, y|v)$ with $N = 144$ votes (4 source positions x 9 listeners x 4 stimuli) filling 16 x 9 bins. For example, the first point on the blue curve of Figure 4.27 is the distance between the density $p(z|y = A, v = \text{mono})$ and $p(z|y = A, v = \text{VBAP})$, along with the standard error. This is repeated for seats $y = B : I$. The abscissa is the seat number. We note that the distance is the smallest for seat E, confirming that even in such a reverberant room, the localization using VBAP and 1st Order Ambisonics becomes closer to that of real sources at the sweet spot, as expected.

The above results indicate that neither VBAP nor 1st Order Ambisonics can deliver equivalent localization accuracy to that of monophonic sources under real listening conditions. As predicted through analysis of the localization vectors, VBAP gives a closer representation to monophonic sources than 1st Order Ambisonics under these conditions. However, the results are not as clear-cut as one may have anticipated from the previous analysis. The subjects were tested using a forced-choice, speaker identification method which could explain the high degree of correlation between the mean results. The standard error does not vary considerably for different listening and source positions which is surprising considering the non-ideal listening conditions. However the mean localization difference does change significantly. We note that the points of greatest localization error for Ambisonics are at listener positions C and G, both of which are at extremities of the array. The poorest localization for VBAP is exhibited at listener position I. Overall, these results are in agreement with the predictions of the localization vectors,

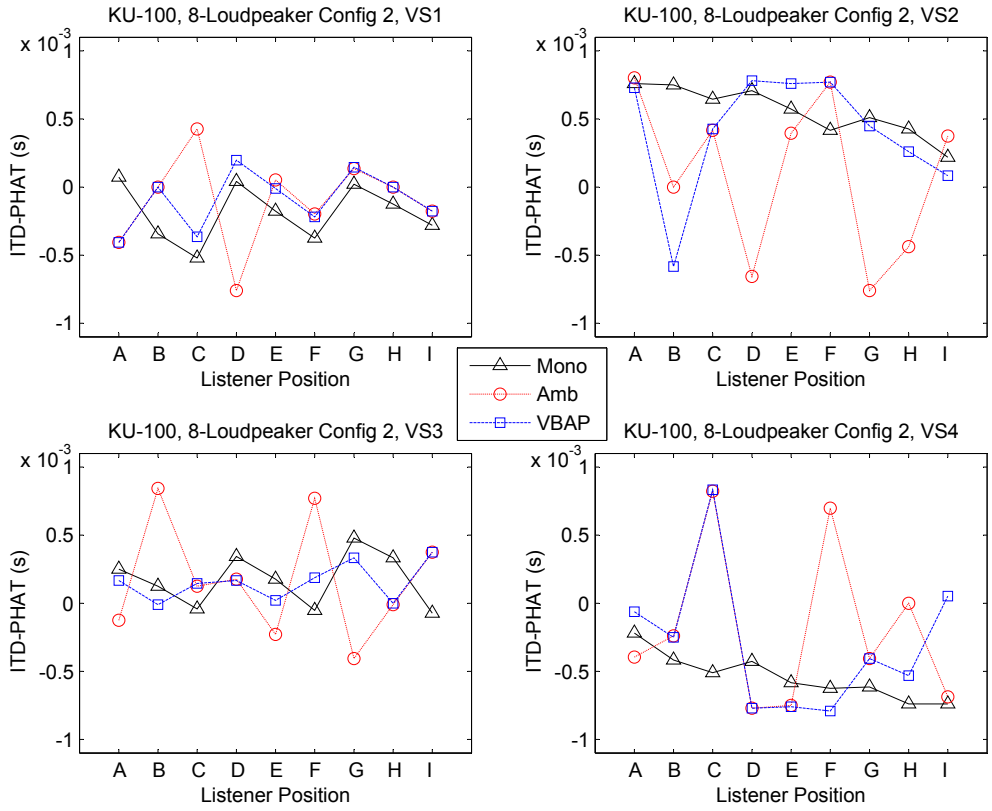


Figure 4.28: *Measured Interaural Time Difference at each listener position for each spatialization method.*

that localization accuracy is degraded for non-central listeners seated close to the array.

4.2.4.3 Room Investigations

Binaural impulse response measurements were taken at each listener point in the hall using a forward facing Neumann KU100 dummy head for the same angular presentations as the subjective experiments. For each measurement, a logarithmic sine-tone sweep was encoded for playback via VBAP and 1st Order Ambisonics. A monophonic presentation at each intended virtual source position was also used as a reference. From these recordings the low-frequency ITDs were extracted for each source location using the IACC-PHAT method introduced in Chapter 2. The results are shown in Figure 4.28. The effect on the ITD at off centre listening is immediately apparent. Spatialized audio presentations using both 1st Order Ambisonics and VBAP result in ITDs which pull the image in the direction of the nearest loudspeaker. For example, errors that occur at listener position C have consistently positive ITDs, showing that they are being pulled to the right, towards loudspeaker 2. Likewise, at listener position D, the ITDs display a negative bias, indicating that the image is being pulled in the direction of loudspeaker 7. We note that large errors do not occur at the sweet spot (position E), but there exists a constant deviation

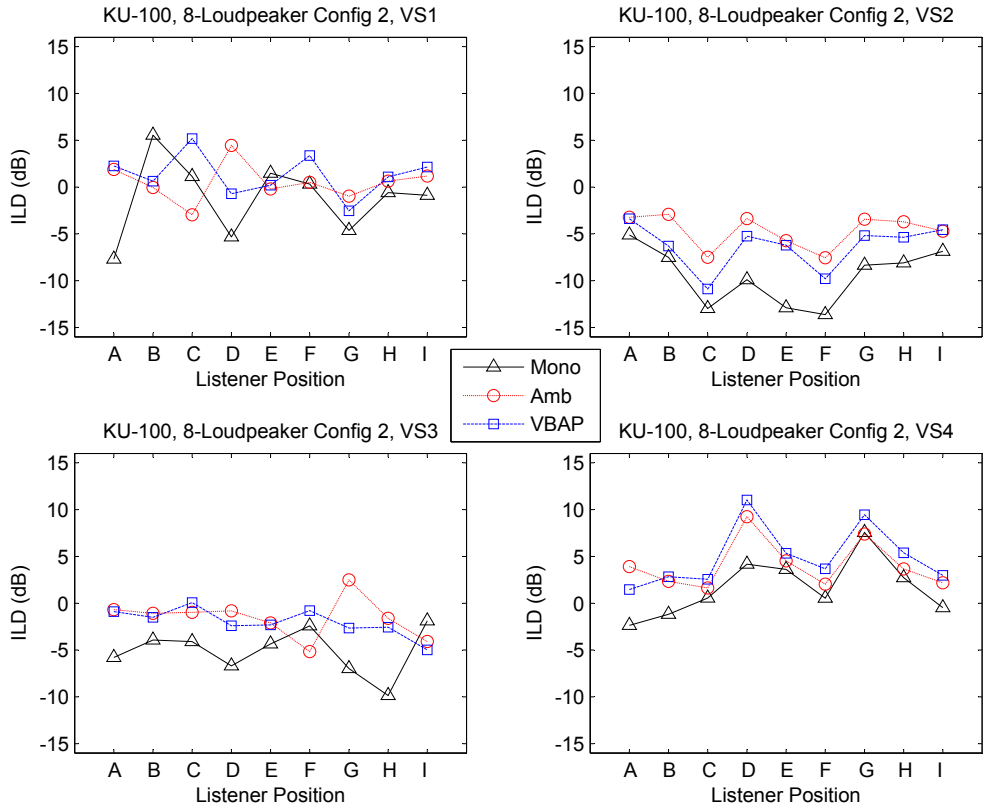


Figure 4.29: *Measured Interaural Level Difference at each listener position for each spatialization method.*

from the true source position. Overall, VBAP does not suffer from large ITD deviations to the same extent as 1st Order Ambisonics does, but the results are far from adequate. The out of phase components in the Ambisonic decode also create localization problems. It is interesting to note that frontal presentations from the direction of VS1 do not exhibit large localization errors, supporting the fact that near-field precedence effects are occurring. However, it is also evident, that there are room effects involved in distorting the ITDs. Such evidence can be seen at listener position B from a presentation intended for the VS4 position. It can be seen that both the VBAP and Ambisonic systems display the exact same angular bias. However, VBAP has no active loudspeakers in the opposed direction of the intended source position. We conclude therefore that significant reflections occur at position C that cause this bias, due to the excitation of multiple loudspeakers in the room.

The high frequency ILDs do not display such large image shifts as those found in the ITDs. These can be seen for each source and listener position in Figure 4.29. In general, the ILD for the spatialization systems follows similar trends to the ILD for monophonic presentations. The performance of 1st Order Ambisonics is comparable to VBAP in this regard, and demonstrates the effectiveness of optimizing Ambisonics for the 'energy decode'. Again, the best localization is

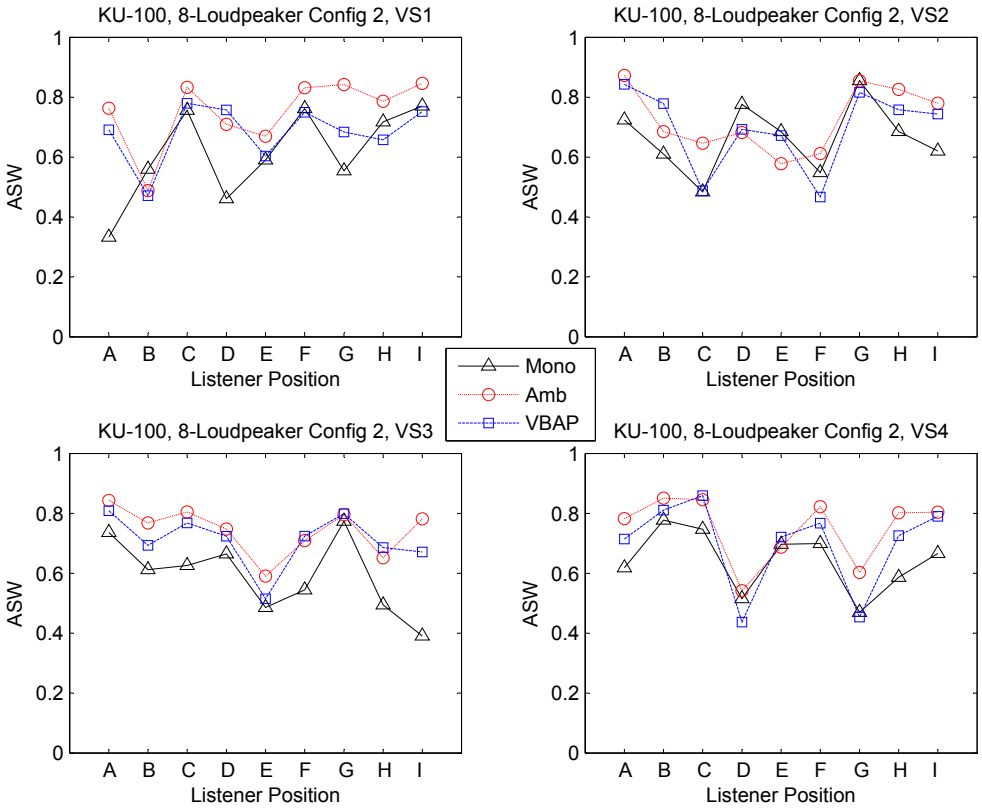


Figure 4.30: *Measured Apparent Source Width at each listener position for each spatialization method.*

achieved at listener position E which confirms both the localization vector analysis and subjective analysis for high frequency spatialization.

The apparent source width at each source and listener position was also calculated from the binaural measurements. The results are shown in Figure 4.30. The monophonic source width is on average approximately 0.6, in good agreement with other standard concert hall environments. Overall, the spatialization systems show similar, but slightly elevated source widths in comparison to the monophonic presentations. This is because the individual loudspeaker contributions do not arrive coherently at off centre listening positions, resulting in a source broadening due to multiple delayed versions of the original source signal at the particular off-centre listening position. This is supported by the fact that the greatest increases in ASW occur at the extremities of the audience area, for example at positions A and I.

An equivalent acoustically modeled environment was implemented using the EASE [Renkus-Heinz, 2009] acoustic simulation tools in order to further study the influence of room effects on the systems. The constructed 3-D acoustic model, shown in Figure 4.31, represents the absorption and reverberation characteristics of the real hall, and its implementation and verification is documented in Appendix A. Impulse response measurements were taken at particular points in

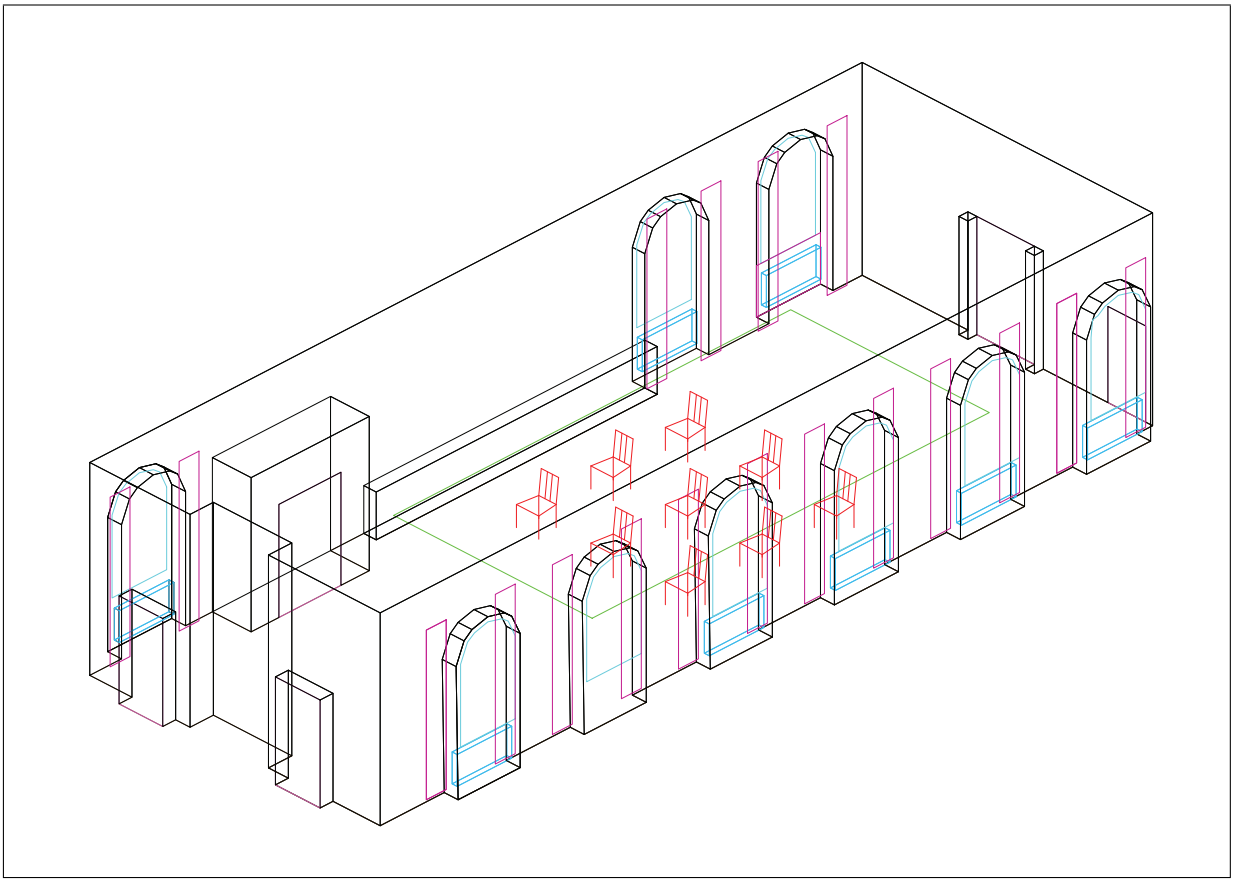


Figure 4.31: *3-D model of Test environment 1.*

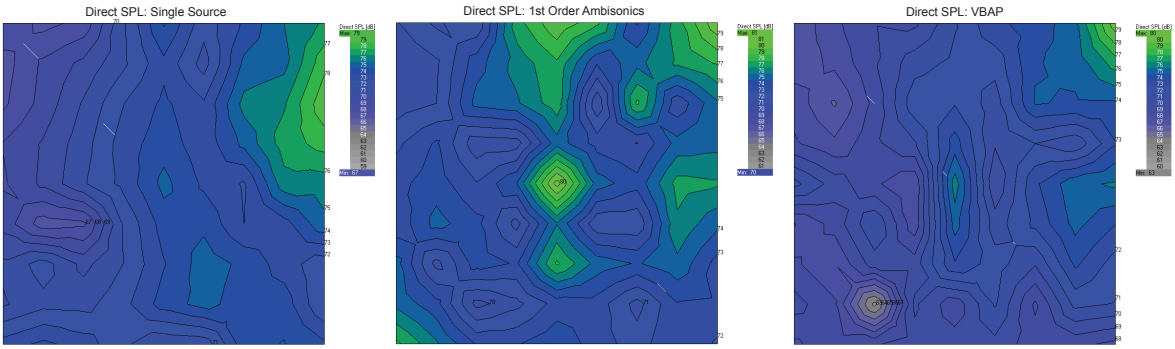


Figure 4.32: *Sound Pressure level after 30ms for monophonic, 1st Order Ambisonic and VBAP presentations from front source position (Loudspeaker 2).*

the hall to verify the accuracy of the model. The measured pressure levels and arrival times of the direct sound and early reflections were found to be comparable to the simulations in EASE.

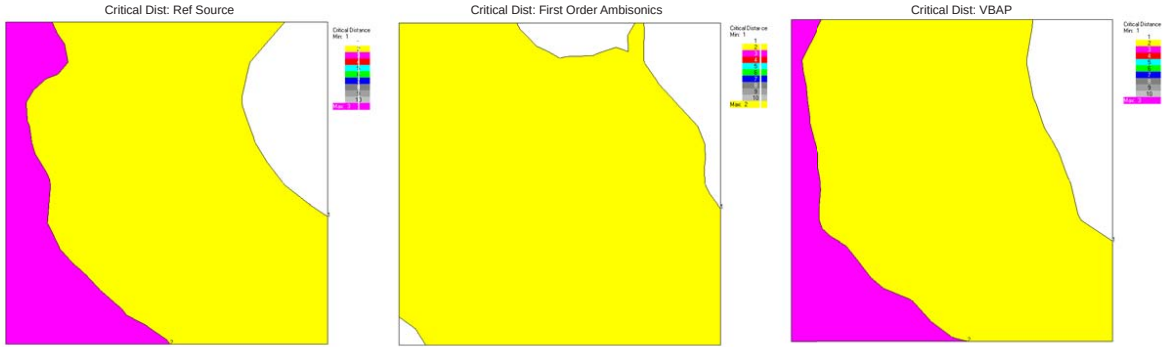


Figure 4.33: *Ratio of direct to reverberant sound for monophonic, 1st Order Ambisonic and VBAP presentations from front source position (Loudspeaker 2) White = 1:1, Yellow = 1:2, Pink = 1:3.*

The direct SPL contributions due to the formation of a virtual source at loudspeaker 2 was calculated in the model for both 1st Order Ambisonics and VBAP, and the results compared to the monophonic presentation shown in Figure 4.32. These graphs have the same orientation as the configuration plot in Figure 4.24.

We can see from the SPLs of both VBAP and Ambisonics that the region at the central listening position receives the highest SPL as expected. Good coverage for a sound reinforcement system should typically be in the region of ±3dB, which is not achieved with the current configuration of these systems. Furthermore, the existence of the large pressure null around the area of seat I in the VBAP SPL plot shows how rear reinforcement may be compromised. We also see that the differential between the maximum SPL seat and the surrounding seats is greatest for 1st Order Ambisonics. Both systems show small areas where the SPL drops significantly (more than 10dB). This can be attributed to the superposition of anti-phase loudspeakers signals as well as excitation of the room modes. Figure 4.33 shows the areas where the direct to reverberant ratio due to the real/virtual source at loudspeaker position 2 is approximately 1:1 (white area indicating direct field), 1:2 (yellow area past the critical distance) and 1:3 (pink area showing complete diffuse field). We see that VBAP behaves similarly to the mono source in this case. This is not surprising, since VBAP is using only two loudspeakers in the direction of the intended source position, in comparison to 1st Order Ambisonics which has contributions from all loudspeakers. A consequence of this is that although there is potential for greater localization inaccuracy, virtual source intelligibility has the potential to be better with 1st Order Ambisonics than with both VBAP or the true source. It is certainly noteworthy that the sweet spot is also in the reverberant field, and therefore subject to localization altering reflections. We can therefore conclude from these simulations that the large SPL fluctuations as well as the direct to reverberant ratio about the audience area are further factors which contribute to the inaccurate localization of sources throughout the listening area. It is clear that each of the given systems

are by no means immune to room effects.

4.2.5 Subjective Listening Experiment II.B: Coloration of Moving Phantom Sources using VBAP and 1st Order Ambisonics under Real Listening Conditions

It is well known that as VBAP sources are panned that the spectral differences due to the comb filtering effects result in audible timbre changes. A good overview of the effect of this phenomena can be found in [Pulkki, 2001a]. A solution to this coloration for VBAP is also presented in this paper, where it is suggested that for moving sources the virtual source should be spread amongst the neighboring speakers, reducing the perceived coloration. However, we will not consider the operation of VBAP in this manner since our aim is to have interactive VAEs where the listener determines when the source should be stationary, and when it should be panned (moved) according to their movements in the interactive VAE. The change between focused sources for stationary sound objects and blurred sources for moving sound objects is not desirable. Furthermore, we have already seen how the localization blur changes as VBAP sources are panned, even at the centre listening position. Audible comb-filtering effects are reduced in 1st Order Ambisonics since there are contributions from a larger number of loudspeakers than with VBAP. What has not been studied however, is whether these differences lead to a subjective preference in decoding schemes in a listening environment which has strong lateral reflections and reverberance. Thus, in a supplementary experiment, listeners were asked to rate their preference for moving sources using both VBAP and 1st Order Ambisonics with energy decode. The hypothesis of this study is as follows:

> There is no listener preference between moving phantom sources generated using VBAP and Ambisonics under real listening conditions in reverberant rooms.

4.2.5.1 Experimental Procedure

In this study a population of 10 experienced listeners was used, and once again each was of good hearing and under 35 years of age. The preference tests were of a A/B paired comparison type. In each test, one of the samples was the VBAP rendered source and the other stimulus was the Ambisonics source. Randomization was used to reduce potential order effects in the testing, as well as to reduce listener fatigue. Three speeds of panning were used to move the sources around the array, 0.033Hz (30 second revolution), 0.066Hz (15 second revolution) and 0.2Hz (5 second revolution). The reader should note that simulated Doppler shift, such as defined in [Chowning, 1971], was not included here.

The subject was asked to identify which test stimulus they preferred, A or B. It was also desirable that the perceptual preference was graded. Both objectives were achieved simultaneously using a preference rating test, where the magnitude of the rating indicates the degree of which the subject likes or dislikes the sound quality of both samples according to an 11-point scale.

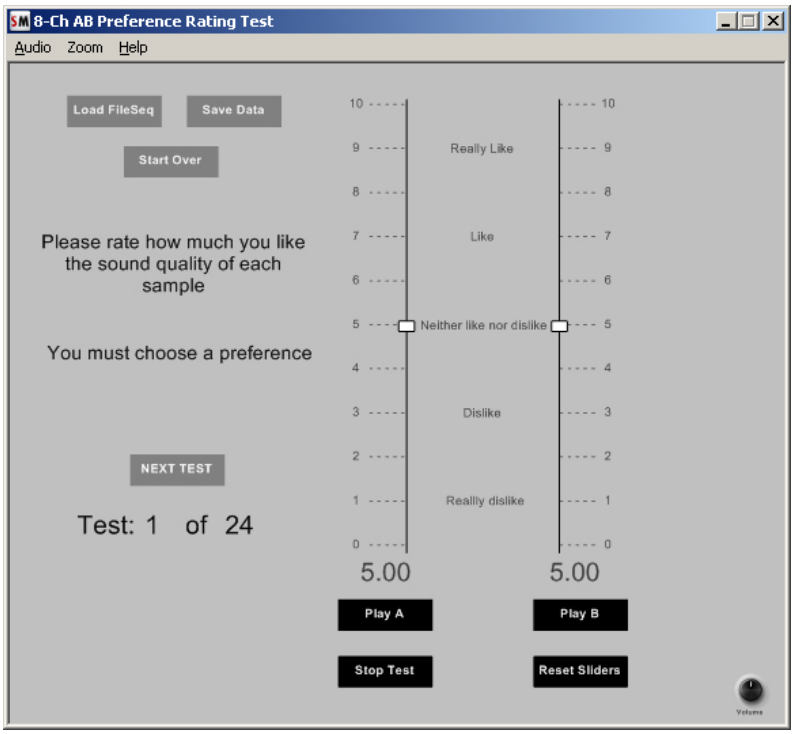


Figure 4.34: *Custom test software designed for subjective preference ratings.*

The scales were implemented using two continuous sliders in the test software, shown in Figure 4.34, and the scores were truncated to 1 decimal place. For each test round the subject was instructed not to give tied scores. If this occurred, the software would prompt the subject to re-enter their scores. This forced choice method ensures that the subject attempted to distinguish between the two stimuli. Subjects are given the following guidelines regarding preference:

- Slight preference: Difference between A and B is at least 0.5
- Moderate preference: Difference between A and B is at least 1
- Strong preference: Difference between A and B is at least 2

These definitions encouraged the subjects to use the scale similarly, such that equal distances of separation between the samples implies the same preference meaning, no matter what part of the scale is used.

4.2.5.2 Results

Two source types were analyzed: female speech and music. The subjective quality ratings were collected for each both stimuli, using both Ambisonics and VBAP at listener positions A, E and I and for each panning speed. Each round was repeated 4 times during the course of each listener trial. For each system, stimuli, position and panning speed we have computed an average

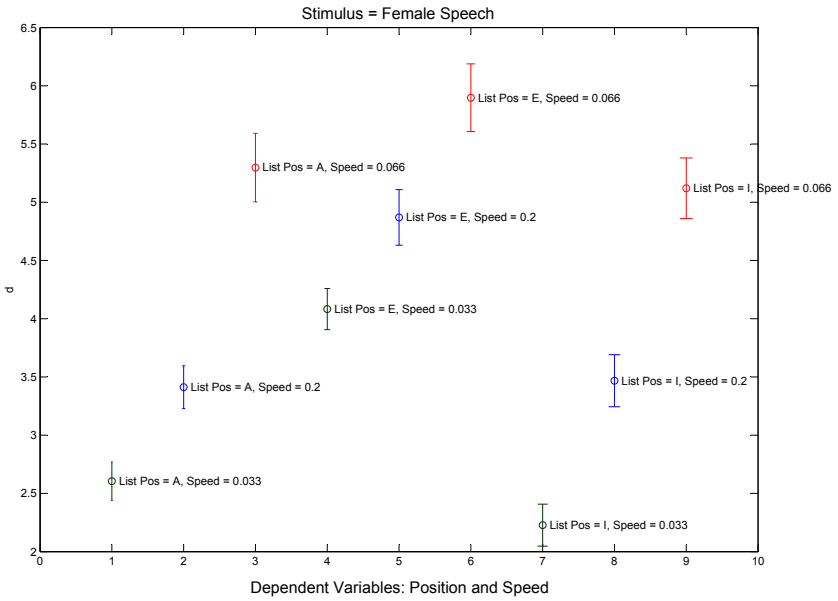


Figure 4.35: *Difference in listener preference between VBAP and 1st Order Ambisonics for moving virtual sources, Stimulus = Female Speech. Positive values of d indicate a preference for Ambisonics.*

---

preference rating $\bar{\alpha}$ (over the 10 subject answers). We then compute the subjective difference in between the VBAP and Ambisonic presentations and the corresponding standard error $se(\bar{\alpha})$. For example, if we wish to consider the subjective preference at the sweet spot for female speech with a panning rate of 0.2Hz, we have

$$\begin{aligned} d \quad = \quad & \bar{\alpha}(\text{system} = \text{`Amb: Energy Decode'}, \text{sound} = \text{Female Speech} , \text{position} = \text{E}, \text{speed} = 0.2) \\ & -\bar{\alpha}(\text{system} = \text{`VBAP'}, \text{sound} = \text{Female Speech} , \text{position} = \text{E}, \text{speed} = 0.2) \end{aligned} \tag{4.10}$$

The results are plotted in Figures 4.35 and 4.36 for each listener position and each panning speed.

In both cases, there is a positive bias in the results clearly showing that Ambisonics is the preferred method for panning virtual sources. It is interesting to note that the speed of panning has a direct effect on this perception, as does off centre listening position. In the female speech case, the largest differences are exhibited at the sweet spot for all panning speeds. The greatest difference occurs with the moderate panning speeds and comparable results are found for music. The greatest noticeable change in timbre with VBAP is when the source 'pans-through' a loudspeaker, i.e. transition from virtual source to mono source and back to virtual source again. At the fast and slow speeds it becomes more difficult to gauge the subjective difference since, on one hand the spectral changes due to comb-filtering effects are too slow for perception of changing timbre, and on the other, changes occur almost too quickly when interacting with the room reverberation. This effect is particularly noticeable at regular intervals (eight times

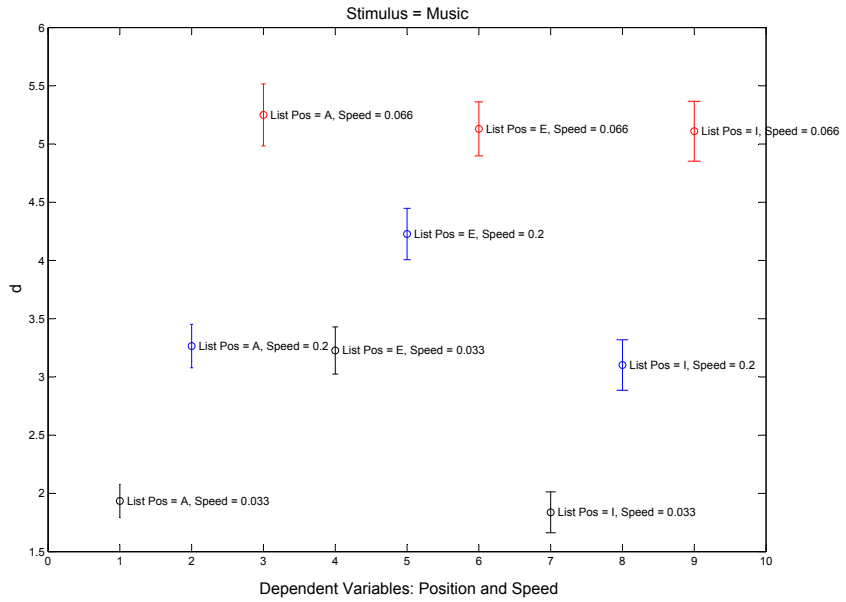


Figure 4.36: *Difference in listener preference between VBAP and 1st Order Ambisonics for moving virtual sources, Stimulus = Music. Positive values of d indicate a preference for Ambisonics.*

---

the panning speed) at the sweet spot. The interaction of the phantom sources with the room plays an important part in this perception. In a highly reverberant environment it is conceivable that the effect of phantom source coloration is lessened due to increased localization blur due to lateral reflections. The differences for off centre listening are somewhat lower than in the sweet spot case. This is partly due to the fact that for a significant portion of a panning cycle the virtual source is further into the diffuse field than in the sweet spot case. When the source moves near the listener, it is more in the direct-field. However, precedence effects in the direct-field lower the overall subjective rating of Ambisonics.

4.2.6 Conclusions from VBAP and 1st Order Ambisonics Analysis

We can conclude the following in response to the hypotheses posed in Section 4.2.1:

1. The localization accuracy of amplitude panned virtual sources is not equivalent to that of monophonic source listening in a real (reverberant) listening environment.

2. The localization accuracy of 1st Order Ambisonic virtual sources is not equivalent to that of monophonic source listening in a real (reverberant) listening environment.

3. The localization accuracy of 1st Order Ambisonic virtual sources at the sweet spot is not greater than that of VBAP virtual sources, when compared to real sources in a real

listening environment.

4. That the localization accuracy of 1st Order Ambisonic virtual sources is not greater than that of VBAP virtual sources for off-centre listening positions in a real listening environment.

Overall, the results show that source localization for non-central listener positions is consistently biased away from the intended image position, irrespective of the spatialization technique or the nature of the source stimulus. Both systems suffer from near-speaker effects which result in a strong localization bias towards the near loudspeakers. However, due to the smaller number of contributing loudspeakers in VBAP, biases towards the nearest contributing loudspeaker do not affect the overall localization performance to the same degree. 1st Order Ambisonics however has more contributing loudspeakers and so off centre localization is severely affected. Of the three main types of Ambisonic decoding the 'Energy' decode provided the best localization accuracy. The localization accuracy of VBAP is directly related to the maximum inter-loudspeaker angle. This is not the case for Ambisonics however, which shows stronger localization errors and greater localization blur. We therefore conclude that for stationary source localization VBAP is a better solution than 1st Order Ambisonics.

However, the coloration of moving phantom sources directly effects the subjective quality of these systems and the preference is in favor of Ambisonics. It is therefore desirable for an effective VAE solution to achieve the localization accuracy of VBAP with the more 'natural' sounding reproduction of Ambisonics.

4.3 Higher Order Ambisonic Decoders for Virtual Auditory Environments

Although VBAP gives a better localization performance than 1st Order Ambisonics, it is possible to improve the localization accuracy of Ambisonic reproduction by increasing the reproduction order, as well as optimizing the decoder. In this section, we will therefore investigate the effectiveness of higher order Ambisonic decoding in terms of localization accuracy in a real listening room. It is important to note that we are not abandoning VBAP as a potential solution to spatial audio reproduction in a VAE. On the contrary, we use it here as a benchmark for localization performance in Ambisonic systems, i.e. for any higher order Ambisonic system to give an equivalent or better localization performance than VBAP, the localization error must be limited to the inter-loudspeaker aperture. In this investigation, we will focus more on the perceptual differences between the different Ambisonic decoder types and orders, rather than the effect of room acoustics. We approach this study with the following hypotheses:

1. That the localization accuracy of Ambisonic systems increases under real listening conditions for all decoder types with increasing order,

Order	Decoder
1st Order	Velocity Energy In-Phase Shelf Decode
2nd Order	Velocity Energy In-Phase Shelf Decode
3rd Order	Velocity Energy In-Phase Shelf Decode

Table 4.1: *Ambisonic Decoders for HOA analysis.*

2. That optimal localization accuracy for an individual listener over an 8-channel array can be achieved using a psychoacoustically optimized decoder.

3. That optimal localization accuracy for a distributed audience is also achieved by using a psychoacoustically optimized decoder.

We will investigate these hypotheses using the analysis framework of the previous study. Localization vectors, binaural simulation and r_v density maps will be used to explore the fundamental localization limitations with HOA systems. A formal listening test will then be used to explore deviations from the theoretical analysis due to real listening conditions.

4.3.1 Objective Analysis of Higher Order Ambisonic Localization

In the previous analysis we investigated the accuracy of VBAP and Ambisonic systems using non-ideal loudspeaker array configurations. It was shown that localization at the sweet spot can be made equivalent to that of regular arrays through proper loudspeaker gain decoding as well as delayed loudspeaker feeds. Furthermore, the localization at off-centre positions is directly related to both the decoding gains and the inter-loudspeaker aperture. In this section, we will concentrate on regular arrays, bearing in mind the limitations imposed for non-ideal loudspeaker setups. In this way, the results presented here can be directly compared to previous work on HOA, and the optimal performance for the virtual loudspeaker approach can be assessed. Based on the previous analysis of horizontal loudspeaker arrays, an 8-channel regular array configuration, shown in Figure 4.37, is chosen for the analysis. Again, 9 listener positions, labeled A to I are used. Given that we are using 8 loudspeakers, we can assess the localization accuracy of Ambisonics up to 3rd Order. We will therefore present an analysis of the Ambisonic decoders shown in Table 4.1.

4.3.1.1 Sweet Spot Analysis

Let us first investigate the theoretical localization accuracy at the sweet spot (Position E). The velocity and energy vectors for each order and decode are shown in Figure 4.38. In the case of the 'velocity' decode, we see that as we move up in order, the level of the energy vector becomes close to 1. Since the velocity vector is maximized at 1st Order, no significant advantage for low frequency localization is demonstrated. However, a binaural simulation of a 'velocity' decode for

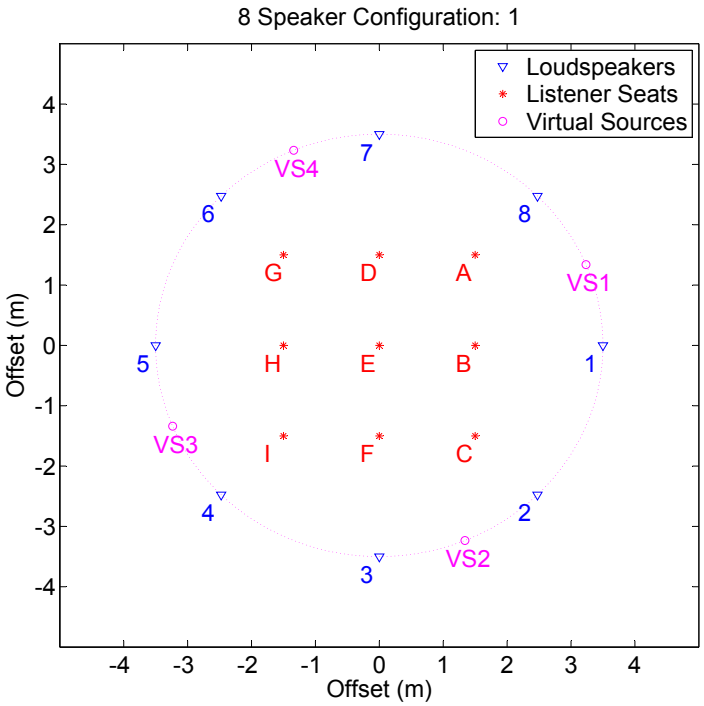


Figure 4.37: *8 loudspeaker regular configuration 1.*

a centrally seated listener reveals that there are marginal improvements in ITD accuracy as we move up in order. Figure 4.39, shows this to be case, and should be compared to the 1st Order 'velocity' decode of Figure 3.26 in Chapter 3. Conversely, with the 'energy' decode, we see that as we move up in order both r_e and r_v increase in value. At 3rd Order both vectors reach a value of approximately 0.92, indicating very good localization accuracy in all directions. The ITD and ILD in Figure 4.40 also show this to be the case, with a good match to the monophonic source at both low and high frequencies. As we might expect, the 'in-phase' decode shows the poorest localization performance, although this does improve significantly as we move up in order, with values of r_v and r_e reaching 0.75 and 0.86 respectively. The resultant ITDs and ILDs shown in Figure 4.41 are smooth, but do not match the monophonic reference source well.

However, the best performance is achieved with a combination of the maximized velocity vector at low frequencies and the maximized energy vector at high frequencies, the fourth row in Figure 4.38. The binaural simulation of Figure 4.42 shows this to be the case also, although the results are highly comparable to the 'energy' decode, illustrating the fact that as we move up in order, the distinction between the different decoder types gets smaller. From this analysis we can conclude that optimal localization for an 8-channel centre listener position in an Ambisonic decode is achieved using 3rd Order max-velocity and 'energy' decode, implemented with shelf filtering.

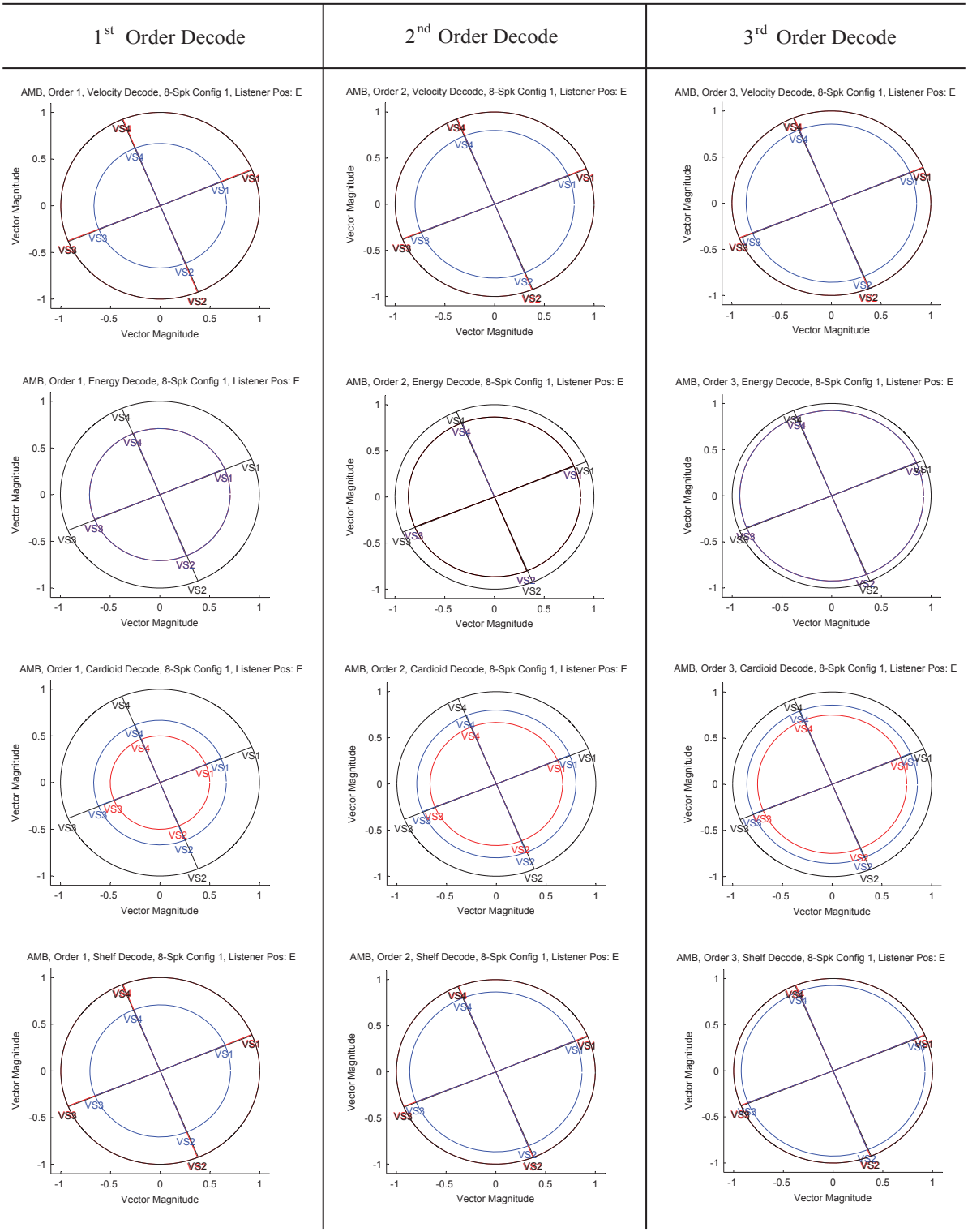
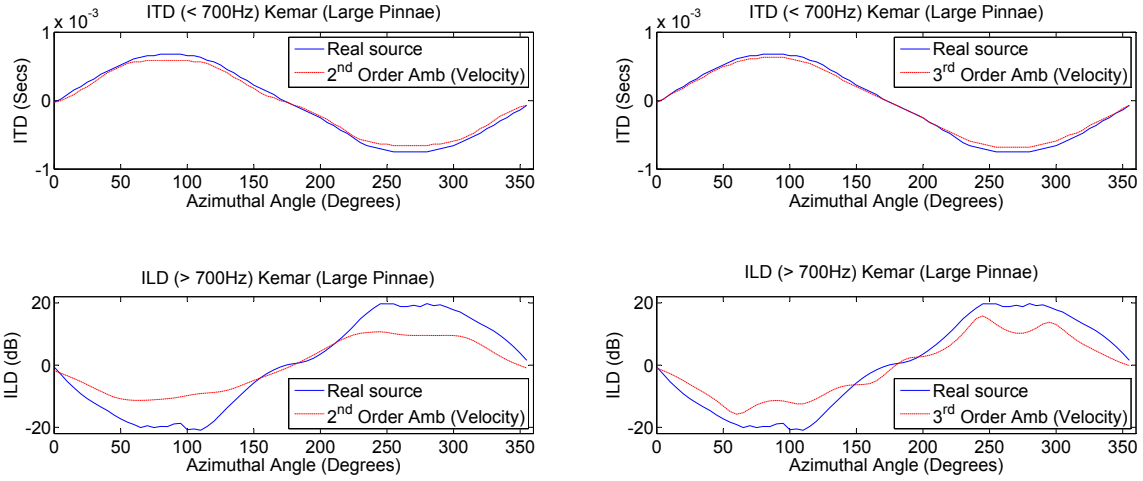
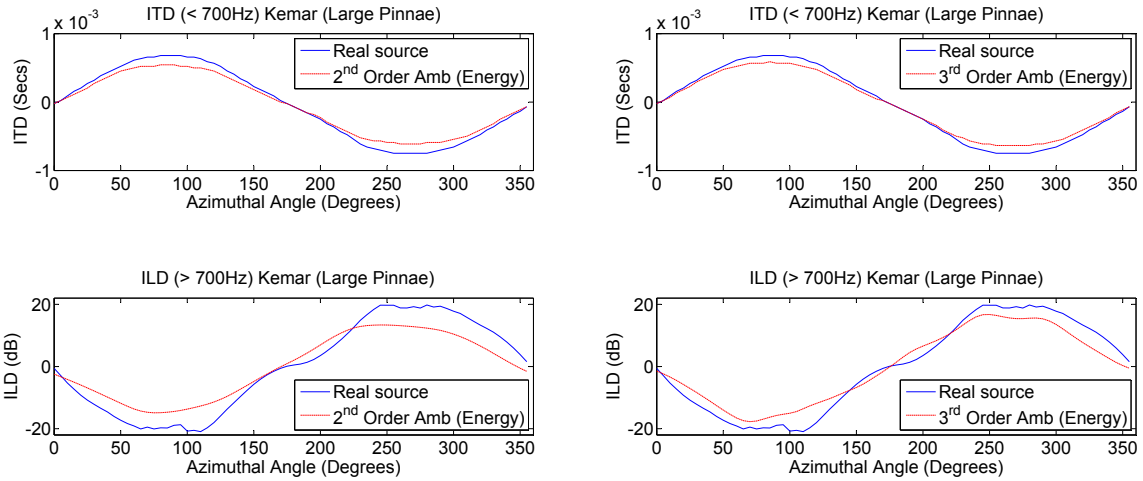


Figure 4.38: *Localization vectors for 1st, 2nd and 3rd Order Ambisonics decoders at the sweet spot position.*



2^{nd} Order 3^{rd} Order

Figure 4.39: *ITD and ILD for 2^{nd} and 3^{rd} Order 'velocity' decode.*

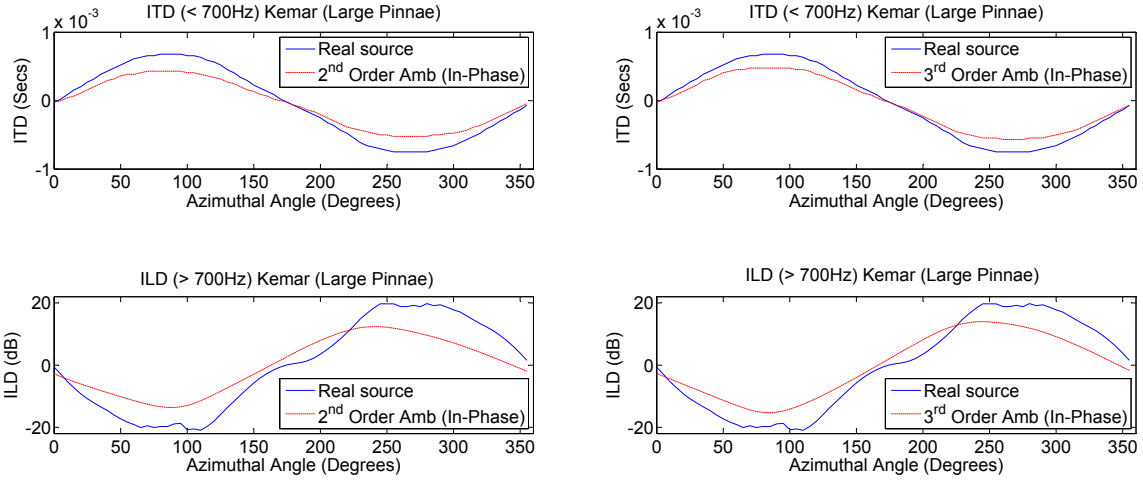


2^{nd} Order 3^{rd} Order

Figure 4.40: *ITD and ILD for 2^{nd} and 3^{rd} Order 'energy' decode.*

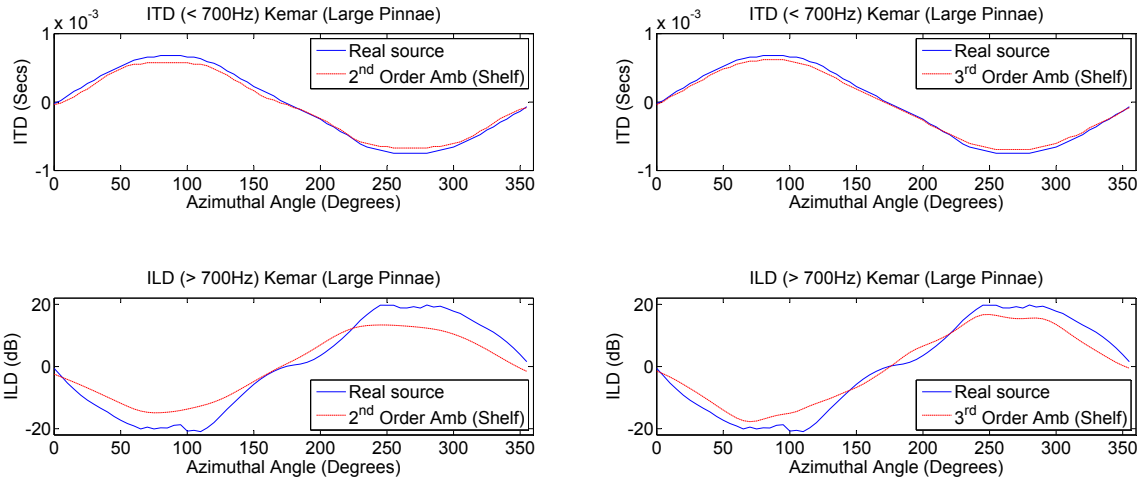
An important consideration in the implementation of shelf filters for a 3^{rd} Order Ambisonic decode is that since higher order Ambisonics is considered asymptotically holographic, then the crossover frequency at which to separate the velocity and energy decoding changes with the Ambisonic order. This is because, as has been shown by Daniel, higher orders give more correct soundfield reconstruction with respect to kr [Daniel et al., 1998]. This reconstruction accuracy can be measured by using the integrated wavefront error, known as the D-Error, introduced by Bamford [Bamford and Vanderkooy, 1995], and given as

$$D\text{-Error}(\mathbf{r}, \psi) = \frac{1}{2\pi P_\psi} \int_0^{2\pi} |p_\psi(\mathbf{r}(\theta)) - p_r(\mathbf{r}(\theta))| \, d\theta \tag{4.11}$$



2nd Order 3rd Order

Figure 4.41: *ITD and ILD for 2^{nd} and 3^{rd} Order 'in-phase' decode.*



2nd Order 3rd Order

Figure 4.42: *ITD and ILD for 2^{nd} and 3^{rd} Order 'shelf' decode.*

---

where p_ψ is the spherical harmonic representation of the plane wave calculated from equation 3.56, p_r is the reproduced wave, P_ψ is the pressure of the plane wave, $\mathbf{r}(\theta) = r.\mathbf{u}_\theta$, the point of measurement and ψ is the plane wave angle. From this measure, we can determine the upper frequency limit with which the soundfield can be reconstructed accurately, within a specific tolerance. Table 4.2 shows these frequencies in relation to 20%, 14%, and 10% error. One should also note that as we move higher in order, the distinction between each of the decoder types gets smaller, and we will find that the need for psychoacoustic optimization is not required at a central listening position.

Tolerated Error Vs Order	1	2	3
20% (-14dB) Order	742.9Hz	1345.8Hz	1059.5Hz
14% (-17dB) Order	613.7Hz	1173.6Hz	1755.0Hz
10% (-20dB) Order	516.8Hz	1033.6Hz	1593.5Hz

Table 4.2: *Upper frequency limit for accurate soundfield construction on an 8.5cm radius area in relation to the D-Error.*

4.3.1.2 Off-Centre Localization

The full set of vectors for all listener positions in the array can be found in Appendix E. Here we will once again concentrate mainly on listener position F for illustrative purposes. The energy vectors for off-centre position F are shown in Figure 4.43. Beginning with the 'velocity' decode, we see that the large localization errors of 1st Order are significantly reduced in higher orders. For example, this is clearly seen with source positions VS2 and VS3 in the 2nd and 3nd Order decode. The localization accuracy of r_e in the 'velocity' decode, increases with order, just as was the case at the sweet spot. However, the localization is not smooth for all angles, and is greatest in the directions of the loudspeakers. The resultant localization blur for each order is shown in Figure 4.44. Once again, the only position which holds a constant localization blur is the sweet spot. Furthermore, although the overall localization blur decreases, the periodicity of the blur, that is, how much it fluctuates around a mean blur value, increases with the order. The minimum blur is with the 3^{rd} Order decode, as expected.

The probability density maps for the 'velocity' decoders are shown in Figure 4.45. As we have seen before for the 1st Order decode, high density regions gather in the directions of the loudspeakers, indicating poor localization. This is reduced somewhat in the 2nd and 3rd Order decodes, but the effect of the negative gains on off centre listening is still apparent as a spread around the direction of localization, with values of r_v ranging from 0.25 to 1. It is also interesting to note that as we move higher in order the highest density region moves from being localized in the direction of a single loudspeaker to being spread between the two loudspeakers encompassing the direction of localization.

The vectors for the 'Energy' decode with increasing order are also presented in Figure 4.43. Here, r_e is less than 1 in all cases. However, it is quite close to 1 for nearby sources and obtains good and smooth maximization of both at 3rd Order. It is noteworthy that the localization of each of the four virtual source directions indicated is always between two loudspeakers for the 3rd Order decode. The associated localization blur is shown in Figure 4.46. Once again, the localization accuracy increases with Ambisonic Order. However, for the 3rd Order decode, the localization blur is now less than 45° at each of the listener positions, again indicating that localization will be guaranteed between the two loudspeakers on either side of the source

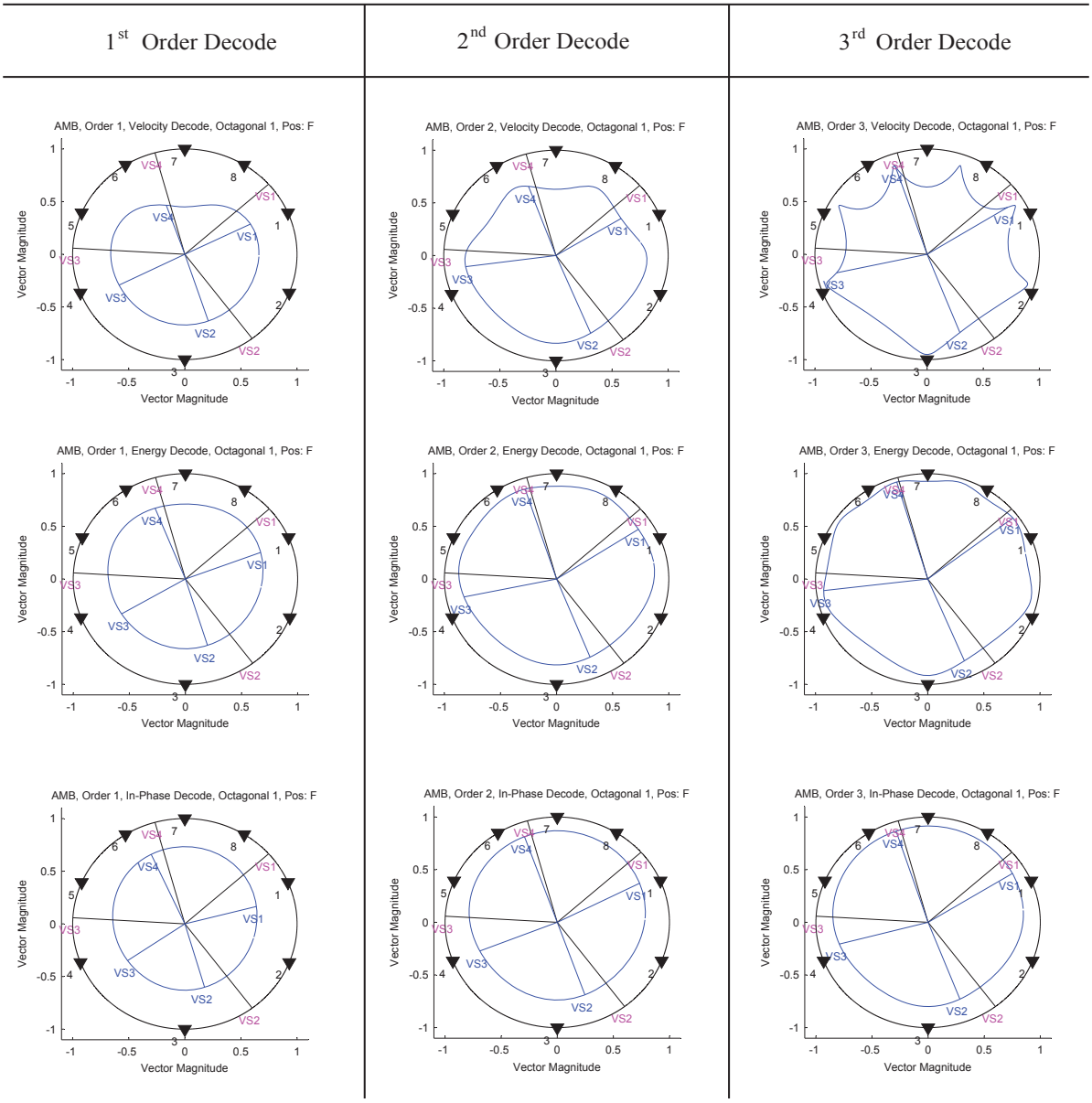


Figure 4.43: *Localization vectors for 1st, 2nd and 3rd Order Ambisonics decoders for off centre listening position F.*

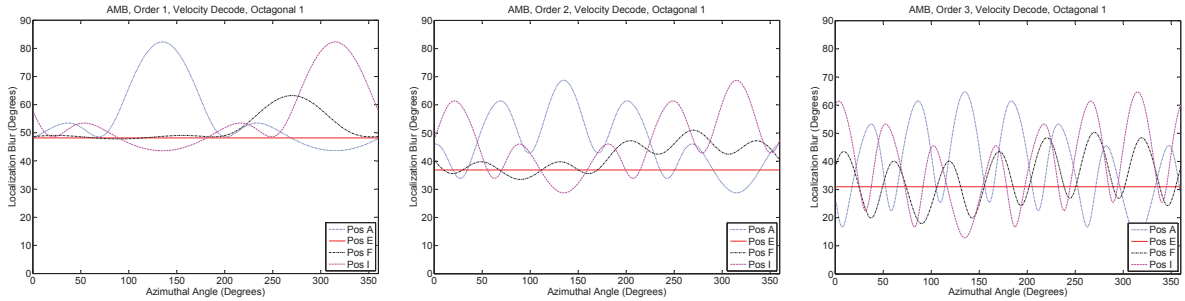


Figure 4.44: *Localization blur for HOA Velocity Decode: 1st, 2nd and 3rd Order.*

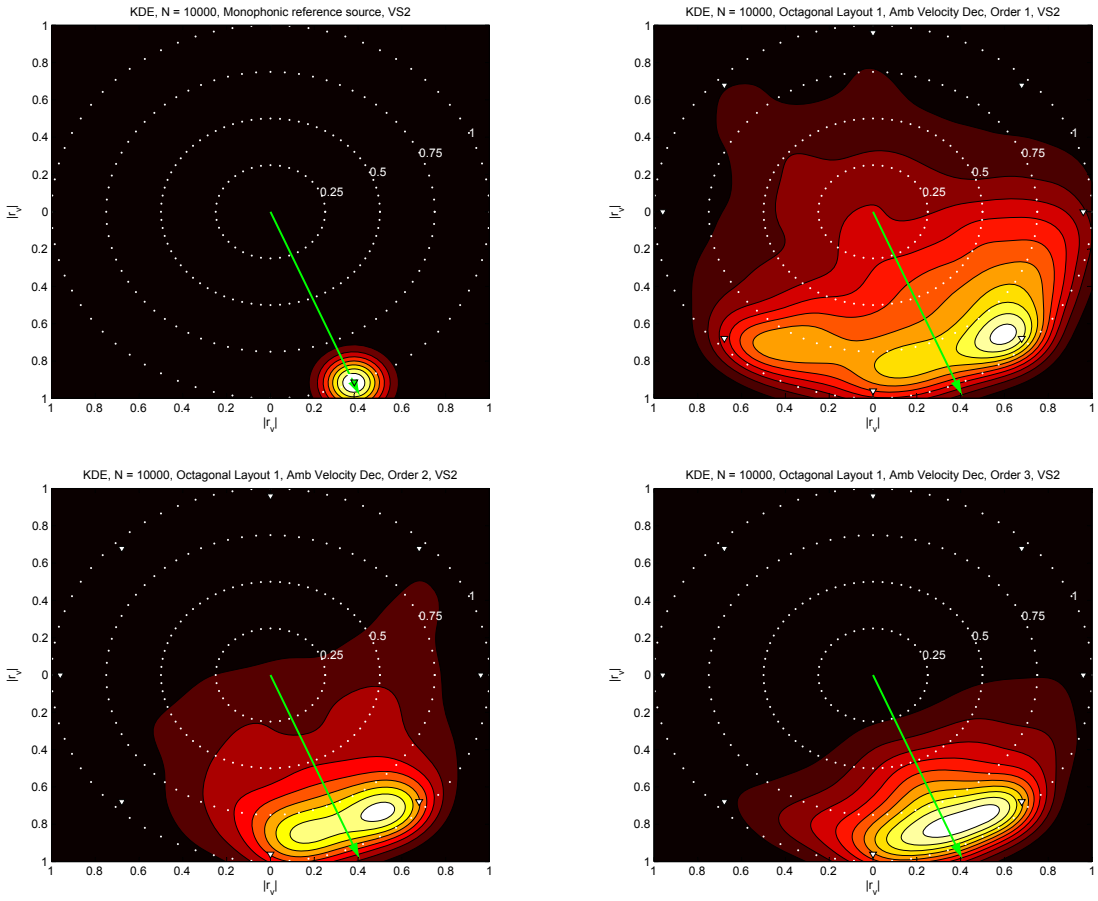


Figure 4.45: *Kernel density estimate of the probability density function for r_v with HOA Velocity Decode: Monophonic reference, 1^{st}, 2^{nd} and 3^{rd} Order. The green arrow represents the intended source direction.*

---

position. The probability density maps also confirm these observations, as shown in Figure 4.47.

Significant improvements are made in the 2^{nd} and 3^{rd} Order decodes. For the 3^{rd} Order decode, localization accuracy is to within the two loudspeakers ($\pm 22.5°$). This is an important result, and shows that from the perspective of a distributed audience, the statistical behavior of the velocity vector in 3^{rd} Order Ambisonics is approximately equivalent to that of VBAP.

The 'in-phase decode' vectors, also shown in Figure 4.43, exhibit the smoothest transition with angle over each order. This is of course due to the spread of the energy over several loudspeakers in the direction of the intended source position. The magnitude of the vectors increases with order, but overall, the localization performance of this decoder in terms of r_e is not ideal. The localization blur plots for the 'in-phase' decode are shown in Figure 4.48, and they demonstrate smoother transitions and smaller blur deviations than with the velocity decode, but the magnitude of this blur is greater than that of the 'energy decode'. The probability density function for r_v depicted in Figure 4.49 also show this to be the case. The density of r_v becomes

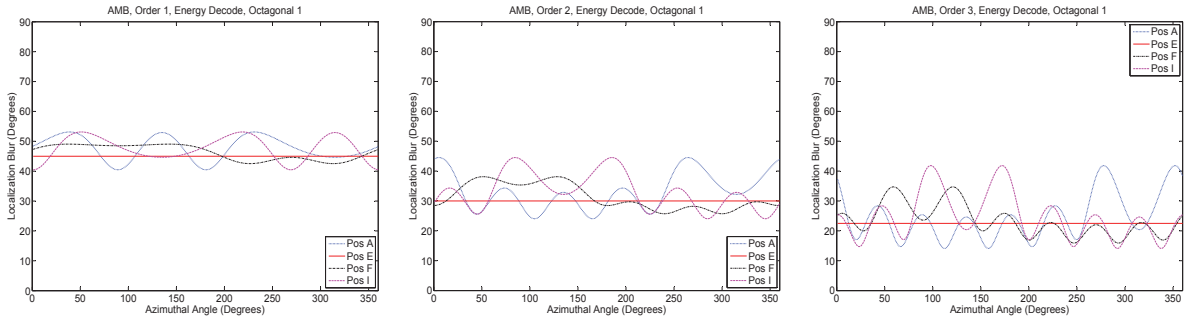


Figure 4.46: *Localization blur for HOA Energy Decode: 1^{st}, 2^{nd} and 3^{rd} Order.*

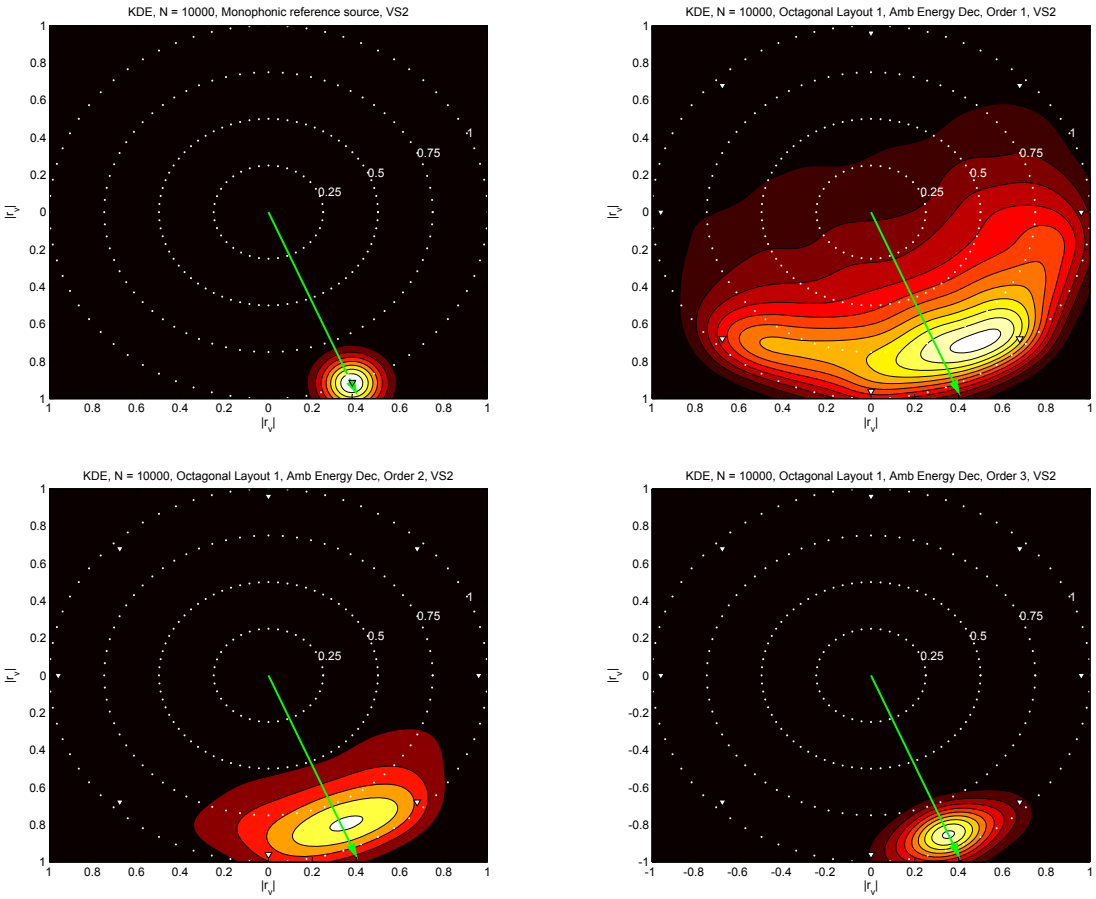


Figure 4.47: *Kernel density estimate of the probability density function for r_v with HOA Energy Decode: Monophonic reference, 1^{st}, 2^{nd} and 3^{rd} Order. The green arrow represents the intended source direction.*

higher with each order, but does not achieve the accuracy of the energy decode.

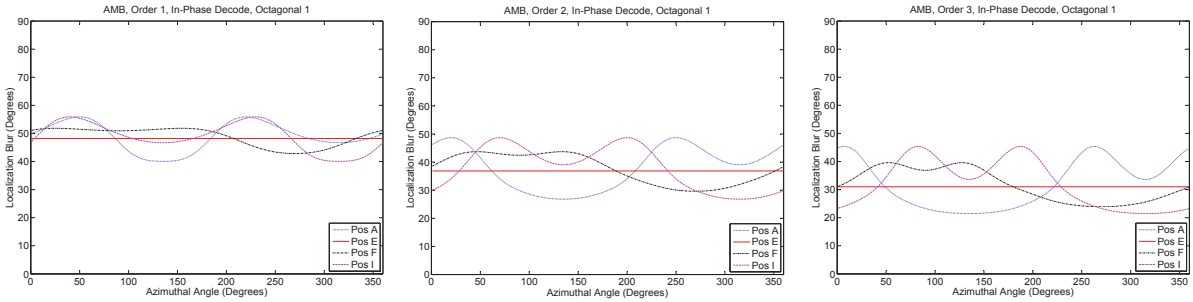


Figure 4.48: *Localization blur for HOA In-Phase Decode: 1st, 2nd and 3rd Order.*

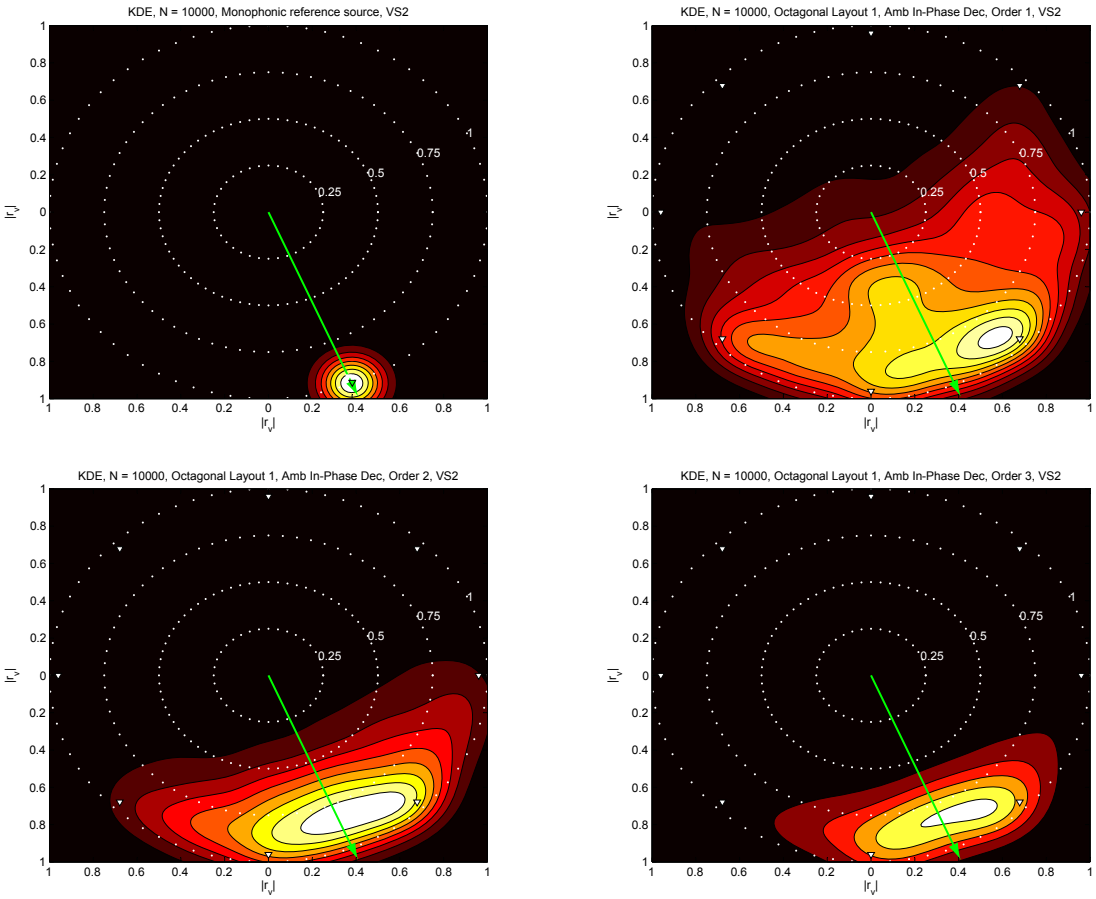


Figure 4.49: *Kernel density estimate of the probability density function for r_v with HOA In-Phase Decode: Monophonic reference, 1st, 2nd and 3rd Order. The green arrow represents the intended source direction.*

4.3.1.3 Comments on Objective Analysis of HOA decoders

From the analysis presented we can conclude the following for both VBAP and 1st Order Ambisonic systems:

1. The localization accuracy of Ambisonics increases for higher orders over the listening area under ideal listening conditions.

2. The optimal localization accuracy for a distributed audience under ideal listening conditions for an 8-channel array is obtained with a 3rd Order Ambisonic 'energy' decode.

3. A 3rd Order Ambisonic 'energy' decode gives equivalent localization performance to VBAP for a distributed audience, over an 8-channel array and under ideal listening conditions.

Based on this analysis, we will now investigate the true subjective performance of HOA decoders in a real listening environment in order to verify the objective results.

4.3.2 Subjective Listening Experiment III: Analysis of Higher Order Ambisonic Localization

A subjective experiment was designed to assess the localization accuracy of the different Ambisonic decoders at higher orders. The primary objectives of this study are to determine how well each of the decoders produces easily localizable phantom sources in a real listening scenario at both the sweet spot and off-centre positions. There are significant differences between the nature of this study and that of the Ambisonics and VBAP comparison presented in Section 4.2. Prior to Experiment 1, the analysis revealed large differences between the localization accuracy of 1st Order Ambisonics and VBAP. For this reason, a forced-choice speaker identification method was deemed acceptable, provided a large amount of listener-data was captured. Here, the objective analysis reveals that the differences in localization between the decoder types and orders are more subtle, and so a different approach must be taken.

The setup for the tests is shown in Figure 4.50. It again consists of 16 loudspeakers, 8 of which are used in the Ambisonic decodes, and 8 of which are used to generate reference monophonic sources. Of the sources generated, we are primarily interested in those from the direction VS1, VS2 VS3 and VS4 (shown in pink) as before. The listening test environment was a medium sized classroom in the Department of Electronic and Electrical Engineering in Trinity College Dublin, which we will term Test Environment 2. The room is approximately 7m x 5.5m x 2.5m with a false ceiling, and displays similar acoustical properties to listening environments such as living rooms or conference rooms, The spatially averaged reverberation time is shown in Figure 4.51, and is approximately 0.35s at 1kHz.

The loudspeakers used were once again, Genelec 1029a models and were positioned 1.2m from floor to tweeter height. Each speaker was placed at 22.5° increments over a circle encompassing the listener area of radius 2m. Acoustically absorbent baffles were placed between the loudspeakers closest to the walls and the walls themselves in order to reduce 1st Order reflection contributions from these directions. Each loudspeaker was calibrated to 80dBA using a real-time analyzer at the centre listening position. Unlike subjective experiment 1, the loudspeakers were

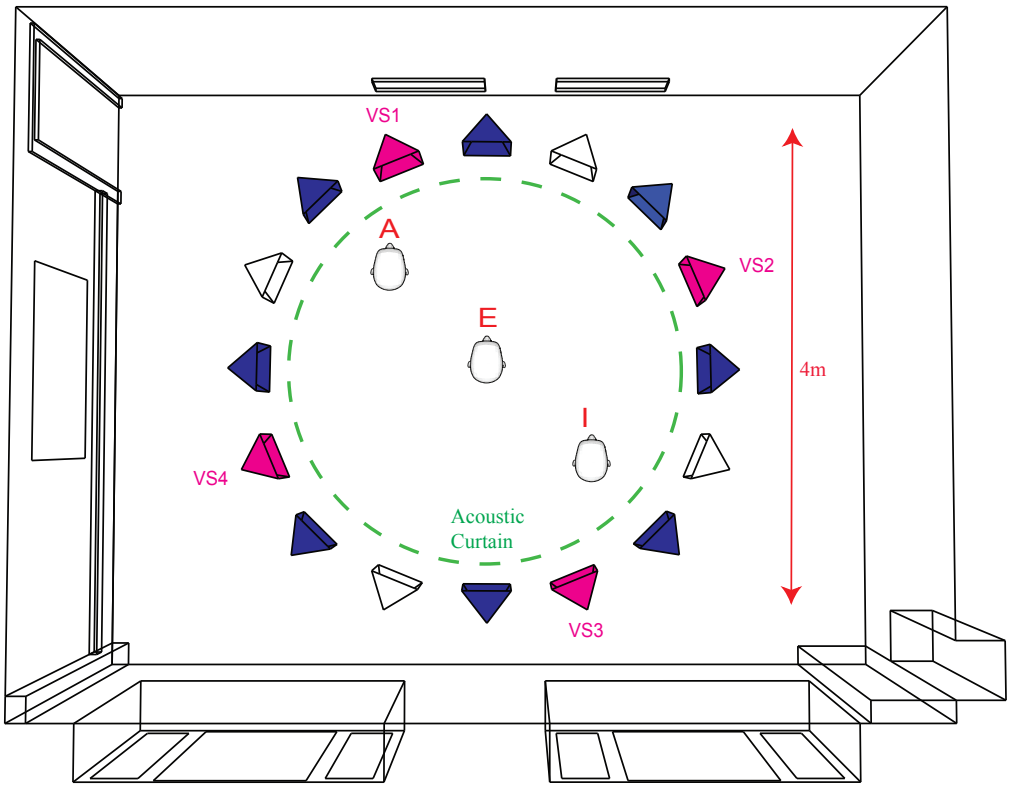


Figure 4.50: *Listening Room 2 loudspeaker configuration.*

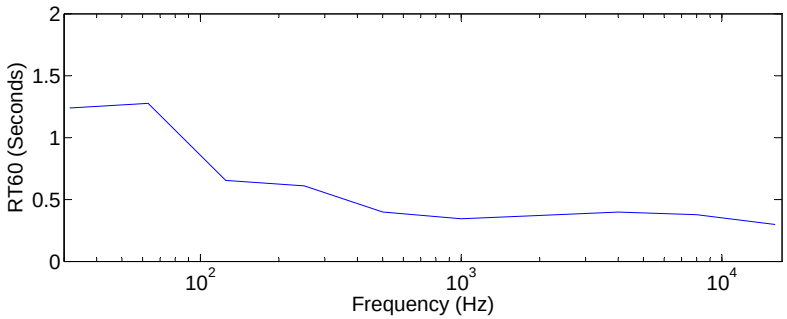


Figure 4.51: *Spatially averaged reverberation time of test environment 2.*

hidden by an acoustically transparent screen, and only guide markers were placed at different angles. Listener positions A, E and I were used in this study.

Two types of stimuli were chosen for the tests: pink noise bursts and female speech. The pink noise bursts consisted of 4 rapid bursts each of 50ms in length, separated by 130ms of silence. The female speech samples consisted of phonetically balanced phrases selected from the TIMIT Acoustic-Phonetic Continuous Speech Corpus database, and re-recorded at higher resolution (96kHz) [Fisher et al., 1986]. The recordings were made in a controlled studio environment using an AKG-C414 microphone. A compensation filter was applied using the methodology

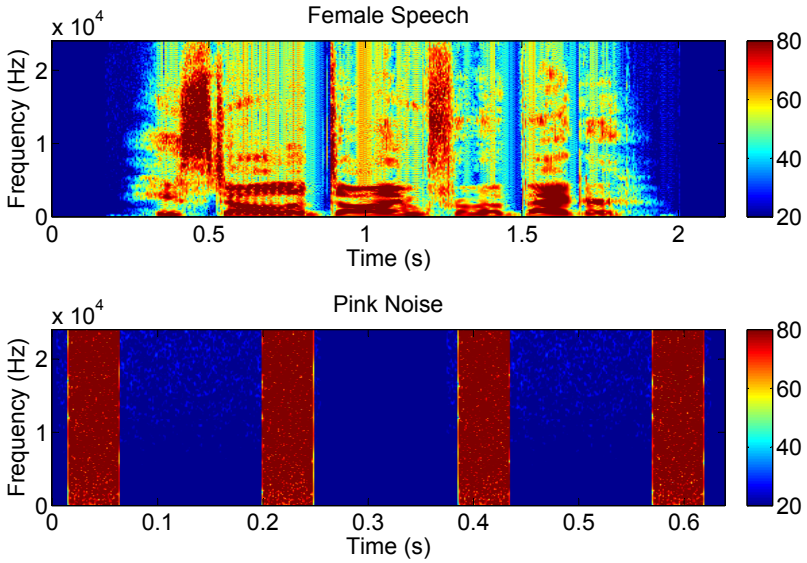


Figure 4.52: *Test data for localization Experiment 2, Top: Female Speech, TIMIT phrase 158, Bottom: Pink noise bursts.*

---

that will be explained in Section 5.3.3 to reduce the coloration of the microphone and the reproduction loudspeakers. The chosen passages are shown in Appendix F. The spectral and temporal characteristics of both types of stimuli are shown in Figure 4.52.

13 subjects participated in the listening tests, each between 18 and 35 years of age and of excellent hearing. The tests were conducted over 3 different sessions for each individual, at the same time of day. At each session, the subject was asked to sit in one of the three listener positions examined. Positions A and I were located 1m from the edge of the array and Position E was at the sweet spot. A dedicated software pointer was developed to allow the listener to choose the direction of each audio presentation. The markers placed on the acoustically transparent screen corresponded exactly with the markers intersecting the perimeter of the circle on the pointer-tool. The three listener positions are also displayed, to allow the subject to orient themselves within the listening area.

Before each test, the subject was trained in the use of the software, and was allowed to become comfortable with the orientation of the markers around the listening area. In total 104 different presentations were made for each subject. The presentations were of the source stimuli pre-rendered for each Ambisonic decoder and source position, as well as monophonic reference sources from each intended virtual source direction. Because of the large number of presentations required, it was decided to perform the tests in three stages per subject, so that listener fatigue was not an issue. The variation in the female speech test signals also reduced listener fatigue. Furthermore, each test presentation was selected from a pre-defined pseudo-random sequence for each listener, so as to remove potential order effects. For each presentation, the listener was allowed to hear the sample as many times as they desired. Once they were satisfied with the

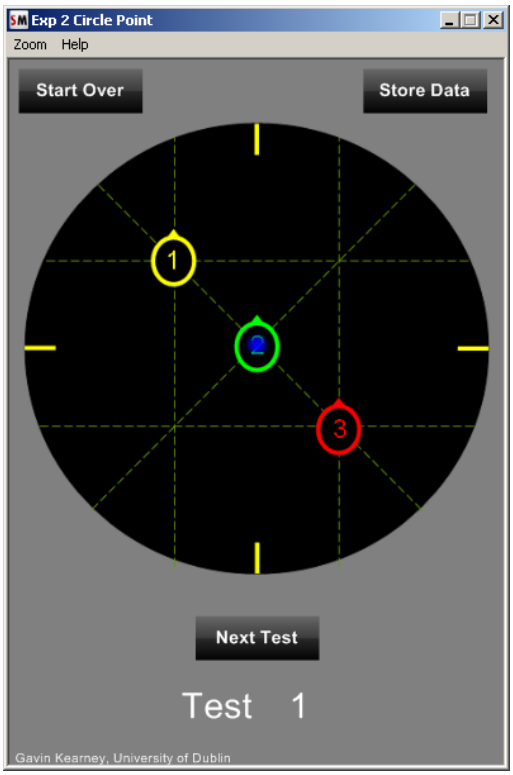


Figure 4.53: *Pointer software developed for subjective experiment 2.*

direction of localization, they moved the blue marker of the test software to the chosen position, and then proceeded to the next test, upon which the software would reset the marker. The size of the marker introduces a $\pm 1.5^{\circ}$ localization error in the tests.

4.3.2.1 Results: Subjective Localization of HOA at the Sweet Spot

In this experiment, the perceived angle of each stimuli (noise and female speech) by 13 listeners was collected for 3 different positions, 4 directions of sound, and 13 different systems (all the decoders and a monophonic reference). For each system, for each position and for each type of sound, we have computed the average angle $\bar{\alpha}$ (over the 13 subject answers) and its corresponding standard error $se(\bar{\alpha})$. We then compute the difference in between the average angle for the (ideal) monophonic source presentations and the average angle for each Ambisonic presentation. For example, if we wish to consider the performance at the sweet spot of the 'Energy' decoder from the direction of VS2 with noise, then we have

$$\begin{aligned} d \;=\; & \bar{\alpha}(\text{system} = \text{'Energy'}, \text{sound} = \text{noise} \ , \text{position} = \text{F}, \text{angle} = 67.5^{\circ}) \\ & -\bar{\alpha}(\text{system} = \text{'Mono'}, \text{sound} = \text{noise} \ , \text{position} = \text{F}, \text{angle} = 67.5^{\circ}) \end{aligned} \tag{4.12}$$

We also compute its standard error $se(d)$. We then sum up d over all of the different angles (VS1, VS2, VS3 and VS4) to give an overall difference d that does not depend on the direction of

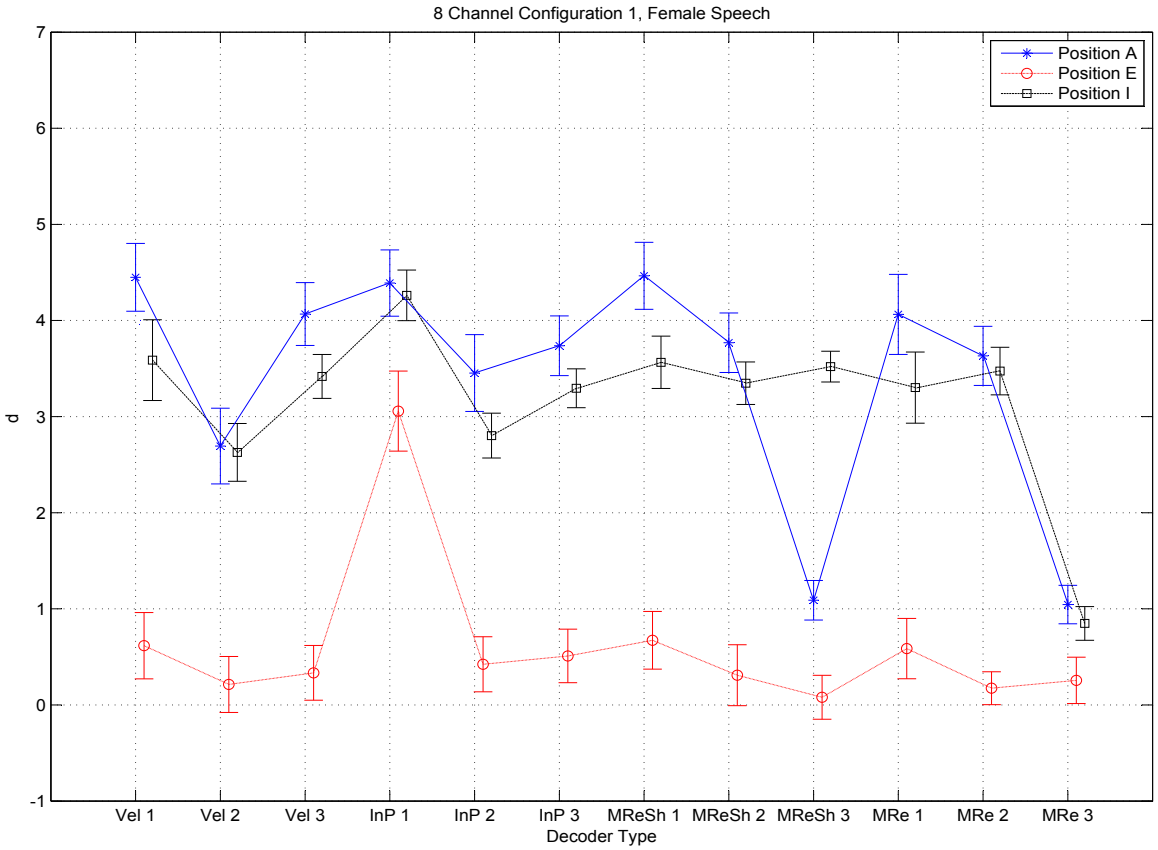


Figure 4.54: *Localization accuracy of* 1^{st} *to* 3^{rd} *Order Ambisonic decoders with female speech.*

---

the sound, and compute the resulting standard error. Figures 4.54 and 4.55 show this measure d computed for each listener position (Position A=blue, E=red, I=black) with the standard errors, for the female speech and pink noise presentations respectively.

The x-axis labels in these graphs can be interpreted as:

- Vel 1: 'Velocity' decode, Order 1
- Vel 2: 'Velocity' decode, Order 2
- Vel 3: 'Velocity' decode, Order 3
- InP 1: 'In-Phase' decode, Order 1
- InP 2: 'In-Phase' decode, Order 2
- InP 3: 'In-Phase' decode, Order 3
- MReSh 1: Shelf-filter decode, Order 1
- MReSh 2: Shelf-filter decode, Order 2

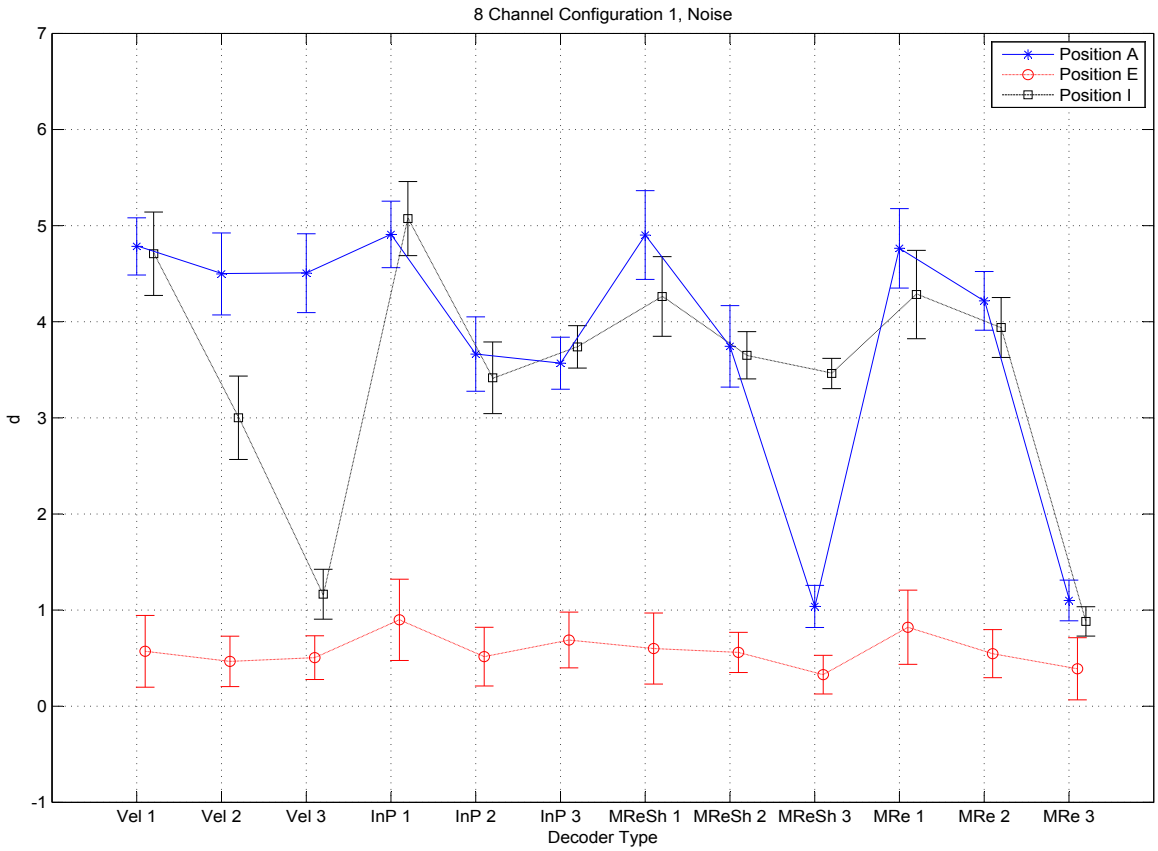


Figure 4.55: *Localization accuracy of* 1^{st} *to* 3^{rd} *Order Ambisonic decoders with pink noise.*

- MReSh 3: Shelf-filter decode, Order 3
- MRe 1: 'Energy' decode, Order 1
- MRe 2: 'Energy' decode, Order 2
- MRe 3: 'Energy' decode, Order 3

For the first order decoders at the sweet spot, we see that the poorest localization performance is exhibited with the 'in-phase' decode. This is particularly so with the female speech stimuli. The other decoders show better performance and are highly comparable. However, with pink noise, both the shelf filter decode and the 'velocity' decode perform marginally better than the 'energy' decode. Since this is not seen with the female speech stimuli, we conclude that low frequency optimisation is highly important Ambisonic decoding. That is, low frequency ITDs should be satisfied more so than high frequency ILD. For the second order decoders at the sweet spot, the performance is largely comparable. We note an overall increase in the localization accuracy, in particular with the 'in-phase' decode. Optimal localization with female speech is demonstrated with the 'Energy' decode. For noise, the best localization is achieved with the

'shelf' decode, which also has the smallest standard error. The third order 'velocity' decoder shows comparable results to the second order decoder, for both stimuli. The 'in-phase' decode is also similar to the second order implementation. However, the localization accuracy of the 'energy' decode is slightly better for the white noise case. Overall the optimal performance is achieved with the 'shelf' decoder, which displays the best overall localization accuracy for both stimuli, and a reduced standard error.

4.3.2.2 Results: Subjective Localization of HOA at Off-Centre Listening Positions

For the off-centre positions A and I, the localization accuracy is quite poor for first order decodes. The poorest localization is achieved with the 'velocity' decode, followed by the 'shelf' decode. This is not surprising, since as we have already seen from Experiment 2, the anti-phase components are optimized for the centre listening position, but give rise to precedence effects for off-centre listeners. Thus, we have the case where the best first order decoders for a centre listener are at the same time, the worst decoders for off-centre listening. Localization with the 'in-phase' decode is also poor. However, one should note that the 'phasiness' often experienced with Ambisonics in non-ideal listening situations is not experienced with the 'in-phase' decoder. Overall, the best 1^{st} Order localization accuracy was found with the 'Energy' decode, reconfirming the results of Experiment 2. As we move into higher orders the localization accuracy improves for all decoders. The best localization is achieved with the 3^{rd} order decoders as expected. Large improvements are seen in the case of white noise for the 'velocity', 'shelf' and 'energy' decodes, whilst the performance of the 'in-phase' decode is comparable to that of the second order case. The best performance is achieved with the 'energy' decode for both stimuli. It is noteworthy that the 'velocity' decode performed far better than the second order case with white noise, but poorer with female speech.

In conclusion, the results presented verify the objective analysis of Section 4.3.1. The best localization performance for a centrally seated listener is achieved using a psychoacoustically optimized decoder, where the low frequency velocity vector and high frequency energy vectors are maximized. However, contrary to popular thought, neither the 'in-phase' or 'shelf' decoders improve localization accuracy over a listening area. The optimal solution is to maximize the energy vector. This produces the least localization error over the entire audience area.

This assertion has also recently been verified by Frank et al [Frank et al., 2008]. In their study 15 subjects assessed the localization accuracy at two off centre positions and the sweet spot of a 12-channel rectangular array. At the sweet spot, they confirm the result that the velocity decode gives marginally better perceptual localization than the energy decode for broadband stimuli. However they did not assess an optimized shelf filter decode. At off-centre listening positions, they too reach the conclusion (by subjective analysis only) that the energy decode is optimal for off-centre listening and that the in-phase decode gives poor results.

4.4 Conclusions

In this chapter we have assessed the subjective and objective performance of VBAP and Ambisonic techniques, in terms of their localization accuracy for virtual sources for a distributed audience in non-ideal listening conditions. First, the localization vector theory was expanded to accommodate off-centre listening, and an objective analysis of VBAP and 1st Order Ambisonic decoders was presented. The results of a novel localization study were shown and these findings were supported by calculated ILDs and ITDs inferred from high resolution binaural measurements recorded in the test environment. An equivalent acoustic model was implemented and used to investigate acoustical aspects of the perceptual findings. From this analysis, it was concluded that whilst VBAP provides superior localization accuracy to 1st Order Ambisonics for stationary sources, 1st Order Ambisonics is preferred for moving sources.

Based on these results an investigation into improvements in the localization accuracy of Ambisonics by using higher order spherical harmonic representation of the soundfield was presented. This was conducted up to 3rd order in a horizontal only Ambisonic system, implemented on an 8-channel regular loudspeaker setup. It was found, in relation to the hypotheses presented in Section 4.3 that

1. The localization accuracy of Ambisonic systems increases under real listening conditions for all decoder types

2. Optimal localization accuracy for an individual listener over an 8-channel array can be achieved using a psychoacoustically optimized decoder, with maximized r_v and r_e at low and high frequencies respectively.

3. Optimal localization accuracy for a distributed audience is achieved by maximizing r_e over the entire frequency range. Decoders designed to optimize r_v result in large low frequency localization errors at off-centre positions.

Furthermore, the objective analysis revealed that VBAP and 3rd Order Ambisonics (with 'energy' decode) encounter similar limitations in the context of distributed audience listening. In both cases, localization accuracy for an 8-channel array is within the loudspeaker aperture of $\pm 22.5°$. The localization accuracy of the 3rd Order Ambisonics 'energy' decode was then verified subjectively as the optimal Ambisonic decoder solution for distributed audience listening.

We therefore justify the use of higher order Ambisonics throughout the remainder of this thesis since:

1. Higher order Ambisonic systems are asymptotically holographic, whereas amplitude panning systems such as VBAP are not.

2. In the context of real room listening, HOA can give equivalent localization accuracy to that of VBAP.

3. Coloration of moving sources is not perceived.

4. Recordings made with Ambisonics can be implemented using commercially available technology[1].

5. Higher Order Ambisonics is fully downward compatible with lower order systems.

We can therefore conclude, based on the simulations and listening test data of this and the previous Chapter, that homogeneous reproduction of soundfields for an individual listener can be achieved *practically* using 8 loudspeaker channels in a horizontal array with a psychoacoustically optimized 3rd Order Ambisonics decoder. This reproduction method can be applied directly to the virtual loudspeaker approach for headphone listening. Furthermore, decoding for distributed audiences should be optimized for energy in the intended source direction, and not velocity. A 3rd Order Ambisonics decoder should also be utilized for this purpose.

However, in order to satisfy the reproduction requirements for VAE presentation, there still exists challenges in content preparation and delivery over HOA systems. In particular, we ask the questions:

- How can we best record acoustic events for 3rd Order Ambisonic reproduction in a VAE?
- Is it possible to adapt 1st Order Ambisonic microphones for this purpose?
- On reproduction, how can we accommodate listener movement in the virtual space using 3rd Order Ambisonics?

Answers to these questions will be the focus of Chapters 5, 6 and 7.

[1]Recordings can also be made with soundfield microphones for VBAP using SIRR (see Section 7.4).

5

Virtual Acoustic Recording

5.1 Introduction

An important consideration in the formation of a virtual acoustic environment is content preparation, i.e. how the source audio is captured and prepared for reproduction in a VAE. Recording an acoustic event represents the first, and perhaps the most important step in any audio production chain. It is well known that the tonal result from the ultimate microphone placement for any close monophonic recording is highly subjective and dependent on many physical parameters, such as the 'proximity effect'. Likewise in the realm of single-perspective Ambisonic recording, there are several significant physical considerations for placement of soundfield microphones which contribute to aesthetically pleasing results. The most important of these is the decision to record in either the *direct* or *reverberant* sound fields. In the context of multi-perspective VAEs a direct/reverberant field approach must also be taken. However, the main challenge is that we must also accommodate listener movement through the virtual space. Based on the psychoacoustic principles outlined in Chapter 2, as a listener moves through a virtual room, the level of the direct sound should change in accordance to the inverse-distance law and the directivity of the source. The early reflections should also change according to the relative position of the source to the listener, and the level of the diffuse-field should remain constant.

Thus, in this chapter, we propose an alternative approach to recording live acoustic events for VAE reproduction. We consider that recording of an acoustic event should be taken in the direct-field and that this recording can later be processed according to measured acoustics at the instantaneous position of the listener. Such acoustical measurements are known as *'spatial room*

impulse responses' (SRIRs), as they preserve the correct spectral and temporal properties of the reverberant soundfield due to a given stimulus at the exact listener position. This methodology is further justified in the context of fixed-perspective scenarios since, quite often, natural acoustic recording in an ideal reverberant space is not practical due to low noise isolation, microphone configuration, equipment and performance issues, or simply budgetary limitations.

This chapter is organized as follows: First we will present a succinct review of traditional recording strategies for natural acoustic recording of live acoustic events. We will then investigate the use of measured acoustic impulse responses in the formation of auditory scenes, with the aim of achieving perceptual equivalence to natural recordings. Finally an objective and subjective evaluation of actual vs. virtual acoustic recordings in a fixed-perspective scenario is presented.

5.2 Natural Acoustic Recording

Before we can address multi-perspective recording for VAEs we must first understand the challenges involved in natural, single-perspective acoustic recording. As an example, consider the orchestral recording scenario shown in Figure 5.1. Here a two or three-channel microphone technique (or even a soundfield microphone) is used to capture the sound field close to the orchestra, i.e. it is the *main microphone*. A secondary microphone array is usually placed further back in the auditorium to capture ambience. This is the *room microphone*. Spot microphones are then used on several key individual instruments, such as a solo violin or a soprano singer, so that their acoustic balance can be controlled more effectively in the final production. Thus, in single-perspective recordings, the spot microphones are used to reinforce the main microphone, with the room microphone adding the desired level of reverberation and acoustic ambience. Indeed, a major advantage to having separate direct-source/ambient microphones is that the level of the reverberant sound field can be controlled post-recording in the mixing studio.

The following considerations therefore apply to such recording:

1. The main microphone should be located close enough to the source in order to capture its *direct* soundfield.

2. The recording angle of the main microphone must be such that a satisfactory stereophonic imaging of the entire orchestra/ensemble is achieved.

3. The spot microphone signals should have sufficient off-axis rejection so that crosstalk between instruments can be reduced.

4. The room microphone should be placed back into the reverberant field at a distance such that satisfactorily decorrelated signals are recorded.

5. On mixing, not only is level difference a key component, but also proper time alignment between the individually recorded elements is crucial to obtaining a satisfactory result.

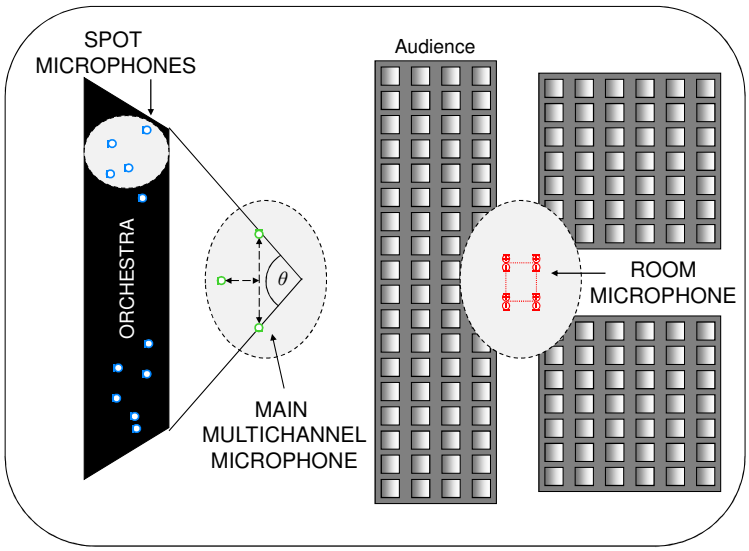


Figure 5.1: *Example orchestral recording setup.*

5.2.1 Direct Field Recording

In a given recording scenario, the critical distance point using equation 2.13 should be estimated before the positioning of any microphones. Invariably, direct-field capture of an acoustic source can be achieved using a monophonic spot microphone. The positioning of the microphone is important not only to the tonal balance, but also to minimize the amount of other instruments (commonly referred to as 'spill') in the recorded signal. In order to achieve such rejection, the directional characteristic of the microphone is extremely important. The directional characteristics of unidirectional microphones (whose equivalent response functions are illustrated in Fig 3.19) are given in Table 5.1. Hypercardioid and supercardioid microphones are frequently used in order to maximize rejection, but the cost of increased directional response can often lead to compromised frequency response in lower grade microphones. Furthermore, the more directional the microphone, the greater the 'proximity effect'. This is an exaggerated bass response, most noticeable at distances less than half a meter. This arises from the difference in the pressure-gradient due to the curvature of the soundfield at small distances.

The proximity effect has particular importance when dealing with vocalists, who often prefer to sing extremely close to the microphone. This results in unnatural sounding recorded vocal which, whilst often desirable for popular music production, is not acceptable for natural sounding reproduction in a VAE application. As a result, the proximity effect needs to be compensated. This is addressed in Section 5.3.3.1.

5.2.2 Applicability of Direct-Field Stereophonic Recording Techniques

Several stereophonic multichannel microphone arrays have been proposed for direct-field acoustic recording aimed at single-perspective reproduction over two and three channel stereo and (over

Direction (relative to micro axis)	0°	15°	30°	45°	60°	75°	90°	105°	120°	135°	150°	180°
Cardioid	0	-0.3	-1	-2	-3	-4	-6	-8	-11	-15	< -18	< -18
Super-cardioid	0	-0.3	-1	-2	-3.5	-5	-9	-14	< -18	< -18	< -18	-11
Hyper-cardioid	0	-0.3	-1	-2	-4	-7	-12	< -18	< -18	-12	-9	-6
Figure of eight	0	-0.3	-1	-3	-6	-12	< -18	-12	-6	-3	-1	0

Table 5.1: *Directivity of unidirectional microphones (dB).*

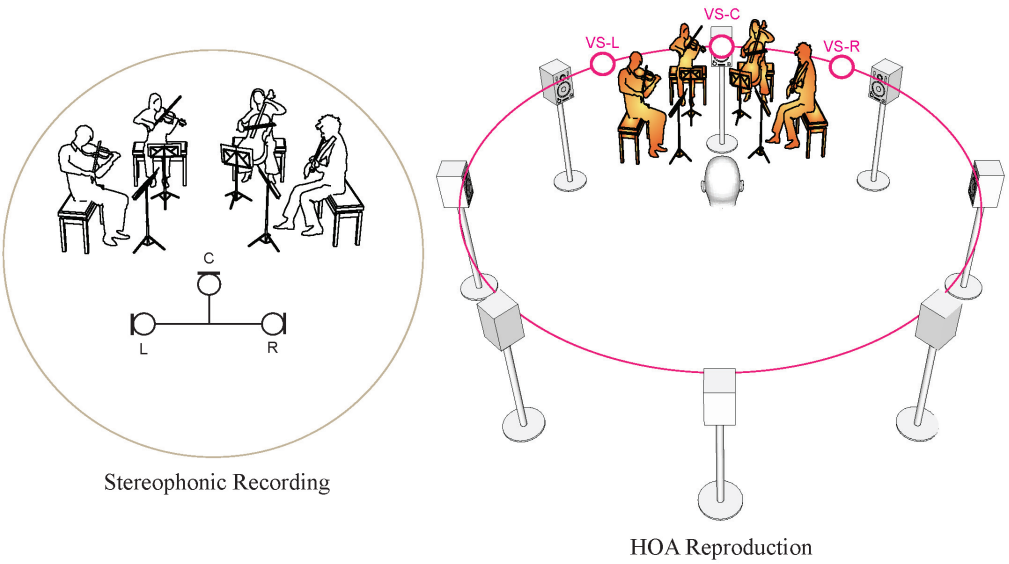


Figure 5.2: *Re-encoding of stereophonic material for Higher Order Ambisonic reproduction.*

the last two decades) with the ITU 5.1 layout. Techniques which have gained popular use include the Optimized Cardioid Triangle (OCT array) [Wittek and Theile, 2002], the Williams array [Williams, 1991] and the Decca Tree [UDK, 1994]. However, if we are to consider multichannel stereophonic techniques for application to VAE presentations then we will need to employ some mechanism that allows for the techniques to be re-encoded to higher order Ambisonics. One solution is to encode stereo virtual loudspeaker positions and feed them the recorded stereophonic signals. Such a method is illustrated in Figure 5.2. Here, a natural acoustic recording is picked up with a three channel microphone technique, such as an OCT array. Ordinarily, the left, centre and right microphone feeds would be presented to the corresponding loudspeakers in a 3-channel stereophonic reproduction system. Instead, we wish to use higher order Ambisonics, and so three virtual sources, labeled VS-L, VS-C and VS-R are created in an Ambisonic field at the same positions that the loudspeakers for the three-channel reproduction would have been located ($-30°$, $0°$ and $+30°$). The microphone signals are then used as the source signals of each virtual source. A similar method has been proposed by [Chen, 2009], known as B+

Format, where the standard 1st Order Ambisonic signals are accompanied with a stereophonic direct-field recordings. These are not elegant methods however, and as we shall see later on from the point of view of VAE reproduction, it is in fact more desirable to have monophonic signals for convolution and directivity enhancement. To this end, we have recently investigated methods of converting stereophonic material to multiple-monophonic sources for remixing in multichannel presentations [Barry and Kearney, 2009]. The notion is that monophonic sources can be extracted from a stereophonic mix using source separation techniques such as the ADRess [Barry et al., 2005] or DUET [Jourjine et al., 2000] algorithms. It was shown that theoretical reconstruction errors associated with the source separation process manifest themselves as image shifts in remixed presentations (a 5.1 example was shown in this paper) and thus leads to perceived localization distortion. The tests conducted indicate that separation algorithms are suitable for remix applications particularly for audience view/ensemble view conversion. This conclusion has also been recently been corroborated by Cobos et al [Cobos and Lopez, 2009] for application of stereophonic to wave field synthesis up-mixing. However, for simplicity, we will only consider here monophonic pickup of direct-sources. The reader is directed to [Kearney and Levison, 2008], [Theile, 2001], [Williams, 1987] and [Wittek and Theile, 2002] for further information on the methodology of direct-field stereophonic recording.

5.2.3 Multichannel Recording in the Reverberant Field

Typically, in standard surround sound presentations of natural acoustic recordings the level of the direct sound in the lateral soundfield should be at least 10dB below that of the frontal direct-sound if good image focus is to be maintained [Theile, 2001]. Since this lateral soundfield is usually the feed from the room microphone, it makes sense to place the room microphone at a position where the direct sound pressure level is at least -10dB relative to the sound pressure level at the main microphone. Therefore, knowing that the direct sound pressure level falls inversely proportional to the distance from the source, we can calculate an ideal placement in the reverberant field for the room microphone. This distance D_{Room} between the main microphone and room microphone is given as

$$D_{\text{Room}} = \sqrt{10^{(SPL_{\text{Dir}} - SPL_{\text{Rev}})/10}} \tag{5.1}$$

where SPL_{Rev} is the desired sound pressure level at the centre of the room microphone (in dB) and SPL_{Dir} is the sound pressure level at the centre of the main microphone in the direct field (in dB). This is illustrated for a critical distance of 2.7 meters in Figure 5.3. Here, -10dB suppression of the direct sound in the room microphone occurs at a distance of 5.25m away from the main microphone. Note that the *optimal* placement of the room microphone will not occur if the engineer empirically *measures* the -10dB level of the room microphone relative to the main microphone, since this sound pressure level will be the result of the direct SPL and the reverberant SPL.

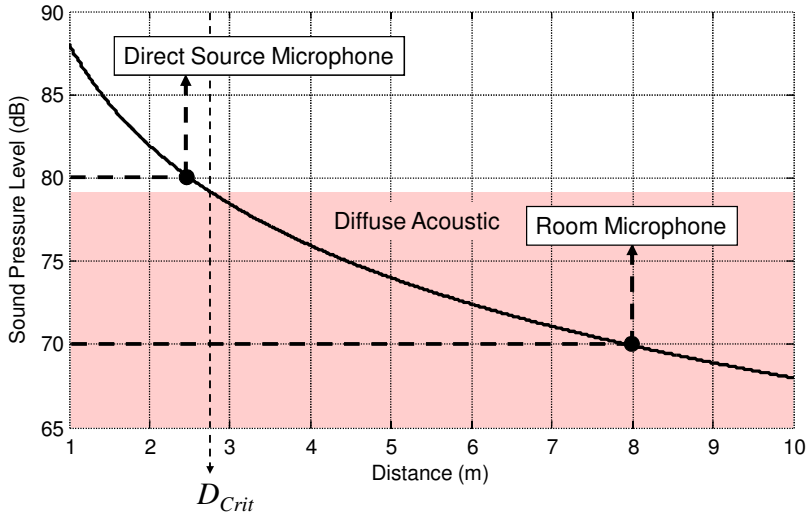


Figure 5.3: *Main and ambient microphone positioning with respect to critical distance (after [Theile, 2001]).*

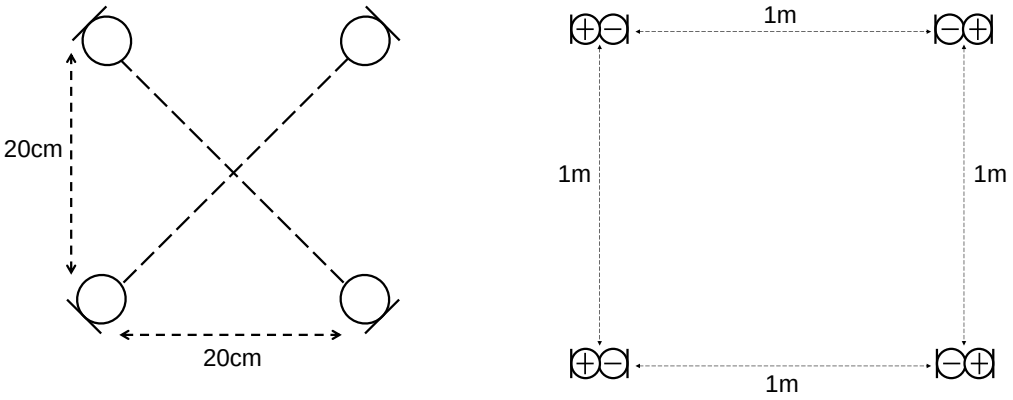


Figure 5.4: *Four channel room microphones: (a) IRT Cross configuration and (b) Hamasaki Square configuration.*

Whilst Ambisonic microphones can be used as room microphones in the reverberant field, for the most part, 4-channel microphones are used with the resultant recorded field mixed with the direct sound field on reproduction. Well known configurations include the IRT Cross [Theile, 2001], which consists of four cardioids arranged quasi-coincident at 90° angles, and the Hamasaki square, which consists of 4 spatially separated fig-8 microphones [Hamasaki, 2000, Hamasaki and Hiyama, 2003]. Both these configurations are shown in Figure 5.4. The Hamasaki Square ensures that sufficiently spatially decorrelated sounds are picked up since the null point of the microphones (the point of zero gain) is faced towards the source. It is therefore well suited to the pickup of room reflections and reverberations.

Given that the reproduced reverberant soundfields are largely diffuse, localization accuracy is not a major concern. Thus, if we wish to include the diffuse field recording from a Hamasaki

square into a 3rd Order Ambisonics recording, we merely have to create four virtual Ambisonic source positions, placed at $\pm 90°$ and $\pm 270°$ and feed the Hamasaki square signals to each virtual source.

5.2.4 Natural Acoustic Recording using Ambisonic Techniques

The optimal positioning of a soundfield microphone (in a single-listener perspective scenario) is at the critical distance. It then is possible to modify the direct to reverberant ratio (to some degree) by using the forward dominance operation of equation 3.22. The soundfield can also be rotated in post-production such that the correct perspective is reproduced. Furthermore, 1st Order Ambisonic recording has several practical advantages over other multichannel recording methods:

1. The use of multichannel microphone arrays means that we must consider playback formats at the recording stage. However, Ambisonics is a method of soundfield encoding that is independent of the reproduction method.

2. The number of recording channels with B-Format Ambisonics is 4 (or 3 if height is not required), whereas the number of channels for both direct and reverberant field pickup in multichannel microphone arrays is never less than 5 for 3/2 playback and increases with the number of output channels required. Thus, the amount of data storage required is greater.

3. Any coincident microphone technique based on 1st order directional microphone patterns can be implemented using a soundfield microphone.

However, there are also several distinct disadvantages to using soundfield microphones in this manner:

1. The use of the forward dominance equations results in spatial image distortion, i.e. with greater forward dominance, the sound stage appears stretched, whereas with less dominance the sound stage reduces in width.

2. Recording with a sound field microphone limits the reproduction accuracy to 1st Order Ambisonics.

3. Multi-perspective reproduction is not possible.

In general, good natural acoustic recordings using either multichannel microphone arrays or soundfield microphones can be difficult to achieve. Quite often, recording environments such as Churches or concert halls can have ideal reverberation characteristics, but low noise isolation from external sources. Furthermore the ideal placement of the room microphones may not be possible during live acoustic performances due to audience considerations. Finally, situations can

arise where an instrument may be required to be overdubbed onto the original acoustic recording. In this case, the logistics of setting up for re-recording in the original acoustic environment may not be practical.

5.3 Virtual Acoustic Recording

A solution to the aforementioned problems in natural acoustic recording is to use *convolution filters* where a dry audio signal is convolved with impulse responses taken from an ideal reverberant environment with the aim of auditory scene synthesis. For example, if we have a measured impulse response, $h(t)$ and a dry source audio signal $s(t)$, then the resultant convolution of the two signals is given by

$$(s * h)(t) = \int_{-\infty}^{\infty} h(\tau) \cdot s(t - \tau) d\tau \tag{5.2}$$

where τ is a time lag. The resultant audio signal now has the characteristics of a source recorded in the measured environment. For large filters, significant gains in computation speed can be achieved by performing convolution in the frequency domain, since

$$\mathcal{F}(s * h) = \mathcal{F}(s) \cdot \mathcal{F}(h) \tag{5.3}$$

where $\mathcal{F}$ denotes the Fourier transform. Convolution reverberation techniques have become increasingly popular over the last two decades and a significant body of work has been presented illustrating the possibilities for multichannel reproduction, for example in [Merimaa and Pulkki, 2005], [Kessler, 2005] and [Kearney and Levison, 2008]. One important, and often overlooked, practical consideration of such audio processing, is the coloration that occurs on the impulse response due to spectral and temporal differences at each stage of the convolution chain. Loudspeaker and microphone responses and their directivity functions, as well as proximity effects and the test stimulus used all contribute to changing the impulse response from a true representation of the acoustic space.

Another significant factor is that the representation of room responses in this manner assumes that the source-room interaction is one that is linearly time-invariant (LTI). In reality, this interaction is far more complex. The impulse response changes significantly with changes in the spatial position between the source and the listener. It has also been reported that small movements of musicians can lead to significant changes in timbre at the listening position [Greisenger, 1989]. Vocalists are a particularly good example of this, since the excitation of the room is directly influenced by the directional orientation of the singer. Finally, although the perceptual changes of reverberant recordings over different loudspeaker layouts has been well documented, the perceptual differences of recordings made with convolution reverberation against natural acoustic recordings has not. Furthermore, there currently exists no formal framework for the recording of live acoustic events such that multi-perspective listening is possible. Such a recording framework will now be proposed and is termed 'Virtual Acoustic Recording'. We define 'Virtual Acoustic Recording' as

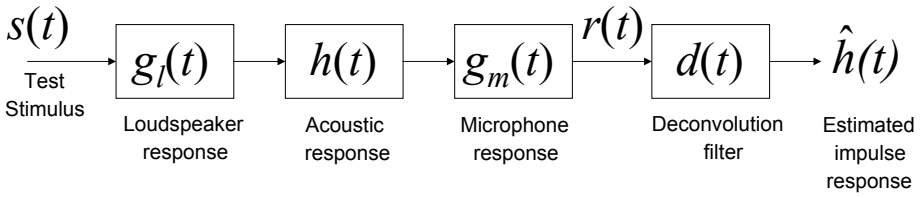


Figure 5.5: *Typical impulse response measurement chain.*

The capture of the spatial and temporal properties of a true acoustic event and its acoustic environment such that synthesis of said event from multiple perspectives in an interactive virtual auditory environment is possible.

The main objectives of the framework are:

- To reproduce acoustic events using convolution based techniques such that perceptual equivalence to natural acoustic recording is achieved.
- To minimize the audibility of coloration due to the convolution process
- To accommodate multi-perspective listening of recorded acoustic events

In this section, we will investigate the plausibility of forming auditory scenes through 'virtual acoustic recording', and the perceptual artifacts that arise from the signal processing techniques employed. We will begin by outlining the process of recording and capturing room impulse responses.

5.3.1 Convolution and Deconvolution in Multichannel Audio

A typical time domain convolution measurement chain is given in Figure 5.5. An excitation signal $s(t)$ is played through a loudspeaker in the acoustic environment and subjected to the temporal and spectral effects of the room. The signal is recorded and $s(t)$ is then deconvolved out of the recorded signal $r(t)$ to obtain the LTI response $\hat{h}(t)$. This is the typical convolution chain that pertains to most common impulse response recording situations. However, in order to obtain a more accurate estimation of the room response, we must compensate for the responses of both the loudspeaker and the microphone. This can be achieved in the deconvolution filter $d(t)$ by computing

$$D(\omega) = S^{-1}(\omega)\ G^{-1}(\omega) \tag{5.4}$$

where $D(\omega)$ and $S^{-1}(\omega)$ are the Fourier domain representation of the convolution filter and excitation signal respectively and

$$G(\omega) = \mathcal{F}\left[g_l * g_m\right] \tag{5.5}$$

The estimate of the room response $\hat{h}(t)$ can then be obtained by

$$\hat{h}(t) = \mathcal{F}^{-1}\left[R(\omega)D(\omega)\right] \tag{5.6}$$

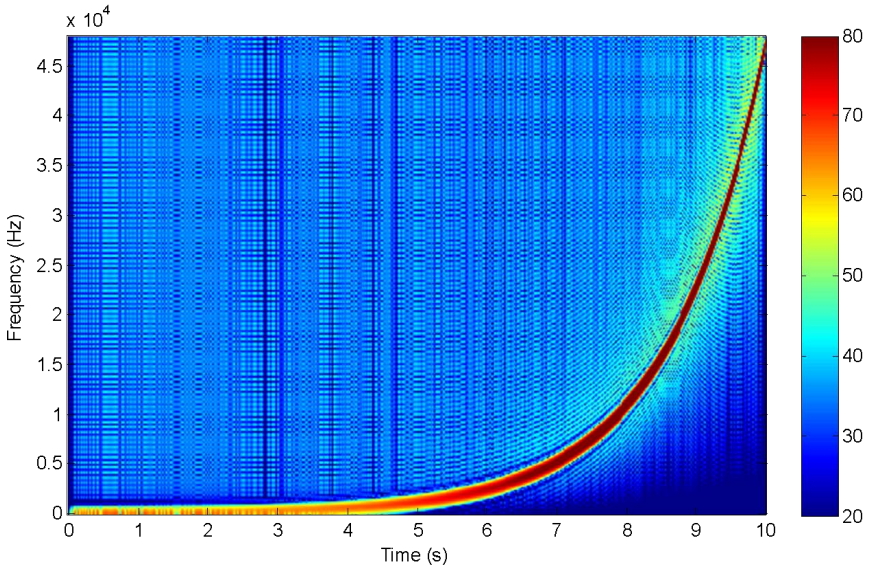


Figure 5.6: *10 second Time Stretched Pulse.*

where $R(\omega)$ is the Fourier domain representation of the recorded signal $r(t)$.

The original excitation signal is of critical importance since it must excite the room response over the entire audible spectrum. Maximum Length Sequences (MLS) are popular choices due to the fact that their autocorrelation is a Kronecker delta function [Borish and Angell, 1983]. However, it is generally accepted that the use of a TSP (Time Stretched Pulse) provides several advantages over the MLS technique, namely that it is more robust to time variances and non-linearities of the room [Farina, 2000], [Kessler, 2005]. Such a TSP is shown in Figure 5.6. TSPs are an exponential sine tone sweep that when played slowly enough, will result in all speaker induced distortion in the impulse turning into pre-delayed signals at the start of the impulse response [Farina, 2000]. The length of the tone must be greater than that of the room's reverberation time multiplied by the number of octaves (taken as 10) [Kessler, 2005].

5.3.2 Spatial Impulse Responses

If we record an impulse response measurement using a single microphone, we can give an impression of the acoustic space on reproduction, but no spatial information will be present. On the other hand, if we use multichannel recording methods, we can faithfully capture the direction of the direct sound and the early reflections. Several different methods of recording multichannel impulse responses have been proposed in the literature. The use of the 5-channel Williams MMA microphone configuration has been shown by Kessler [Kessler, 2005] to provide a good general approach to impulse response recording for the ITU 5.1 layout and is based on the well known Williams Curves for recording angle calculation [Williams, 1987]. Farina [Farina and Tronchin, 2004] suggests a combinational method of using a binaural head, an Ambisonics

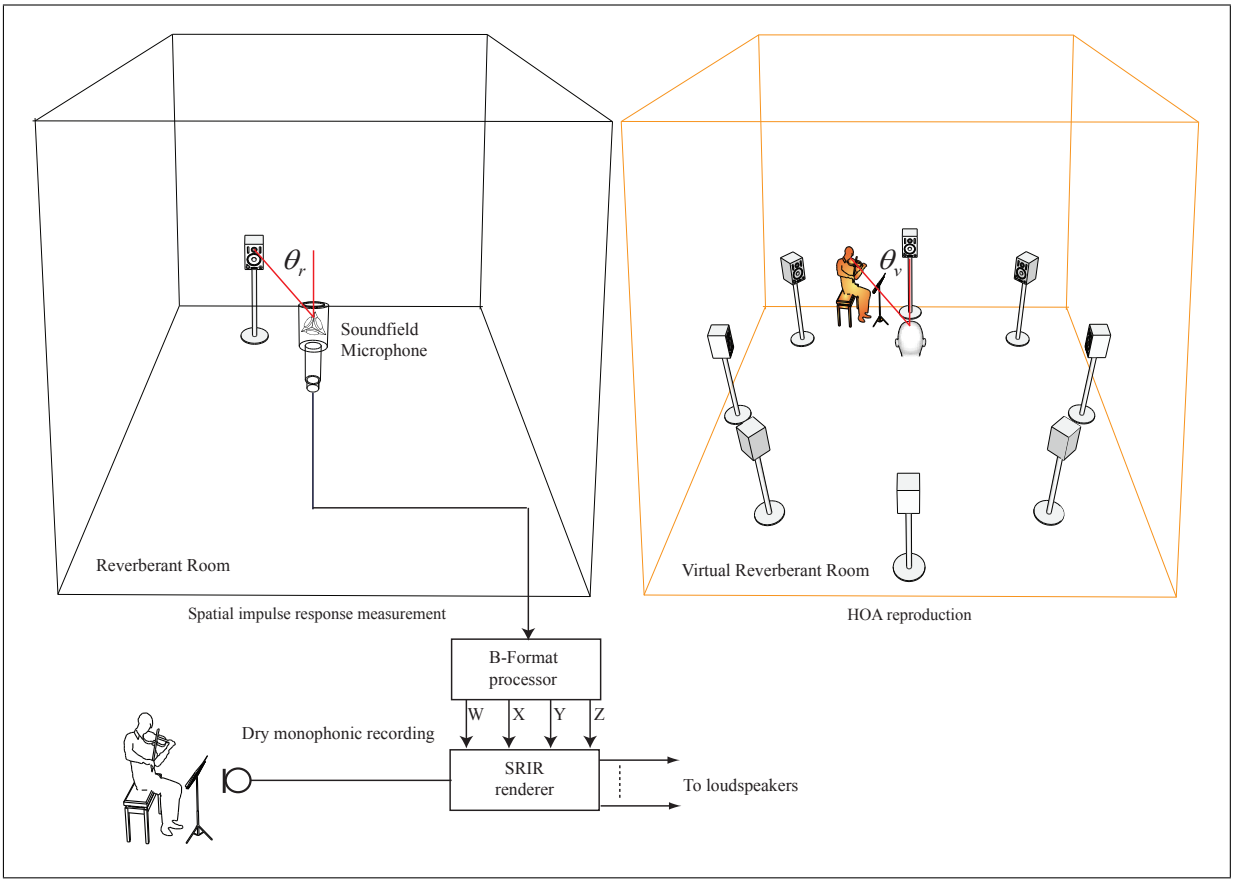


Figure 5.7: *SRIR capture with a sound-field microphone.*

Soundfield microphone and an ORTF pair[1] for full periphonic impulse response capture. The author has also suggested the use of the OCT and Hamasaki square arrays for use with 7.1 reproduction systems [Kearney and Levison, 2008].

However, approaches to impulse response capture that utilize multichannel stereophonic arrays are limited in that arrays such as the Williams MMA or the OCT are optimized for specific multi-channel setups and not for higher order Ambisonic representation. Whilst this can be overcome by re-encoding the loudspeaker feeds to Ambisonics, it is not optimal. Furthermore, in any spatial impulse response measurement, if the amplitude and arrival times of the impulse are preserved (i.e. no normalization) the source-receiver angles are fixed for each measurement. This concept is illustrated in Figure 5.7. Here a soundfield microphone is used to capture the impulse response due to the loudspeaker in the measurement room. Processing of the 1st Order Ambisonics channels, W, X, Y, and Z results in impulse responses h_w, h_x, h_y and h_z. Without any transformations, these impulse responses can be convolved with a monophonic source signal

[1]ORTF: A stereophonic microphone technique developed by Office de Radiodiffusion Télévision Française [Streicher and Everest, 2006].

S, resulting in

$$\hat{s}_w = h_w * s \tag{5.7}$$

$$\hat{s}_x = h_x * s \tag{5.8}$$

$$\hat{s}_y = h_y * s \tag{5.9}$$

$$\hat{s}_z = h_z * s \tag{5.10}$$

These new B-Format signals can then be decoded using the operations of Section 3.4.2. The result is the placement of the monophonic performance at the same azimuthal angle of incidence as the measurement loudspeaker was to the measurement microphone. Furthermore, the spatial characteristics due to the measurement loudspeaker (reflections and diffuse field reverberation) are also now characteristics of the reproduced source i.e. the reproduced source sounds like it was recorded in the measurement environment. The reader should note that although a 3-D 'virtual room' is depicted here for illustrative purposes, the resultant soundfield is, of course, only 2 dimensional with a horizontal only array. Thus, if the wish is to construct equivalent recordings from both 'live' and 'dry' acoustic environments, impulse measurements must be taken at each of the performers positions in the live room using the same microphone array.

1st Order Ambisonics offers the ability to capture directional impulse responses via a Soundfield microphone in the X, Y and Z directions and is therefore an elegant method for recording impulse responses. However, if the desire is to reproduce 3rd order sound fields, we must consider ways to re-purpose 1st Order impulse responses for higher order Ambisonic reproduction. This is addressed in Chapter 7. However, since we wish to investigate the subjective attributes of actual vs virtual acoustic recordings, we will continue to utilize 1st Order soundfields, due to the commercial availability of 1st Order microphones.

Given the ideal characterization of room responses in the aforementioned manner, it is prudent to ask if the compensation for distortions from the loudspeaker response and microphone response are a major perceptual feature, since the majority of impulse responses in recording productions are taken and applied without such compensation. This is, of course, a function of the quality of the loudspeaker and microphones used, but the extent to which this may or may not be perceived given reasonable commercially available recording equipment has not yet been investigated. Thus, we will now consider the coloration that occurs through each step of the convolution chain.

5.3.3 Equalization of the Convolution Chain

Starting from impulse response measurement, through to final convolution, shifts in timbre of the final audio from its desired tone can be caused by:

1. The free-field response of the measurement loudspeaker
2. The free-field response of the measurement microphone

3. The free-field response of the source audio microphone
4. Proximity effect due to source audio microphone directivity and close-source recording
5. The free-field response of the reproduction loudspeaker
6. The acoustic response of the reproduction environment

Since we wish to compare true acoustic recordings against virtual acoustic recording then we will assume that the main measurement/recording microphone has a sufficiently linear frequency response up to 20kHz and will not bias in any way the overall perceptual differences. This assumption can also be made for the reproduction setup. However, spectral distortions due to cases 1, 3 and 4 cannot be ignored.

We can consider the total response of each of these effects as a cascade of convolution filters, each one affecting the overall magnitude and phase response of the IR, represented in the frequency domain as

$$H_e(f) = H_l(f) + H_m(f) + (H_m(f) - H_p(f)) \tag{5.11}$$

where $H_e(f)$ is the overall system response, $H_l(f)$ is the response due to the measurement loudspeaker, $H_m(f)$ is the omnidirectional response of source recording microphone and $H_p(f)$ is the response due to proximity effect of the source to the source recording microphone. If we know each of these parameters, then it becomes possible to construct an inverse filter that can be applied to the monophonic source audio prior to convolution. However, unless an anechoic chamber is available to the engineer, it is difficult to obtain the required free field measurements. Struck and Temme [Struck and Temme, 1994] have presented a framework for performing simulated free-field measurements in regular environments that give equivalent results to anechoic measurements, which will be utilized here.

5.3.3.1 Measurement of Convolution Coloration

In a typical loudspeaker response measurement, the microphone is in the far field from the loudspeaker (usually $<$ 1m) and the room is totally absorptive. If the room is reflective then we must truncate the measured impulse response before the first arriving reflection. This in turn, determines the frequency resolution of the loudspeaker measurement, where the lowest frequency is given by

$$f_{low} = \frac{1}{T_1} \tag{5.12}$$

where T_1 is the time of arrival of the 1st order reflection. Let us assume that the nearest reflecting surface is the floor, and that the loudspeaker and measurement microphones are at a height, h from the floor and distance d from each other, then equation 5.12 becomes

$$f_{low} = \frac{c}{2\sqrt{\frac{d}{4}^2 + h^2}} \tag{5.13}$$

where c is the speed of sound. The direct sound is truncated from the impulse response by multiplying the measured impulse response by a rectangular time window of length T_1 with leading and trailing Hanning ($\cos^2$) time windows at 10% of T_1. The low frequencies that are below f_{low} can be measured by moving the microphone as close as possible to the driver. According to Struck [Struck and Temme, 1994], the following empirical formula results in measurement errors of less than 1dB:

$$d < 0.11a \tag{5.14}$$

where a is the radius of the loudspeaker cone. Since the microphone is placed right up to the centre of the cone, reflections are physically eliminated. The upper frequency limit at which this technique can be performed is where $ka = 1$, where k is the wavenumber $(2\pi f)/c$, or

$$f_{high} = \frac{c}{2\pi a} \tag{5.15}$$

$$= \frac{10949.86}{2a} \tag{5.16}$$

for a cone diameter $2a$ in centimeters. Since this measurement is made in the near-field, the sensitivity will appear much higher, and needs to be matched to that of the high frequency far-field measurement, such that

$$H_{FF}(f) = H_{NF}(f) - 20\log_{10}\frac{4d}{a} \tag{5.17}$$

where $H_{FF}(f)$ is the far-field measurement and $H_{NF}(f)$ is the near-field measurement. Time delay of the near-field measurement must also be implemented, so that discontinuities in the phase spectrum can be avoided. The amount of delay is related to the distance between the near and far-field measurements and the loudspeaker

$$\tau_d = d_{FF} - d_{NF} \tag{5.18}$$

Once the magnitude spectrum of the near field measurement has been scaled an overlap region should be clearly visible between f_{low} and f_{high} and a suitable crossover frequency f_x can be determined. Phase alignment can then be implemented by introducing a further delay τ_x such that

$$\tau_x = \frac{1}{360}\frac{\phi_{NF} - \phi_{FF}(f_x)}{f_x} \tag{5.19}$$

Of course, to obtain the true measurement of $H_l(f)$, we must have a high precision microphone with a linear frequency response. However for the application presented here a precision microphone is not necessary. In fact, it is more advantageous to use the microphone used to record the source audio, such that we obtain the combined response function

$$H_c(f) = H_l(f) + H_m(f) \tag{5.20}$$

An example of this combination is shown in Figure 5.8. Here the responses of near field and far field measurements of a Genelec 1029a loudspeaker measured with an AKG-C414 microphone,

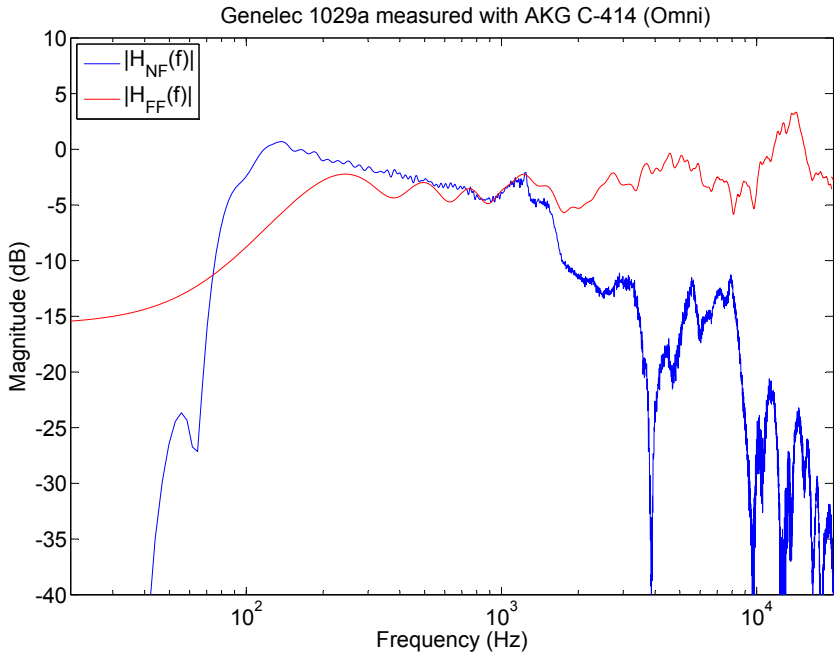


Figure 5.8: *Near and far field frequency response measurement for Genelec 1029a loudspeaker measured with AKG C-414 microphone (omnidirectional setting).*

set to omnidirectional are plotted. The far-field measurement was taken at 2m at a height of 1.24m from the floor, limiting the low frequency resolution to around 638Hz. The far field measurement is scaled and we clearly see the existence of an overlap region at approximately 1kHz.

We can further extend this notion to include proximity effect, if we measure the near-field response using a unidirectional microphone. The notion is that we can also compensate for the exaggerated low frequency response in the direct field recording of performers such as vocalists. Ideally then, the unidirectional microphone setting employed for the measurement should be the same as that employed for the recording of the performance. Such a low-frequency response due to the proximity effect is given in Figure 5.9. Shown is the response of the Genelec 1029A measured with the AKG-C414 microphone first set to omnidirectional and then with cardioid directionality. We see that as the microphone becomes more directional, the level of the bass response increases by 6dB at approximately 105Hz. This exaggerated low frequency response can now be combined with the far-field measurement to obtain a total system response as shown in Figure 5.10. Thus, we can now measure simultaneously all coloration effects in the convolution chain, in a regular environment.

Once the transfer function describing the coloration in the convolution chain has been determined, an appropriate inverse filter must be derived. Several solutions to the design of inverse filters exist and a good review can be found in [Farina and Ugolotti, 1998] In one method, proposed by Neely and Allen [Neely and Allen, 1979], a minimum phase inverse filter is formed by taking the FFT of the original impulse response and then spectrally smoothing the magnitude.

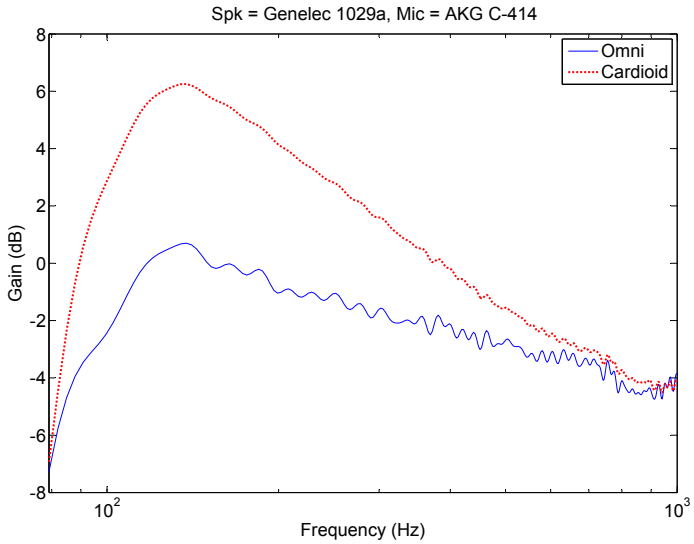


Figure 5.9: *Proximity effect for AKG-C414 using a Genelec 1029A as a source (r = 1cm).*

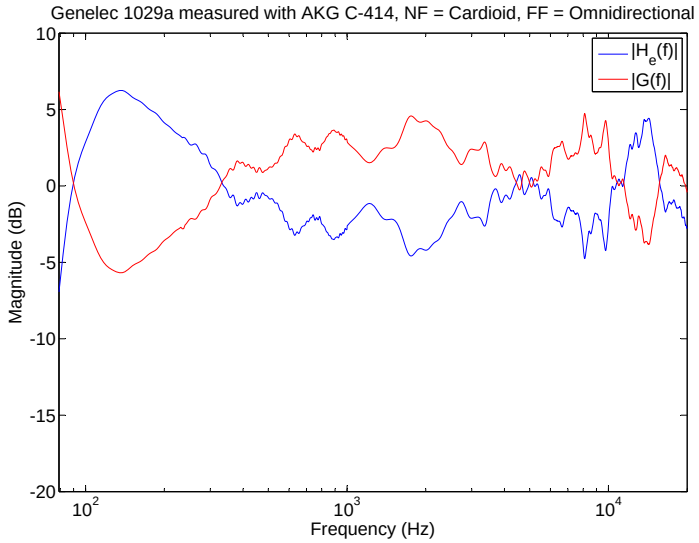


Figure 5.10: *Total system response and inverse using Kirkeby regularization [Kirkeby and Nelson, 1996].*

The IFFT then yields an inverse filter function which sufficiently flattens the magnitude spectrum. However, the phase of the complex spectrum is omitted. Mourjopoulos [Mourjopoulos, 1994] proposes the use of a time-domain least squares inversion of the impulse response, such that when convolved with the original impulse, approximates a Dirac function. However, if the spectrum of the impulse has sharp dips, then the inverse filter will have extremely sharp peaks which can often lead to audible and undesirable artifacts. Instead, it is best to use some form of frequency domain regularization, such that the denominator in the inverse filtering function

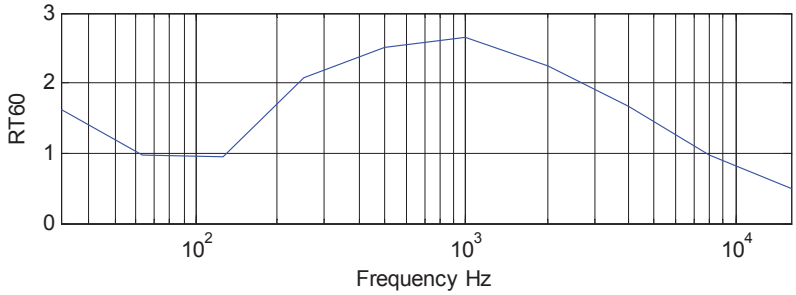


Figure 5.11: *Spatially averaged RT_{60} of Trinity College Chapel.*

does not allow the magnitude spectrum to become too large. Kirkeby [Kirkeby and Nelson, 1996] has put forward a formulation where the inverse filter is given as

$$G(f) = \frac{conj[H(f)]}{|H(f)|^2 + \Gamma} \tag{5.21}$$

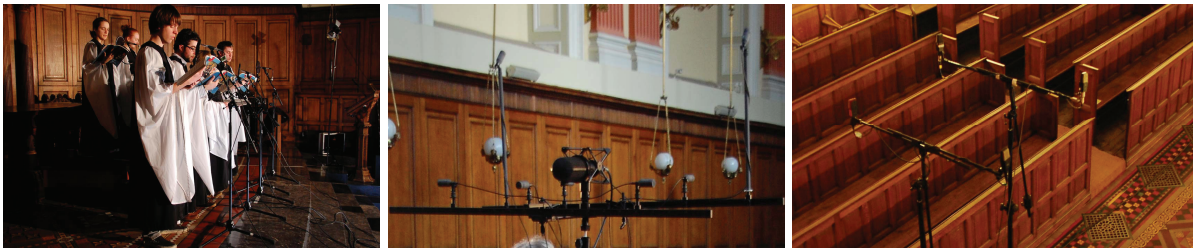
where Γ is the regularization parameter. After the inversion in the frequency domain, the filter is transformed back to the time domain, where a half-length rotational shift is applied to the end of the response. Various attempts at different values of the regularization parameter are generally undertaken before a final inverse filter is designed. The frequency domain inversion of $H_e(f)$ using this method is shown in Figure 5.10, with $\Gamma = 1$ and inversion up to 16kHz only.

5.3.3.2 Perceptual Effects of Coloration in the Convolution Chain

It is certainly prudent to question the perceptual relevance of removal of coloration in the convolution chain, in particular, where good quality recording and measurement equipment is used. This topic has recently been studied by the author in [Kearney and Levison, 2008], the main results of which are presented here. In this study, natural room recordings were directly compared to close microphone recordings of the same performance, convolved with spatial impulse responses. This method allows for direct comparison of the perceived audio without variation in the artistic performance.

For this study, both actual and virtual acoustic recordings were made at the chapel in Trinity College Dublin, Ireland of the resident Chapel Choir. This chapel is a highly reverberant and diffuse performance environment with a volume of 5600m^3. The spatially averaged reverberation time was measured at 8 different points in the room (3 stage/altar source positions, 5 audience receiver positions and 3 stage/altar receiver positions) using the Schroeder inverse integral method [Schroeder, 1965] and is shown in Figure 5.11. The room measures a spatially averaged RT60 of 2.66 seconds at 1kHz. The chapel, as with most non-studio recording environments does not have ideal noise isolation from outside traffic and other street noises. Measurement of the long-term averaged A-weighted (LA_{eq}) levels showed a 44dBA noise floor inside the church.

The choral ensemble consisted of seven singers, performing in front of the altar. Close microphones (Rode NT5s) were used to obtain as dry a version of the performance as possible



(a) (b) (c)

Figure 5.12: *Microphone Setups: (a) Choir close microphones, (b) OCT 7 and Soundfield MK5, (c) Hamasaki Square.*

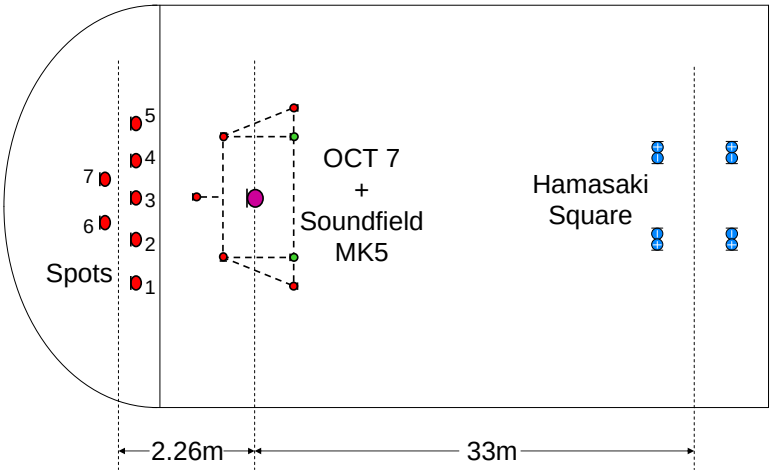


Figure 5.13: *Microphone layout over chapel.*

from each of the singers. The main microphones used to capture the performance were an OCT 7 array (consisting of Schoeps CCM-4 cardioids and CCM-41V supercardioids) and a Soundfield MK5, with the microphones placed at the critical distance of 2.6m from the centre singer. A Hamasaki Square comprised of 4 AKG C-414 microphones was placed high up and at the back of the room at a total distance of 33m away from the main microphones, as shown in Figures 5.12 and 5.13. Recordings were made of several choral pieces with all 7 singers, as well as a solo recording with one singer located at position 3 (centre position in front of the array). The choral pieces recorded were:

- Sicut Cervus - Giovanni Pierluigi da Palestrina
- Sanctus - Thomas Tallis.

After the performances were recorded, impulse responses were captured using the TSP technique with 30 second long sine sweeps. A Genelec 1029A was used for playback of the sweep tones, and measurements were taken for the source at each singer position. Care was taken to ensure that both the height and position of the source matched that of the corresponding singer.

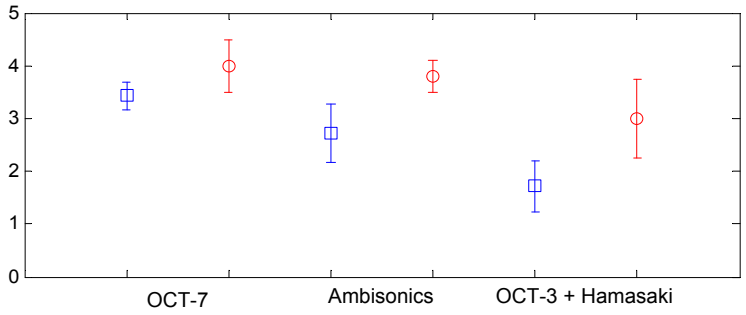


Figure 5.14: *Perceptual difference between actual and virtual audio sample pairs: (1 = Completely Different, 2 = Dissimilar, 3 = Similar, 4 = Very Close, 5 = Identical). Blue □ = μ (Uncompensated), Red ∘ = μ (Compensated), ⊢⊣ = Standard Error.*

For each measurement, recordings were taken using the OCT, Soundfield and Hamasaki arrays. From these measurements a data set of 105 deconvolved impulse responses was obtained. For comparative purposes the recorded impulses were compensated for loudspeaker response only and not for coloration of the room microphones. However, the recorded spot microphone signals, which would later be used as the feeds to the convolution were compensated for the on-axis frequency response of the Rode NT5. The compensation filters were created using the methodology shown in Section 5.3.3.1.

In total, three entire sets of mixes were made for each set of impulse responses:

1. The actual room recordings from the multichannel arrays in the chapel
2. Virtual recordings, created from uncompensated impulse responses from the multichannel arrays
3. Virtual recordings, created from compensated impulse responses from the multichannel arrays.

The recordings were played back in a treated listening room space over a 7.1 'Front-High' layout [DTS, 2007] as shown in Figure 3.10 comprising of 7 Genelec 1029A loudspeakers and a Genelec 7050A sub-woofer. Each of the loudspeakers was calibrated to 79dBC (with +10dB in-band gain on the sub-woofer) at the centre listening position. The listening room is a good monitoring environment with a spatially averaged reverberation time of 0.3 seconds at 1kHz. 10 experienced listeners were chosen for the tests. Each listener was under 35 years of age, of excellent hearing, and well experienced in musical production. The tests were of a paired comparison type and the sequence in which each audio sample pair was played was random. Background noise at the same signal to noise ratio measured in the hall was introduced to each of the virtual recordings to remove any bias towards the natural recordings in the tests. The participants were given full control of the test and were allowed to play back/compare each sample pair at their leisure.

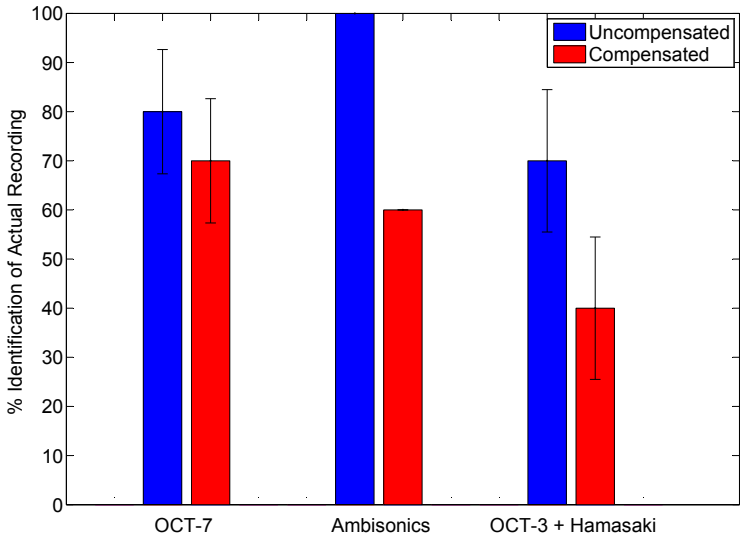


Figure 5.15: *% identification of actual recording against uncompensated (blue) and compensated (red) virtual recordings (95% confidence Interval).*

The results presented in Figure 5.14 show the extent of the perceptual difference between the natural and uncompensated virtual recordings, with the greatest difference and largest deviation from the mean μ occurring for Ambisonics. In contrast, with the compensated recordings, the distinction between the actual and virtual recordings becomes quite unclear, and higher values of μ are obtained in all cases.

Each listener was then asked to identify which was the natural recording and which was the virtual uncompensated one. The results, shown in Figure 5.15 showed that all listeners could identify the actual Ambisonics recording, with 70% identifying the actual OCT-3-Hamasaki recordings and 80% the natural OCT-7 recording. However, when the actual recordings were compared against the virtual compensated recordings, each listener found it more difficult to completely distinguish between the two, in particular the OCT-3 + Hamasaki Square.

It can be concluded from this initial study that

- The timbral differences between compensated and uncompensated virtual acoustic recordings can be perceived such that identification of a virtual recording is possible.
- The inclusion of a highly diffuse field blurs the distinction between compensated and uncompensated recordings. The results obtained for the OCT-3 and Hamasaki square demonstrate this, since the role of the Hamasaki square in this case is to introduce a more diffuse field to the mixes.
- Perceptual differences between compensated virtual recordings and real recordings exist, indicating the existence of other parameters that must be considered before true emulation of acoustic events is possible.

5.4 Evaluation of Virtual Acoustic Recording

Based on the preliminary analysis of the previous section, we will now perform further investigations into the objective and subjective differences between actual and virtual acoustic recordings. We approach this analysis with the following hypotheses:

1. The perceptual differences between actual acoustic recordings and virtual acoustic recordings change with the direct to reverberant ratio.

2. The subjective differences between actual and virtual acoustic recording (other than coloration in the convolution chain) can be quantified as objective attributes.

3. Actual recordings are always preferred to virtual recordings.

In the same manner as the work presented in Section 5.3.3.2, natural room recordings will be directly compared to close microphone recordings of the same performance, convolved with spatial impulse responses recorded from a soundfield microphone. Objective comparison of the recordings through binaural measurements is followed by listening tests to investigate the perceptual attributes between the actual and virtual audio.

5.4.1 Database Acquisition

In order to investigate the hypotheses presented in the previous section, a series of recordings was taken in Test environment 1. These recordings consisted of short separate performances from a female singer and a violin player. 4 positions of the performer were recorded, located 1m, 2m, 4m and 8m into the room with reference to the microphone array, as shown in Figure 5.16. An Ambisonics soundfield microphone and a Neumann KU100 mannequin were used to record the performances at the array position. An OCT array was also set up, but was not used in this study. For the spot microphone recording an AKG C-414 was used. At each position the vocalist and violin player individually performed short 30 second pieces which were

- Vocal piece: Perfume de Carnaval

- Violin piece: Brahms, Hungarian Dance No. 5

Reference A-weighted SPL measurements of the performers playing/singing a tuning note were taken at the 1m position, for calibration at playback. For consistency, the performers were presented with a tuning note before each performance, and a click track consisting of pulsed fundamental tones.

After the performances were recorded at each position, impulse responses were captured using the TSP technique with 30 second long sine sweeps. A Genelec 1029A was used for playback of the sweep tones, and measurements were taken for the source at each of the performance positions. Again, the height and position of the measurement loudspeaker matched that of the

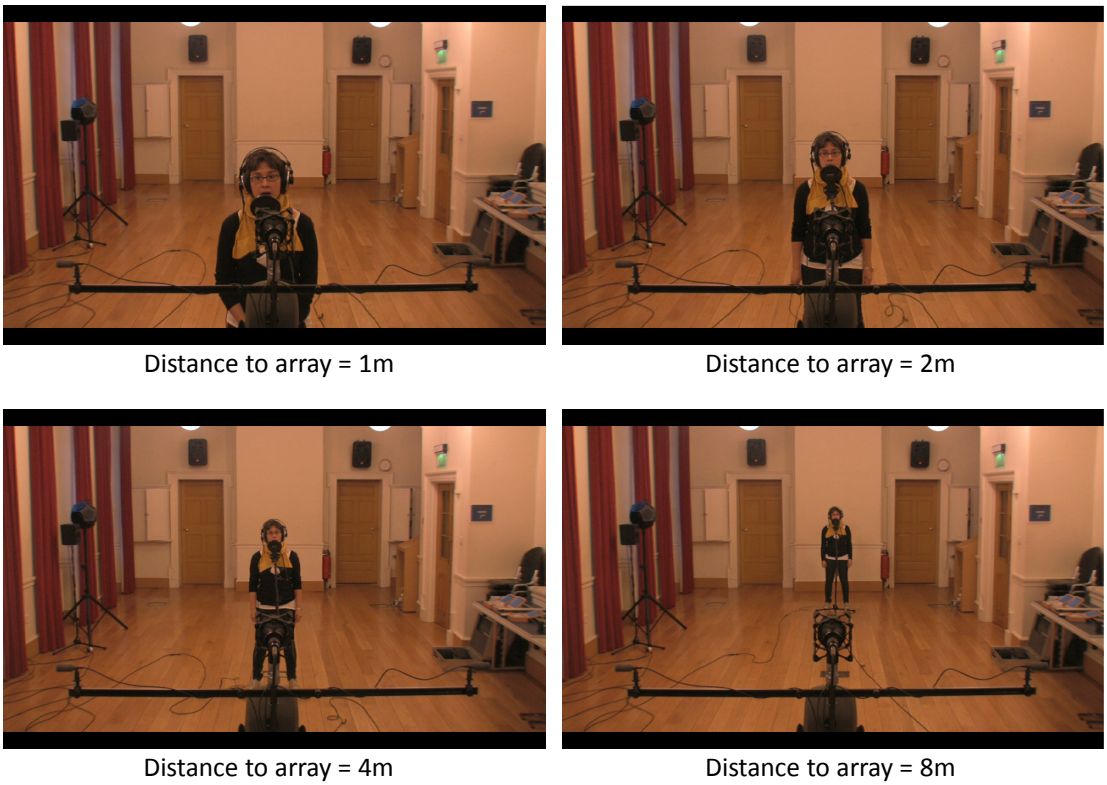


Figure 5.16: *Female vocalist positioning with respect to microphone array.*

corresponding performer. For each measurement, recordings were taken using the Soundfield and binaural microphones. The recorded impulses were then compensated for loudspeaker response as well as the spot microphone signals using the method described in Section 5.3.3.1.

The test environment itself was not completely acoustically isolated from the outside world, and had a long-term averaged A-weighted level of 42dB. Due to this intermittent external noise, several performances had to be repeated until a satisfactory result was achieved.

5.4.2 Binaural Comparison of Reproduced Recordings

The objective differences in these recordings can be first investigated through examination of the signals recorded in the test environment with the dummy head and the direct source signal convolved with the impulse response measurements taken using the dummy head. The recordings were first compared using the normalized Interaural Cross Correlation Function (IACF) of equation 2.14. Conventionally, this measure is implemented on sets of binaural impulse responses, but as suggested by Mason [Mason et al., 2003], it can also be used as a continuous measure over the length of a musical passage.

For each of the recordings the IACC was computed in 35ms Hanning windows with 50% overlap. The results at 1m for both the actual and virtual recordings, as well as their difference, for 10 seconds of the female singing example are plotted in Figure 5.17. It can be seen that the

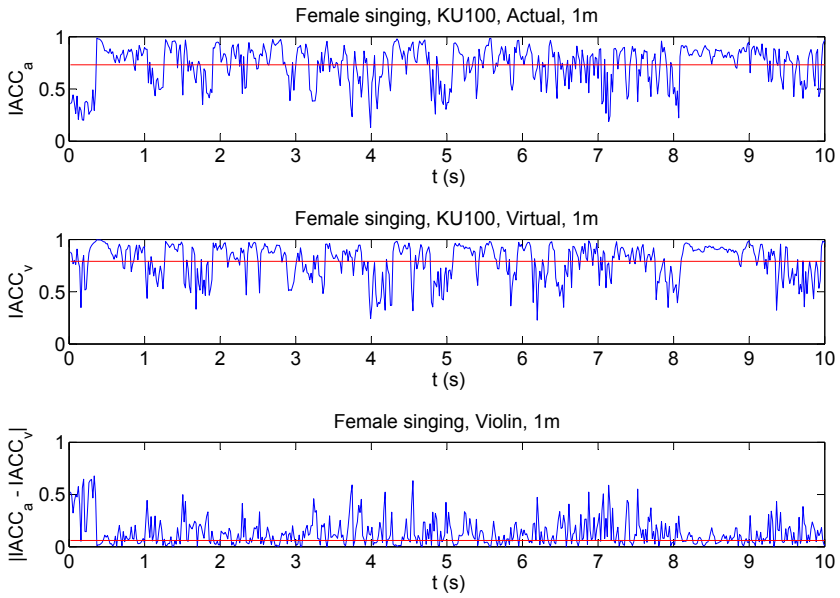


Figure 5.17: *Running-IACC and IACC error for actual and virtual binaural recordings of female vocals at 1m.*

results are very close in terms of the magnitude of the IACC but there exist definite differences between the actual and virtual recordings. The fluctuations in the IACC are due to the musical phrasing, where changes from high to low correlation occur in between each phrase (i.e. as the amplitude of the direct sound rises and falls). If the actual and virtual recordings are completely identical, no difference should be exhibited here. However, this is not the case, as is shown by the running IACC error function $IACC_e = |IACC_a - IACC_v|$. The average error in the IACC for female singing at 1m is approximately 0.08, indicating that the actual recording may be perceived as having a different source width from the virtual. We note that in fact $|1 - IACC|$ is measured higher for the actual source. As we move further into the diffuse field this IACC error decreases. This is shown in Figure 5.18. This trend is also exhibited with the violin samples. However, here the magnitude of the error is greater in the near field than in the case of female singing, with an average value of 0.27. As a result, a perceptual difference will also exist between actual and virtual recordings in the context of overall source width with this instrument.

We can also use the IACC-PHAT method introduced in Chapter 2 to determine the source movement. As with the running IACC measurement, we again employ 35ms windowing in the estimation of the IACC-PHAT. We also low pass filter the signal such that we are only estimating the ITD below 700Hz. In each window, we extract the lag of the peak IACC value indicative of the interaural time delay.

The resultant continuous ITD for the violin at 1m is shown in Figure 5.21. The ITDs for the actual recording show significant fluctuations, which again occur mainly between the musical phrases. Such varying fluctuations are not seen in the case of the virtual recording however. This is due to the fact that convolution with a binaural impulse response yields a completely

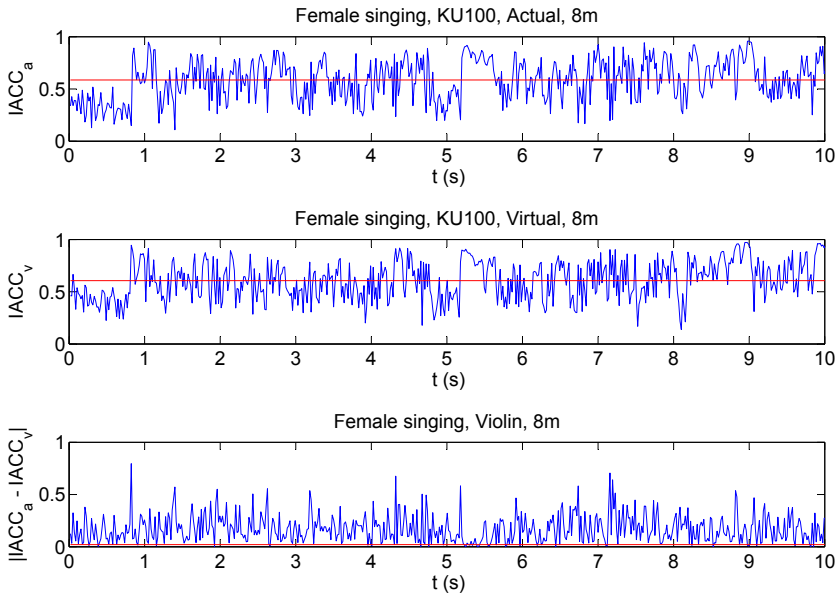


Figure 5.18: *Running-IACC and IACC error for actual and virtual binaural recordings of female vocals at 8m.*

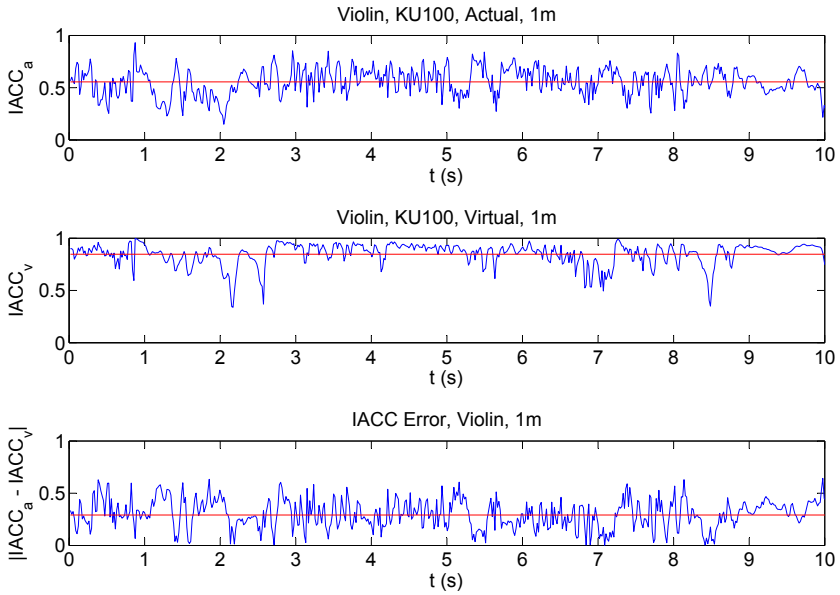


Figure 5.19: *Running-IACC and IACC error for actual and virtual binaural recordings of violin at 1m.*

static virtual image. Furthermore, the impulse response measurement can be considered as a time-frequency average (over the length of time it took to create the measurement) which assumes linear time invariance. Overall the average ITD difference between both the actual and virtual recordings is small.

At a distance of 8m, the virtual ITD again displays little fluctuation. Peaks appear in the running-ITD, again due to the musical phrasing. However, according to the studies of Klumpp

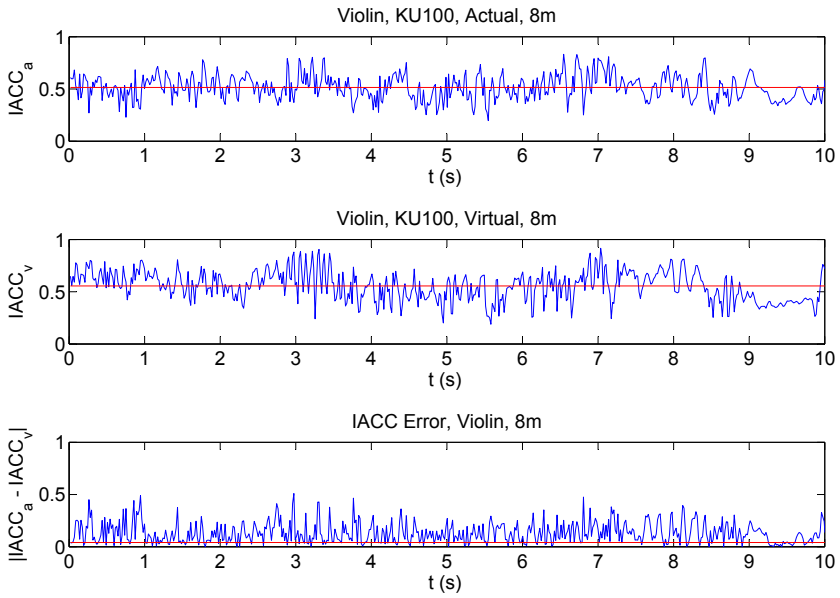


Figure 5.20: *Running-IACC and IACC error for actual and virtual binaural recordings of violin at 8m.*

and Eady [Klumpp and Eady, 1956], outlined in Section 2.2.4.2, these small time delays will be perceptible, since the human auditory mechanism can detect changes in the ITD of the order of 9μs in low frequency bands. As was also shown from Figure 2.11, ITDs corresponding to as much as 50μs relate to approximately 5° difference in the frontal plane, depending on the temporal and spectral properties of the source.

It is also useful to objectively look at the level differences between each binaural pair. This will indicate the high frequency directional characteristics of the source. Measurement of the Interaural Level Difference (ILD) is accomplished by estimating the power spectrum ratio of the ear signals above 700Hz. Figure 5.23 shows the ILDs for the actual and virtual compensated violin recordings at 1m. Also presented here is the magnitude of the ILD error in dB between the actual and virtual recordings. Once again, we see fluctuations in error function, but the average magnitude of this error is less than 1dB. As we move into the reverberant field this average error changes between 0 and 2dB. Ideally, the diffuse field should result in an averaged ILD of 0 for both the actual and virtual recordings. Although this is not the case here, we note that the difference between the actual and virtual ILD measures becomes smaller with greater window sizes. Nevertheless, the results presented here indicate that differences in source position may be apparent over the duration of the musical passage due to ILD.

Overall, the averaged binaural results for female vocals and violin at each distance can be found in Tables 5.2 and 5.3 respectively. For the IACC, we conclude that errors exist between the actual and virtual recordings, which can lead to differences in the perceived apparent source width, but these errors are reduced the further we move into the reverberant field. The ITD in both cases does not lead to significant localization differences, but it is noteworthy that the

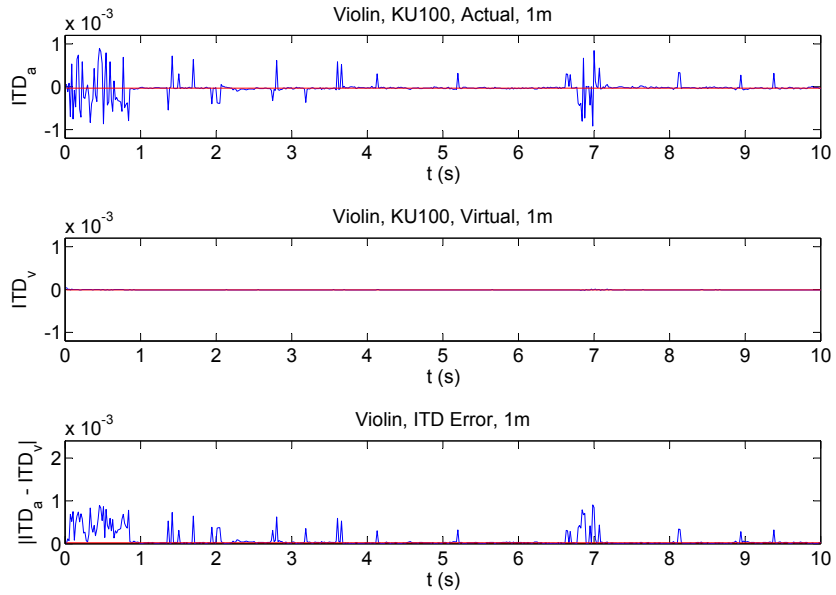


Figure 5.21: *Running-ITD and ITD error for actual and virtual binaural recordings of violin at 1m.*

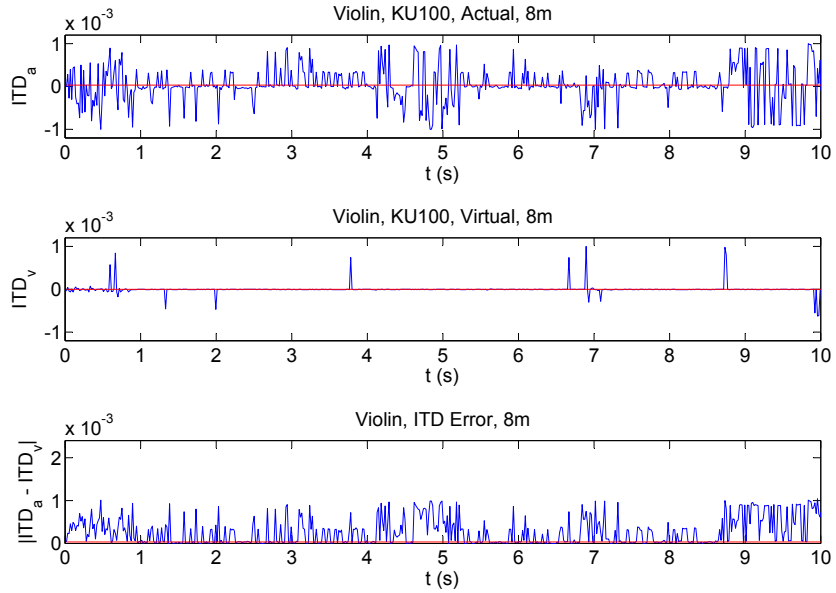


Figure 5.22: *Running-ITD and ITD error for actual and virtual binaural recordings of violin at 8m.*

errors are somewhat greater in the case of the violin. The ILDs in both the actual and virtual recordings change between 0 to 2dB over the different distances. This error does not follow any significant trend and appears highly dependent on the position in the room and the source type.

It can be concluded that small fluctuations in the IACC, ITD and ILD contribute to perceived changes in the convolved source. Since such errors exist after compensation for coloration in

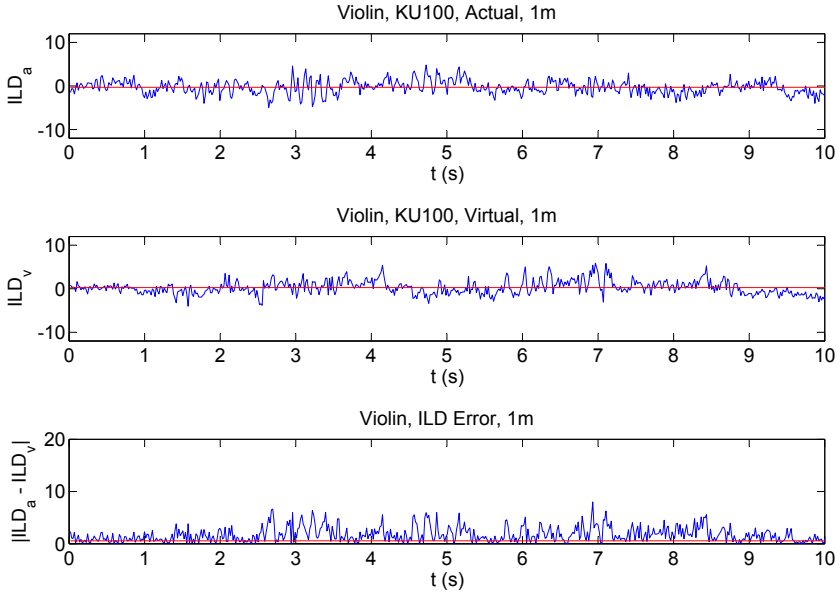


Figure 5.23: *Running-ILD and ILD error for actual and virtual binaural recordings of violin at 1m.*

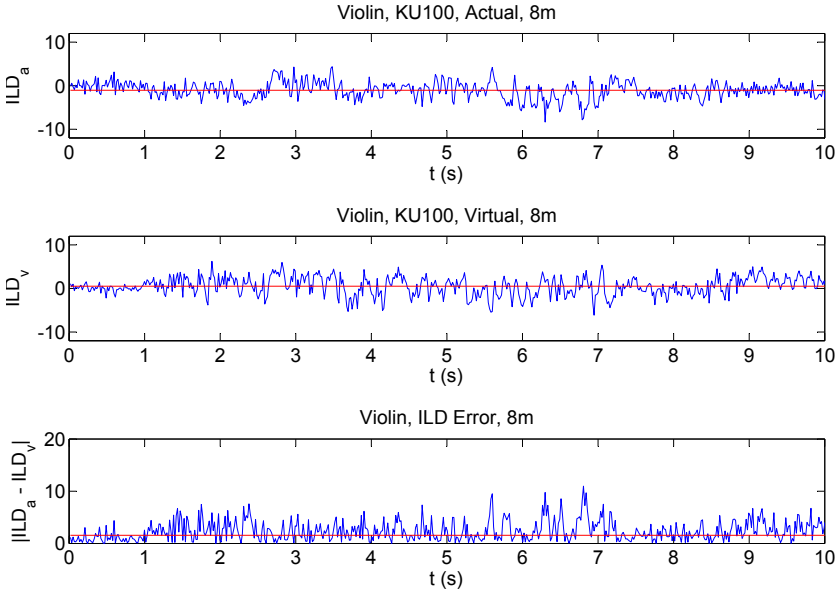


Figure 5.24: *Running-ILD and ILD error for actual and virtual binaural recordings of violin at 8m.*

---

the convolution chain, then other factors, such as the directional response of the performer are responsible for these binaural differences. Furthermore, it is necessary to determine whether these discrepancies can be perceived in the context of music performances, and whether they contribute to an overall subjective preference of actual recordings over virtual ones.

	Actual recording	Virtual recording	Error
IACC, 1m	0.73	0.79	-0.06
IACC, 2m	0.64	0.68	-0.04
IACC, 4m	0.53	0.58	-0.05
IACC, 8m	0.59	0.61	-0.02
ITD (x 10^{-5})s, 1m	-2	-1	-1
ITD (x 10^{-5})s, 2m	-1	0	-1
ITD (x 10^{-5})s, 4m	0	0	0
ITD (x 10^{-5})s, 8m	1	0	1
ILD (dB), 1m	-0.37	-1.21	-0.84
ILD (dB), 2m	-1.19	-1.22	0.03
ILD (dB), 4m	-0.08	-1.27	0.39
ILD (dB), 8m	-1.28	-1.12	-0.16

Table 5.2: *Averaged binaural analysis of 10 seconds of female singing performance in Test environment 1*

	Actual recording	Virtual recording	Error
IACC, 1m	0.56	0.85	-0.29
IACC, 2m	0.53	0.62	-0.08
IACC, 4m	0.50	0.56	-0.06
IACC, 8m	0.51	0.56	-0.04
ITD (x 10^{-5})s, 1m	-3	-1	-2
ITD (x 10^{-5})s, 2m	-6	-1	-5
ITD (x 10^{-5})s, 4m	-3	0	-3
ITD (x 10^{-5})s, 8m	-3	0	-3
ILD (dB), 1m	-0.31	0.26	-0.57
ILD (dB), 2m	-0.09	-2.03	-1.93
ILD (dB), 4m	-0.01	-1.83	1.81
ILD (dB), 8m	-1.04	0.51	-1.56

Table 5.3: *Averaged binaural analysis over 10 seconds of violin performance in Test environment 1*

5.4.3 Subjective Experiment IV.A: Subjective Attributes of Actual Vs. Virtual Convolution Based Recordings

In this experiment, the subjective attributes of the anomalies identified with the binaural recordings are investigated. For this we will utilize the musical examples of the binaural analysis, this time decoded from the B-Format microphone recordings for 1st order Ambisonics reproduction. An 8-channel regular array was set up in test environment 2 for the tests. A shelf filter Ambisonic decode was utilized with maximized r_v at low frequencies and maximized r_e at high frequencies. Three types of recordings were assessed in this study:

1. Actual 1st order Ambisonic recordings of violin and female singer respectively
2. Virtual 1st order Ambisonic recordings of violin and female singer respectively created from measured spatial impulse responses in the recording environment
3. Virtual 1st order Ambisonic recordings of violin and female singer respectively created from artificial spatial impulse responses.

The virtual B-Format recordings were synthesized from spatial impulse response measurements of the test environment and the direct field performance recordings. The artificial spatial impulses responses were created using the image source method from a simple model having the same geometric parameters as the true test environment. Fourth order reflection image source impulse responses were created and added to a randomized reverberation tail. The image source impulses were rendered to give similar reverberation decay rates and energy in each octave band in comparison to the room impulse response at each position. The ISO-3382 [ISO, 2009b] acoustical parameters for the B-Format omnidirectional component for both the reference acoustical measurement and image method simulations are shown in Appendix G.

A listening session was set up prior to the formal testing phase to determine a consensus vocabulary with which to describe the subjective attributes under test. Here, a small listening panel consisting of 5 experienced listeners, were individually allowed to play through all the test material using the custom software shown in Figure 5.25. They were then instructed to write down how they felt the different source material differed for each other.

After analysis of each participant's answers, duplicate attributes relating to the binaural analysis of Section 5.4.2 were grouped together, and in a subsequent group session the listening panel concluded that the following descriptors should be given to each subject prior to the test:

- **Reverberance:** *The duration of the reflected sound. Think of the sound that continues in a room after a hand clap. Is there more or less of this reverb present in the samples? Listen to the tails of notes, ends of phrases etc.*
- **Source-Width:** *The space a source occupies within the recording. Is the source bigger or smaller than in the reference sample?*

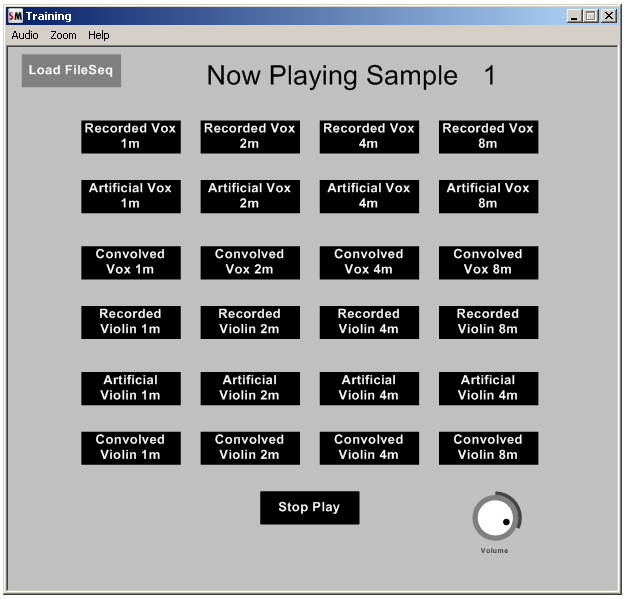


Figure 5.25: *Custom test software for derivation of subjective attribute descriptions.*

- **Source-Clarity:** *How well-defined and intelligible are the source sounds? Are they more or less clear than the reference? Does they sound more muffled than the reference?*
- **Source-Movement:** *Musicians make small movements when performing. Are these movements more or less pronounced when compared to the reference?*
- **Natural-Timbre:** *Accurate and realistic tonal quality. In overall tonality, are the samples more or less realistic than the reference?*

The tests were designed as grading tests, where each subject was asked how much they thought a particular attribute changed in the test samples A, B and C in comparison to the reference. A 7-point hedonic categorical scale, ranging from 'Far Less' to 'Far More' was used for the rating. In a similar manner to the grading tests of Section 4.2.5, subjects were given the following guidelines regarding the ratings,

- Slight difference between two samples: Difference between samples is at least 0.5
- Moderate difference between two samples: Difference between samples is at least 1
- Strong difference between two samples: Difference between samples is at least 2

In these tests, the artificial samples were colored with background noise recorded from the original acoustic environment, so as to not introduce any unnecessary bias in the testing of the subjective attributes concerned. Prior to recording the data from the tests, each subject was allowed to familiarize themselves with the sample material and the grading scales were explained.

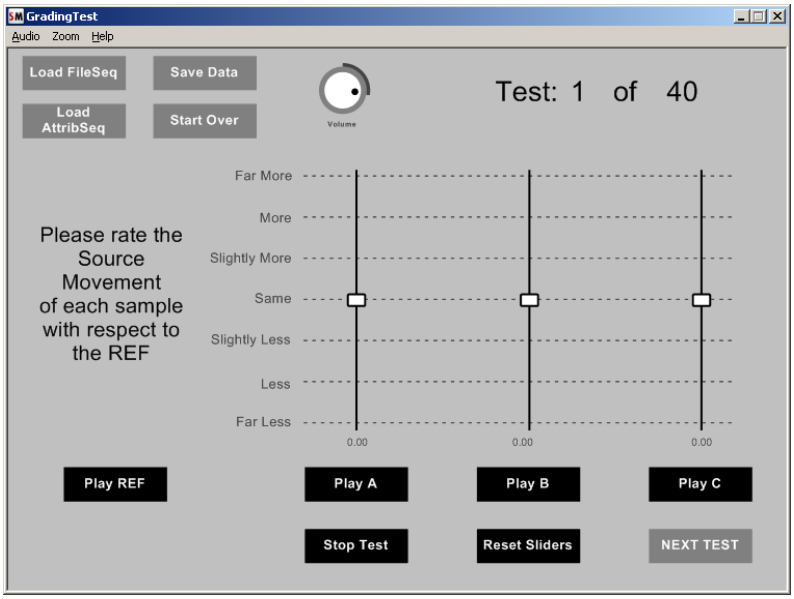


Figure 5.26: *Custom test software for subjective comparison of actual and virtual recordings.*

Listeners were allowed to play each sample for as long and as many times as they wished before deciding on the grades and moving to the next phase of the test. Unknown to the subject, the reference sample used was always the original acoustic recording. However, subjects were aware that one of the samples was the same as the reference, hidden as samples A, B or C. The subjects were therefore forced to rank one of the samples as the 'same' as the reference, before the test software allowed them to proceed to the next test iteration. In total, there were 40 grading tests (2 sources, 4 positions and 5 attributes).

5.4.3.1 Results

The average scores for each listening test attribute are shown in Figure 5.27. In these plots, the y-axis labels can be read as follows:

- R: Identification of the reference
- A: Actual acoustic recording
- I: Image source method generated recording
- V: Virtual acoustic recording using measured spatial impulse responses.

For each of the tested segments the reference source was correctly identified by all listeners. For the 'reverberance attribute' in the case of the violin, the samples are not statistically different. In the case of the female speech however, the subjects perceived the image source recording as being more reverberant, and the actual and virtual acoustic recordings as being equivalent.

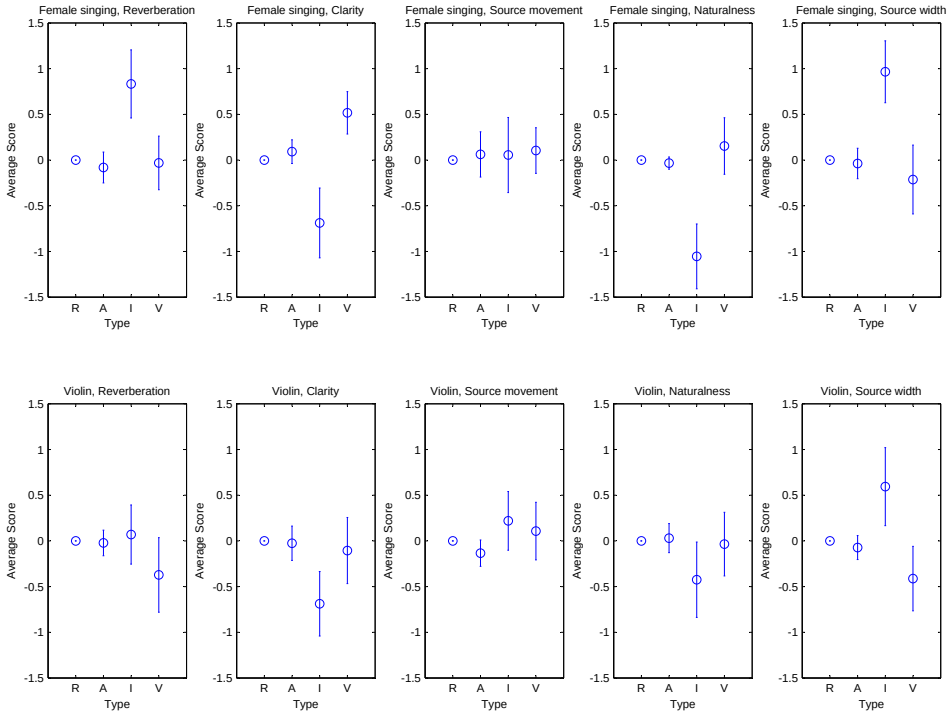


Figure 5.27: *Average scores for actual vs virtual recording (95% Confidence Interval).*

In the case of 'clarity', the image source simulation was perceived as being less clear than the other sources. In the case of female singing, the virtual acoustic recording showed a significant improvement in the clarity over the actual recording. For both source types, no change in source movement is perceived overall. This corroborates the findings of the ITD binaural study. No significant difference is found between the actual and virtual source recordings in terms of natural-timbre. The image source method however is rated to have a significant difference in natural-timbre. Finally, the source width is perceived as being larger for the image method source, but smaller for the virtual acoustic recording for the violin sample, in comparison to the reference source. This is not surprising since the image source method simulation assumes an omnidirectional source, whereas the virtual acoustic recording is utilizing the directional properties of the measurement loudspeaker and not the natural directional characteristics of the original source audio. This result is exactly as was predicted in the binaural assessment.

The overall capabilities of the virtual acoustic recording methodology are therefore summarized in Table 5.4. The main distinction between the actual and virtual acoustic recordings lies with the source width, which is related to the directional response of the original acoustic source and the measurement loudspeaker. A strategy is therefore required such that the directional properties of the original source can be modeled via loudspeaker measurements. This is addressed in Section 5.5.

Matching Parameter	Actual recording	Virtual recording	Image Source Method
Source Width	Yes	No	No
Reverberance	Yes	Yes	Yes
Clarity	Yes	Yes	No
Naturalness	Yes	Yes	No

Table 5.4: *Summary of correctly perceived attributes between actual and virtual sources*

5.4.4 Subjective Experiment IV.B: Subjective Preference of Actual Vs. Virtual Convolution Based Recordings

It is important to note that all of the aforementioned attributes are no real indication for overall listener preference. Thus in a supplementary listening test, each listener was also asked which of the audio sequences they preferred. The interface to the custom designed software used for this forced choice test is shown in Figure 5.28. The samples were of the same musical presentations as in the previous studies. However, the background noise in the virtual recordings that was present in the previous study was not included here, since in real-world applications of the method this background noise would not be present. Each of the test stimuli were randomly mapped to sample buttons A, B and C during the tests. The test was repeated for each subject at each of the different distances and 10 subjects participated in the tests.

The results of this study are shown in Figure 5.29. For the female vocals, there is a strong preference at all distances for the virtual acoustic recording, with over 60% of the vote. No listeners chose the artificial recording. For the violin however, the actual recording was preferred at all distances except at 8m. It is interesting to note that as we move into the diffuse field, the % vote for the actual recording reduces. At 1m, the actual recording is significantly different from the artificial and virtual acoustic recordings. This indicates that correct near field auralization is an important aspect of listener preference. This is not the case at 2m (and beyond) however, where highly diffuse recordings are preferred using either natural or virtual acoustic recordings and not image source representations.

5.4.4.1 Conclusions on Actual Vs. Virtual Recording Assessment

In response to the hypotheses presented in Section 5.4, it was shown

1. The perceptual differences between the actual and virtual acoustic recordings do change with the direct to reverberant ratio. This is particularly noticeable with the source width and ILD measures.

2. The subjective differences between the actual and virtual acoustic recordings can be quantified as objective attributes. The source width is directly related to the directional characteristics of the source and its interaction with the room. The coloration effects of the

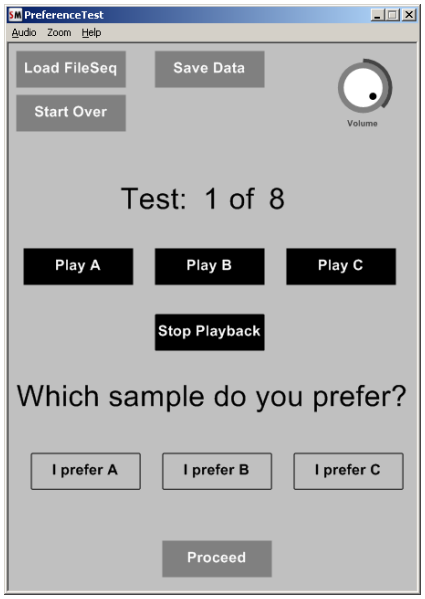


Figure 5.28: *Test software for listener preference tests.*

convolution chain are related to the perceived timbre. Finally, the ITD and ILD estimates are related to the perception of source movement.

3. Actual recordings are not always preferred over virtual recordings. The results presented here favor the virtual recording approach due to the reduction of external noise and the increase in source clarity. Since the virtual acoustic recording does not suffer from background noise issues, it is an ideal method of recording in reverberant environments that do not have good acoustic isolation. Preference towards the actual recordings occur when the directional characteristics of the source are not correctly modeled. Because of this, the technique works well for voice sources, but less so for sources with complex directional responses, such as the violin.

Furthermore, the presented experiments quantify the differences between actual recordings, artificially generated recordings and virtual acoustic recordings compensated for coloration effects in the convolution chain. It is shown that the virtual acoustic recording approach favors well against the natural acoustic recordings. The perceived timbre, level of reverberation, and source movement are equivalent to natural recordings. There were few listeners who preferred the image method auralizations over the natural or virtual acoustic recordings. This is attributed to the directional characteristics of the source audio, which is confirmed by the binaural measurements of source width.

We conclude that virtual acoustic recording shows remarkable potential for perceptually equivalent recordings provided the directional characteristics of the source are taken into account. A novel approach to this issue will now be addressed.

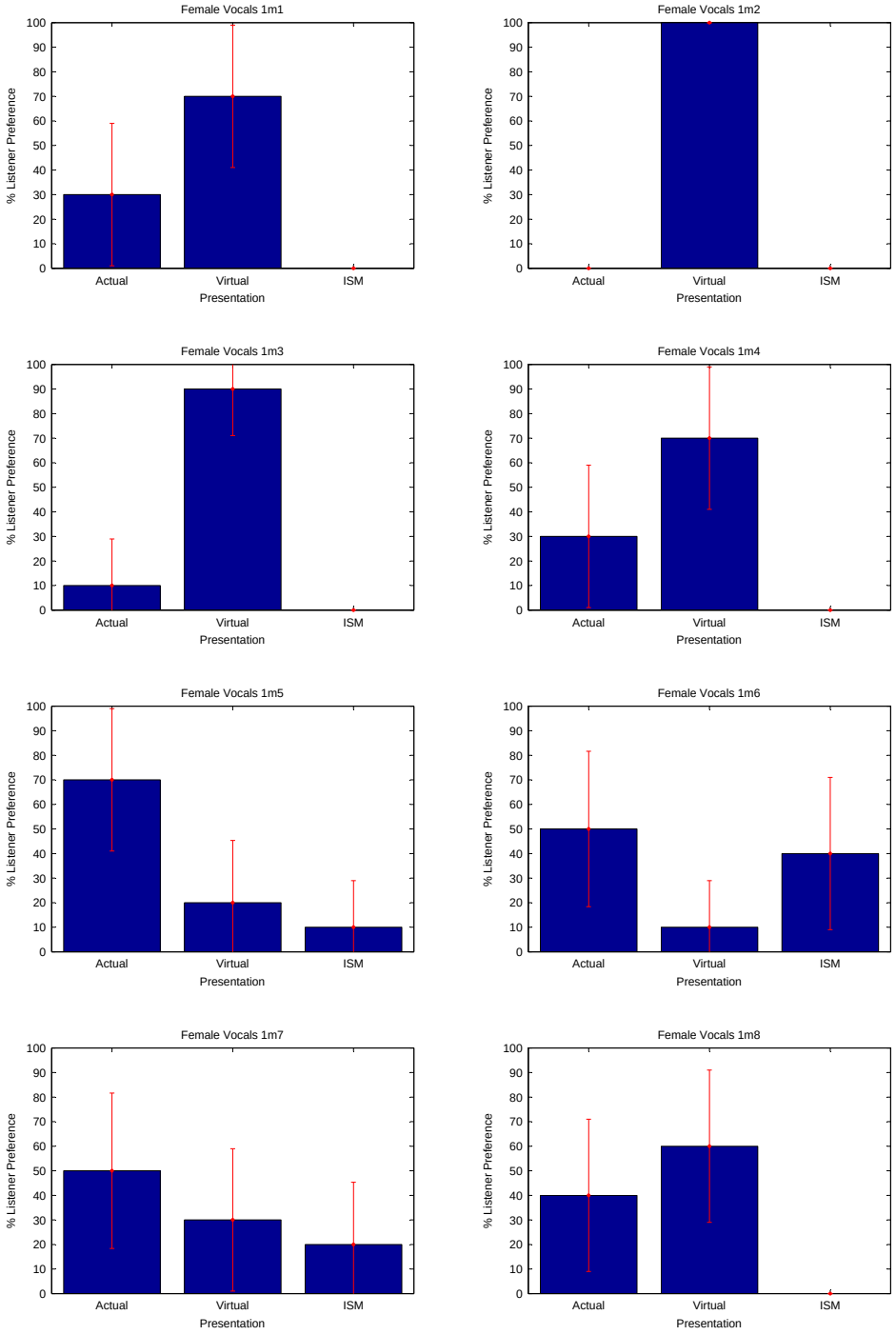


Figure 5.29: *% Listener preference between Actual, Virtual and Image Source presentations for each source type and position (95% Confidence Interval).*

5.5 Directivity Enhancement for Virtual Acoustic Recordings

Based on the conclusions of the previous section, it is clear that for any virtual acoustic recording to be convincing, the directional properties of the source audio must be considered. Several approaches to measuring and synthesizing source directionality have been proposed on the literature, primarily with regard to wave field synthesis reproduction. The capture of source directivity is traditionally achieved using arrays of microphones surrounding a performer and databases of such measurements are available [Physikalisch-Technische-Bundesanstalt, 2009]. Directivity filters can then be applied to a single monophonic recording of a performance to simulate the change in frequency response with source/listener movement. A simple approach has been proposed by Giron [Giron, 1996], where the interference of several monopole virtual sources is used to synthesize the directivity of real sources. However, the resulting frequency dependent directivity does not behave like that of real world sources. A further approach is that the array of microphones used to capture the directivity measurements can actually be used to capture the performance (in an anechoic chamber) entirely, and virtual loudspeakers can be synthesized at reproduction using monopoles or virtual cardioids [Jacques et al., 2005]. This is not a practical solution to a real performance situation however. Another promising method is the decomposition of the directional response into spherical harmonics, which has been proposed for computational based auralization in numerous papers most notably by Menzes [Menzies, 2007] and Spors [Ahrens and Spors, 2007].

However, it has been found by Jacques et al. [Jacques et al., 2005] that the directivity characteristics of a natural musical instrument do not have to be completely objectively identical to the original response, provided that the end-listener is not familiar with every single aspect of the *particular* instrument's directional response. A close approximation is generally sufficient to create a plausible reproduction. Kessler [Kessler, 2005] has broadly outlined a method of approximating source directivity by taking multiple spatial impulse response measurements at the same position, but at different loudspeaker rotations. An approximated directional characteristic can then be derived by simply exciting the room in different directions from the source position, capturing the SRIRs in each source direction and combining the impulse responses in a given ratio for auralization. Here, we too utilize a similar approach to Kessler's, with the exception that we employ a least-squares optimisation to the measured directional signals and the corresponding SRIRs, such that a plausible directional characteristic is obtained.

5.5.1 Source Simulation Radiators

Consider the loudspeaker layout given in Figure 5.30. Let us assume that we have measured the directional characteristic of each of the loudspeakers, which results in three sub-cardioid radiators, angled 120° from each other, as shown in Figure 5.31(a).

We will define the matrix containing the directivity vectors of each of the loudspeakers as $\mathbf{D} = \begin{pmatrix} \mathbf{d}_1 & \mathbf{d}_2 & \mathbf{d}_3 \end{pmatrix}$. The directivity of our virtual source must match that of real source

directivity measurements $\mathbf{p}$ such that

$$\mathbf{D}\begin{pmatrix}\alpha_1\\ \alpha_2\\ \alpha_3\end{pmatrix} = \mathbf{p} \tag{5.22}$$

where $\mathbf{a} = \begin{pmatrix}\alpha_1 & \alpha_2 & \alpha_3\end{pmatrix}^T$, the gain factors to be applied to the loudspeakers such that the desired directional characteristic is obtained. Thus if

$$\mathbf{D}^T\mathbf{D}\begin{pmatrix}\alpha_1\\ \alpha_2\\ \alpha_3\end{pmatrix} = \mathbf{D}^T\mathbf{p} \tag{5.23}$$

then we can solve for $\mathbf{a}$ with

$$\mathbf{a} = \begin{pmatrix}\alpha_1\\ \alpha_2\\ \alpha_3\end{pmatrix} = (\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{p} \tag{5.24}$$

For example, imagine we wish to approximate a source with a supercardioid directivity. Starting with sub-cardioid responses, the above optimization process will then yield $\mathbf{a} = \begin{pmatrix}-0.4211 & 1.0000 & -0.4211\end{pmatrix}$, where a negative gain represents a phase inversion. The resultant supercardioid is then obtained as shown in Figure 5.31(c). Also shown here is the directional responses of 3 Genelec 1029a loudspeakers arranged as in Figure 5.30. Optimisation then gives the close to supercardioid response, also shown in Figure 5.31(c). The gain factors can be used in two ways: the first is to determine the optimized gains from anechoic or simulated free-field measurements of both the instrument directivity vector $\mathbf{p}$ and the loudspeaker matrix $\mathbf{D}$ prior to the impulse response measurements. The correct ratio can then be used to play back the TSP signals for impulse response measurements. The second method, which is far more flexible, is to take the spatial impulse response measurements from each of the three loudspeakers individually. At post-production, the correct ratio of the SRIRs is exactly the same as the ratio of the loudspeaker gains, i.e. an impulse response h captured at position i is the linear summation of three impulse responses due to the separate excitation of each loudspeaker in the source simulation array, given by

$$h_i(t) = \alpha_1\ h_{s1}(t) + \alpha_2\ h_{s2}(t) + \alpha_3\ h_{s3}(t) \tag{5.25}$$

However, true directional excitation of complex sources is also frequency dependent. Consequently, we propose that the above optimisation be performed in perceptual frequency bands, such that a more realistic estimation of the source directivity is obtained.

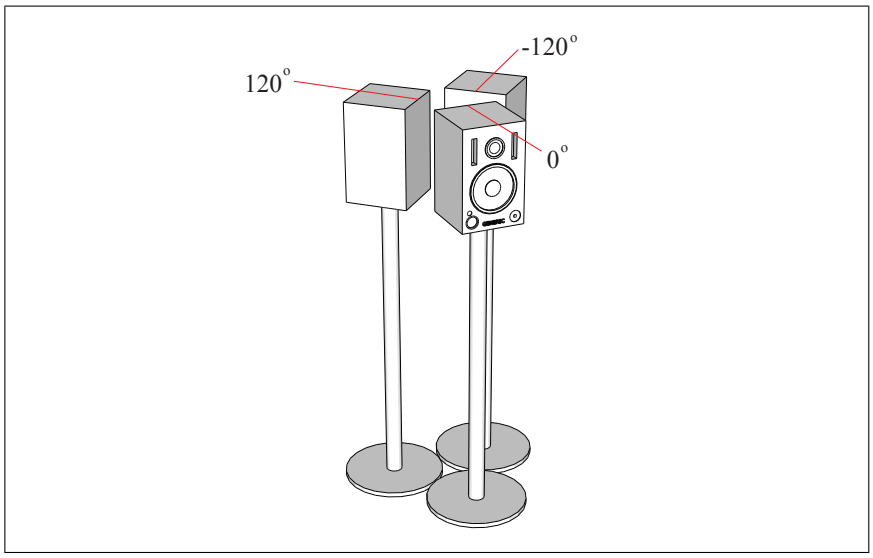


Figure 5.30: *Source simulation radiator layout.*

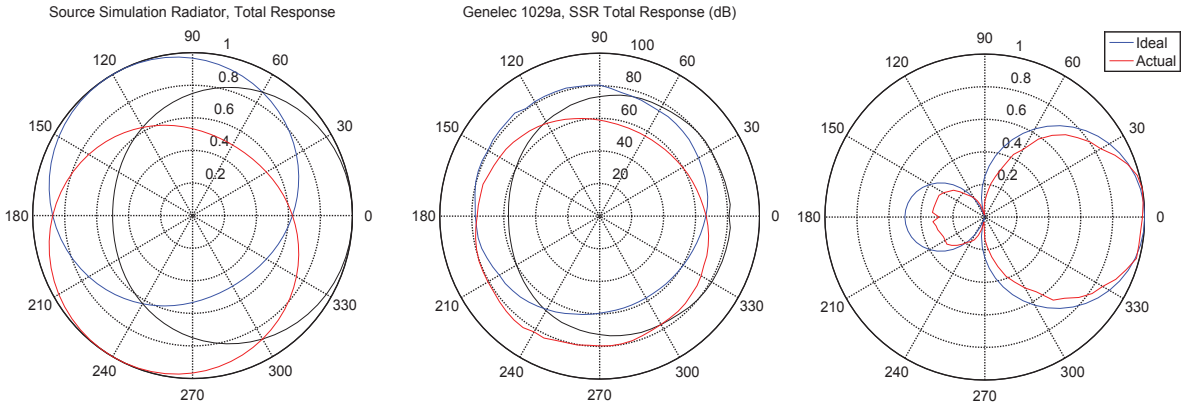


Figure 5.31: *Source simulation radiation example: (a) Ideal sub-cardioid response (b) Measured SSR response using Genelec 1029a loudspeakers (c) Approximated source directivity.*

5.5.2 Measurement of Source Radiation

For this study, power spectrum measurements of the directional characteristics of a violin and female vocals were taken in controlled studio environment. Horizontal-only directional measurements were taken using nine equi-spaced Rode NT5 condenser microphones arranged in a 1m radius semicircle as shown in Figure 5.32. We justify utilizing only the horizontal plane since, as was shown in Chapter 2, it is our most sensitive plane of hearing, and also since horizontal only reproduction is the main focus of this work. However, for more accurate measurements that include directional responses over a sphere, or for periphonic reproduction, more measurement



Figure 5.32: *Measurement of horizontal violin radiation characteristic.*

microphones are required.

The performers were asked to first play scales to a click track in the forward direction and then to repeat the performance turned 180°, so that a full horizontal dataset was measured. The effective sound pressure measured at each microphone was then calculated in 25 rectangular critical bands based on the Bark scale, and the data was linearly interpolated to give 360 horizontal measurements. The resultant directional spectra are shown as the top plots in Figures 5.33 and 5.34. We see in the case of the female vocals that the synthesized power spectrum (shown as the bottom plot in Figures 5.33) for each direction and in each band is well matched. The violin response, shown in Figure 5.34 is more irregular than the female voice, but is still represented well by the source simulation process. We note the synthesized response is smoother than the true violin response.

An important aspect of filtering direct-source audio with the final SRIR is de-emphasis of the directional responses. Calculation of the gain factors to satisfy the directional response functions of Figures 5.33 and 5.34 assumes that the magnitude spectrum of the source audio for convolution is flat, and that application of the gain factors will yield the appropriate directional *magnitude* response. This would require the direct source signal to be filtered with the inverse magnitude response pertaining to the angle at which the direct-source microphone picked it up, thus creating a neutral magnitude spectrum ready for the direction-dependent characteristic associated with the optimized impulse responses. However such an undesirable processing stage can be avoided if we normalize the gains in each band according to the gains due to the centre loudspeaker in the SSR array. That is, the directivity of the source is simulated by keeping the

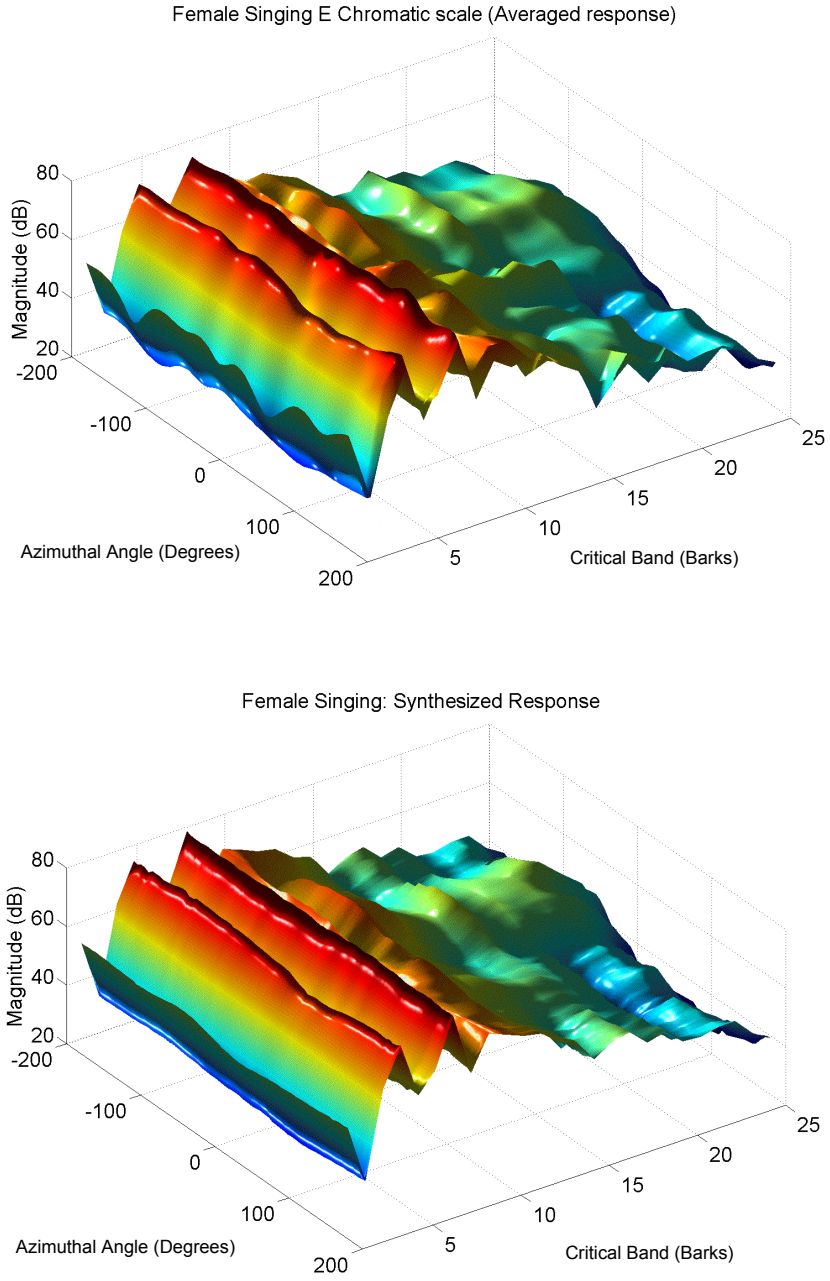


Figure 5.33: *Female singing directional response and synthesized response.*

SRIRs due to the centre loudspeaker constant and adjusting the relative levels (in each band) of the SRIRs due to the rear loudspeakers. This assumes that the forward directional characteristic of the SSR array yields the correct magnitude response, which is justified if the source audio is already compensated for the on-axis directional response of the loudspeaker (in the removal of convolution chain coloration). Thus, the estimated gain reduction factors in each band for each

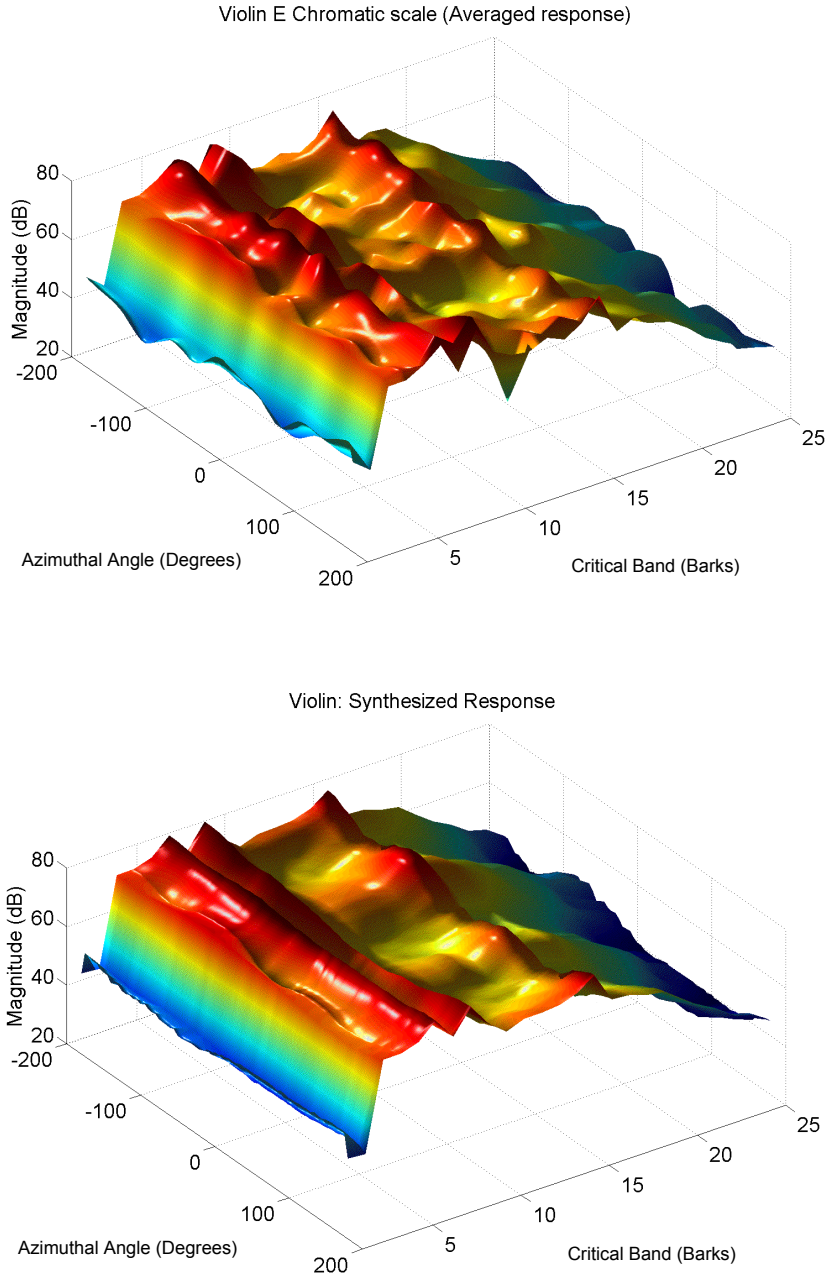


Figure 5.34: *Violin directional response and synthesized response.*

loudspeaker for the female vocal response are represented by the bar plot in Figure 5.35. The resultant in-band gains for the violin response are shown in Figure 5.36. We note the presence of negative gains, which represent phase inversion, used to form the complex directivity pattern.

To summarize, the steps for employing source simulation radiators in impulse response measurements are:

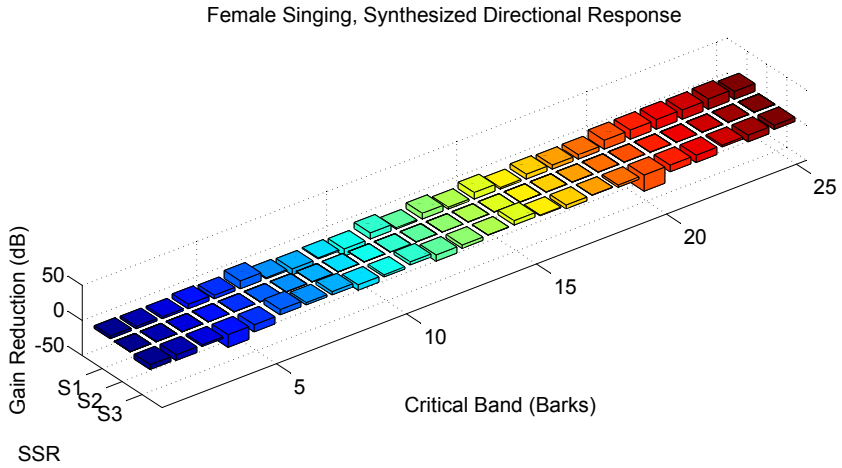


Figure 5.35: *Gain compensation in critical bands for synthesized female vocal response.*

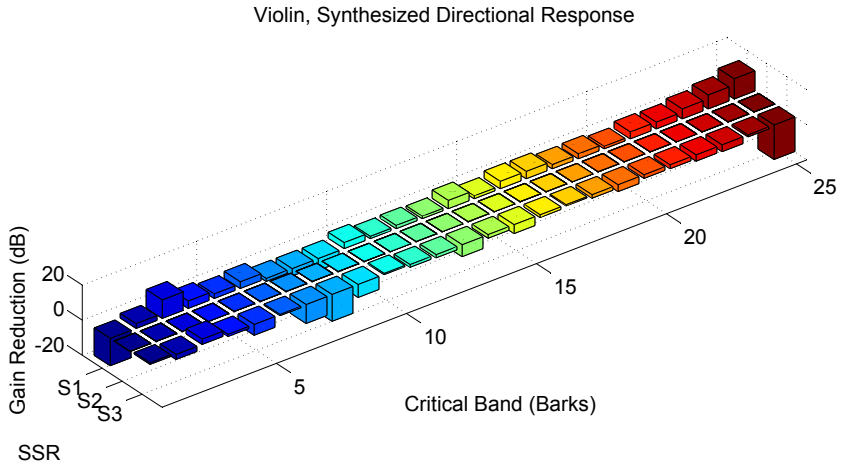


Figure 5.36: *Gain compensation in critical bands for synthesized violin response.*

---

1. Measure the directivity characteristic of the source in a free-field.
2. Measure the directivity characteristics of the source simulation array in a free-field.
3. Use least-squares optimisation to determine the frequency dependent gain at which each of the loudspeakers must be to yield an approximation of the source directivity and normalize according to the centre loudspeaker gains.
4. Capture three individual SRIRs from the separate excitation of each of the loudspeakers in the reverberant environment.
5. Filter the three SRIRs into perceptual bands and apply the corresponding gain coefficients
6. Sum up the SRIRs to obtain the new optimized SRIR.

5.6 Conclusions

In this chapter we have introduced the concept of virtual acoustic recording, a method of capturing acoustic events and representing them in a virtual environment with an equivalent perception to real world recordings. The method utilizes B-Format Ambisonic impulse responses to capture both the spatial and temporal properties of the soundfield. The removal of coloration effects due to the convolution chain was found to be an important factor in identifying virtual recordings from actual ones. Furthermore, binaural analysis as well as subjective listening evaluation demonstrated the effectiveness of the method in comparison to real acoustic recordings. Source width was flagged as an important subjective and objective characteristic, and a solution to the enhancement of measured B-Format responses using a source simulation array and a least square optimisation process was proposed. The method was demonstrated to give a good approximation of the source directivity, in particular for complex sources such as stringed instruments.

The principle focus of this chapter has been on single-perspective listening. We must now consider how to expand the virtual acoustic recording framework to satisfy multiple listener positions as well as walk through scenarios. However, in order to achieve this, we must have a large number of spatial impulse responses. Furthermore, the representation of 1^{st} Order soundfields in 3^{rd} reproduction systems must be considered. These issues will be addressed in the following chapter.

6

Spatial Impulse Response Interpolation

6.1 Introduction

Real-time convolution with finite room impulse responses for interactive audio-visual presentations is still a highly problematic area. As was shown in the previous chapters, there is the drawback that a single convolution measurement reflects only a static source and receiver (microphone) position. This is also the case in the multichannel sense, since, for example a stereo capture of a SRIR will still only reflect a static source position on playback. This is acceptable for the realm of music production, but not for convincing interactive real-time audio visual presentations, such as teleconferencing or gaming. In such cases the RIR needs to change according to the instantaneous source-listener position in the virtual space. This would lead to more convincing auralization, but to achieve this, several RIR measurements are required at spatially separated locations in a chosen room. The challenge then is to reduce the number of measurements for optimal storage and minimum reproduction effort whilst maintaining perceptually correct spatialization. To this end, we investigate here the interpolation of RIRs to aid in measurement reduction of RIR datasets.

We must also consider the spatial resolution required for a SRIR measurement dataset. The current state of the art in 'walk-through' computer-based auralization employs grids of RIRs, where the number of simulated RIRs decreases proportional to the distance from the source. Such an example is shown in Figure 6.1, from the work of Dalenback [Dalenbäck and Strömberg, 2006]. Here, in a chapel simulation, sparse SRIR computations are made over the main listening area and the SRIR density increases as we move closer to the source. Linear interpolation is

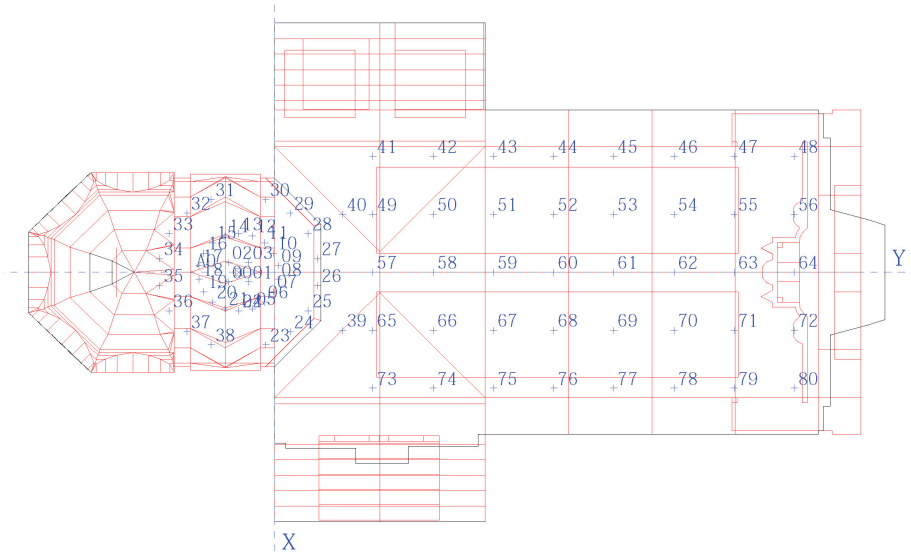


Figure 6.1: *Example of an RIR grid used in room acoustics auralization [Dalenbäck and Strömberg, 2006].*

applied at run-time on the RIRs and low latency convolution is implemented. Since a high density of simulated RIRs is easily achieved in computational-based auralization, there exists no formal framework for establishing the number or position of such measurements in relation to a given source position. However, this is not feasible for real-world measurements.

It is possible that interpolation can be used to create new SRIRs 'in between' measured responses. For this, it is unnecessary for any interpolation algorithm to be employed for the diffuse-field. Interpolation is only relevant for the direct sound and the early reflections, which contain the most important information regarding localization. The reverberation tail, which is considered stochastic, can be synthesized using decorrelation methods. Thus, the transition region between the early reflections and diffuse decay must be considered.

This chapter is therefore organized as follows: We will begin by exploring the spatial resolution required for capture of an impulse response dataset in a reverberant environment based on perceptual localization. This is presented in the form of a case study for Test Environment 1. A novel interpolation algorithm will then be introduced that reduces the measurement effort in the capture of real-world SRIRs, and limitations of this method are evaluated. Finally, several measurement topologies are presented based on the accuracy of auditory localization in a reverberant environment and the performance of the interpolation algorithm.

6.2 Spatial Resolution of SRIR Datasets

In the realm of computational based auralization, the number of impulse responses to be calculated for an effective walk-through simulation is in general determined on a trial and error basis,

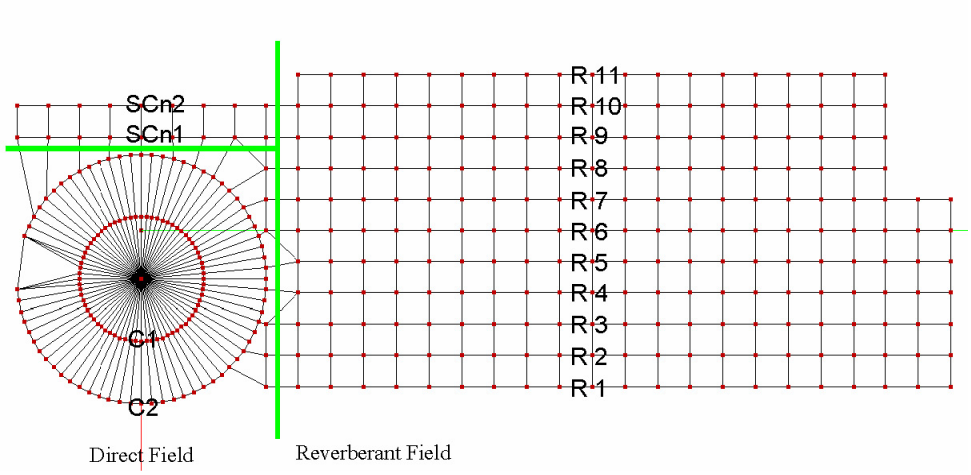


Figure 6.2: *IR grid for one source in test environment 1: The grid is divided into relevant regions: Measurement lines R1 to R11 defining the main audience grid, measurements C1 and C2 defining the direct-field source measurements and measurement lines SCn1 and SCn2 defining supplementary corner measurements.*

---

where the ultimate accuracy of the auralization is judged subjectively. It is surprising that there is relatively little psychoacoustic foundation for RIR grid topologies, other than diffuse field assumptions. This is largely due to the fact that the off-line computation of geometric based RIRs is trivial (apart from the trade-off in computation time and RIR accuracy) in comparison to real-world RIR measurements. Furthermore the relationships between direct to reverberant ratio, onset duration, listener distance, lateral energy level, IACC, ITD and ILD are highly complex for any particular environment, making it difficult to form strategies for RIR measurement. In such cases, it is better to overestimate the number of measurements required to construct the auditory scene. However, this leads to a large number of measurements, in particular if we wish to apply the source simulation radiator approach of Section 5.5.1. As an example, consider the measurement of Test Environment 1 for walk-through auralization purposes. An example of a complete measurement grid for one source in this room is shown in 6.2. The measurement grid is based on 5° circular measurements at 1m and 2m around a source, with a corresponding grid of impulse response measurements in the reverberant field, each spaced 0.5m apart. The capture of this grid, shown in Figure 6.3, was conducted over several days, using a Soundfield ST350 and 3 Genelec loudspeakers forming a source simulation radiator. Logarithmic sine swept tones were used as the test stimulus, and the resultant averaged pressure fields from the W component for the main grid are shown in Figure 6.4. Provided the measurements are carefully taken, highly accurate results can be obtained from data-based auralization, as has previously been shown by many authors, such as in [Karamustafaoglu et al., 1999], [de Vries and Baan, 1999], [de Vries and Hulsebos, 2004]. However, the drawbacks of measurement time, post-processing time and the tedious nature of the IR capture make such an undertaking a difficult task for the audio engi-

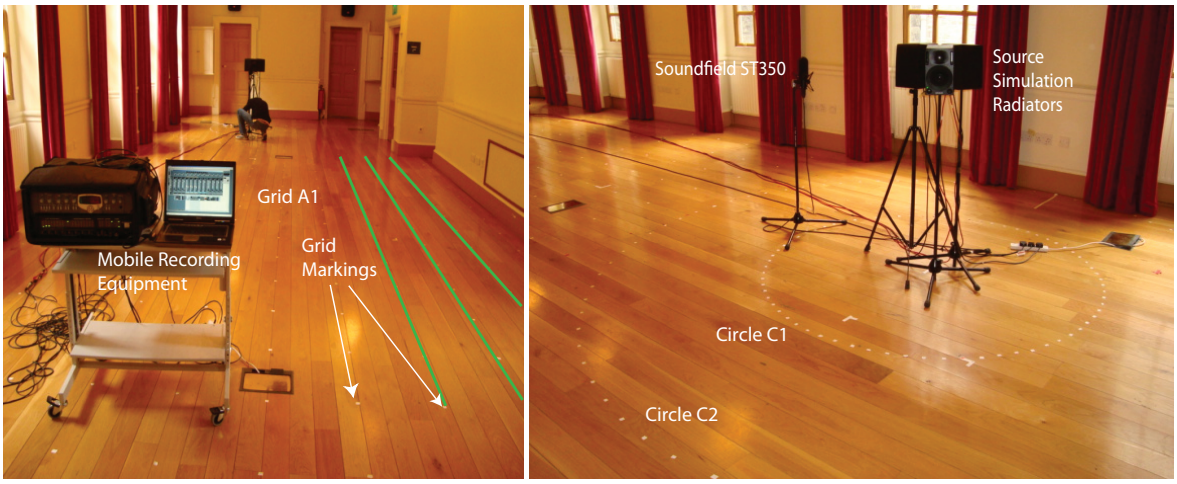


Figure 6.3: *IR grid measurement in test environment 1.*

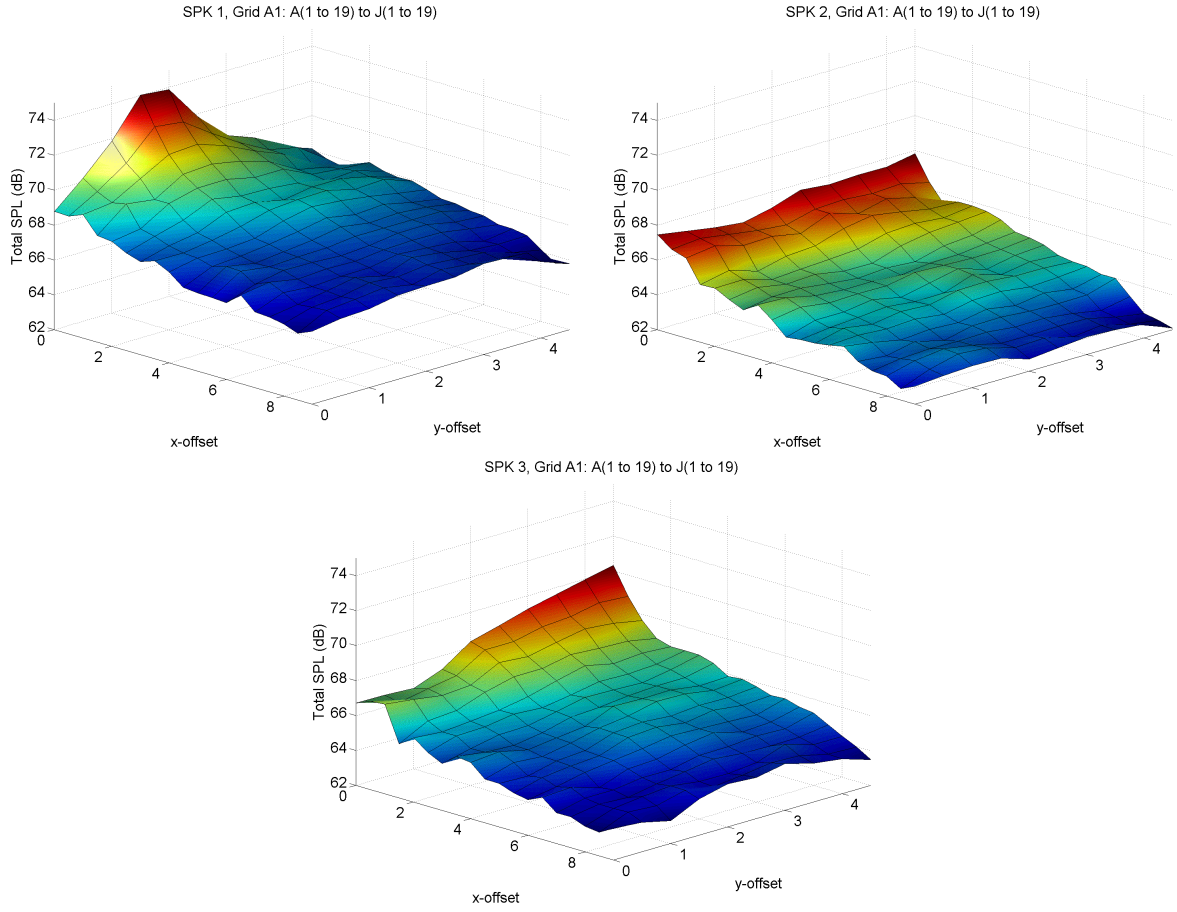


Figure 6.4: *Sound pressure level due to source simulation radiators, S1, S2, S3 over main grid in Test Environment 1.*

neer. The number of measurements can be reduced however, if we consider the psychoacoustical perception of sources within the environment. There are two key elements in this analysis

1. The localization blur in the horizontal and vertical plane
2. The perception of source depth in a reverberant field

For the first consideration, the spatial resolution of any virtual environment that allows the listener complete freedom of movement is dependent on the region of highest localization accuracy, i.e. the forward facing direction. In other words, since the listener can face any direction, it is important that the density of SRIR measurements in any forward facing direction is sufficient for good localization and spatial impression to be achieved. It was shown in Chapter 1 that the minimum discriminable angle in the forward facing direction can be as low as 1° to 2° depending on the spectral and temporal content of the source. However, the majority of studies reported on angular discrimination have been conducted under free-field conditions. We have seen in the previous chapter how the apparent source width increases as the source moves further into the reverberant field i.e. the IACC decreases when the level of the direct sound drops in relation to the diffuse field. It therefore follows that the localization blur will increase as the direct sound level decreases with respect to the level of the diffuse field.

In the second case, the perception of distance has been shown to be one that is not linearly proportional to the source distance. For example, both Nielson et al [Nielsen, 1993] and Gardner [Gardner, 1969], have shown that the localization of speech signals is consistently underestimated in an anechoic environment. This underestimation has also been shown by other authors in the context of reverberant environments, both real and virtual. In [Bronkhorst and Houtgast, 1999], Bronkhorst et al. demonstrate that in a damped virtual environment, sources are consistently perceived to be closer than in a reverberant virtual environment, due to the direct to reverberant ratio, as shown in Figure 6.5. Here the room simulation is conducted using simulated BRIRs created from the image source method. They show how perceived distance increases rapidly with the number and amplitude of the reflections. This effect for a source, 2m away from the listener is shown in Figure 6.5 (b) for regular reflection amplitude, constant reflection amplitude and constant amplitude raised by 3dB. In a similar study, Rychtarikova et al [Rychtarikova et al., 2009] investigated the difference in localization accuracy between real rooms and computationally derived BRIRs. Their findings show that at 1m, localization accuracy in both the virtual and real environments is in good agreement with the true source position. However, at 2.4m, the accuracy degrades, and high frequency localization errors were found in the virtual acoustic pertaining to the difference in HRTFs between the model and the subject. In the same vain, Chan et al [Chan et al., 2009] have shown that distance perception using recordings made from the in-ear microphones on individual subjects again lead to underestimation of the source in virtual reverberant environments, more so than with real sources.

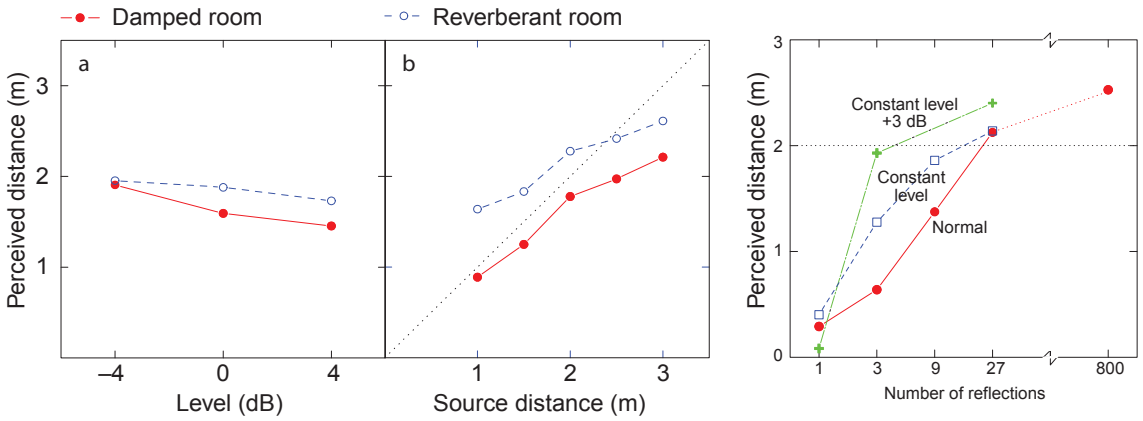


Figure 6.5: *Effects of level and reverberance of source on perceived distance: (a) Difference between damped and reverberant rooms and (b) with increasing reflection density and level [Bronkhorst and Houtgast, 1999].*

Waller [Waller, 1999] has identified that one of the key factors in distance perception is the importance of listener movement in the virtual space, which has not been considered in the previous studies. Since the extent of the effect of cross modality between the visual and auditory stimuli is still very much an open question, it is therefore important that in any depth perception study both the auditory and visual cues are congruent. That is, the correct audio and visual motion parallax cues must be presented to the listener if the correct perception of depth is to be achieved. A subjective localization experiment, which utilizes these cues will now be described.

6.2.1 Subjective Experiment V: Spatial Resolution in Virtual Environments with Motion Parallax

In this experiment, we assess the spatial resolution of depth perception in virtual audio systems in comparison to real world stimuli. The main hypothesis of this experiment is

> That the spatial resolution of both real and virtual sound sources in the diffuse field is significantly poorer than in the direct field when correct motion parallax cues are presented to the listener.

This analysis will directly inform how we should craft a topology of SRIR measurements for walk-through auralization. In this study, we consider that the depth perception cues and the visual cues must be non-contradictory if accurate localization data is to be achieved. For this reason, we utilize Wave Field Synthesis for virtual source reproduction. This is further justified since

1. The correct parallax cues can be achieved when a listener moves in the reproduced wavefield

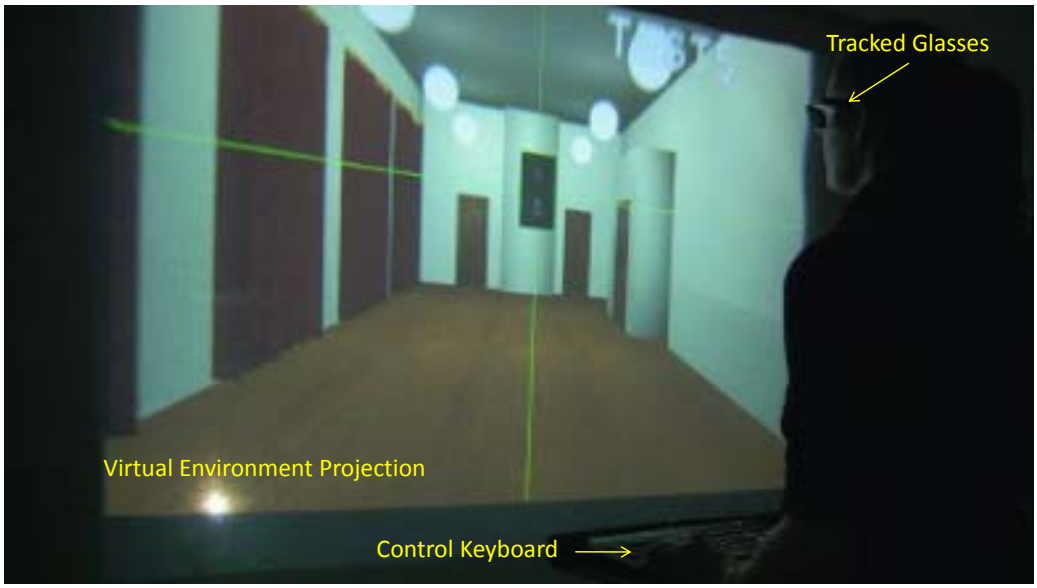


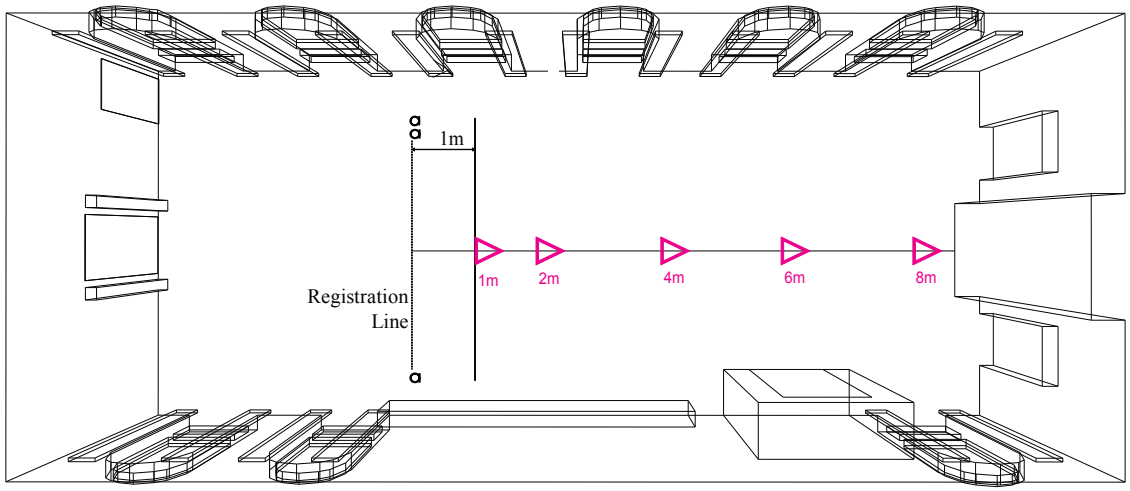
Figure 6.6: *Head tracked virtual display for depth perception analysis.*

2. As was shown in Chapter 3, Higher Order Ambisonics approaches the performance of Wave Field Synthesis for very high orders of reproduction.

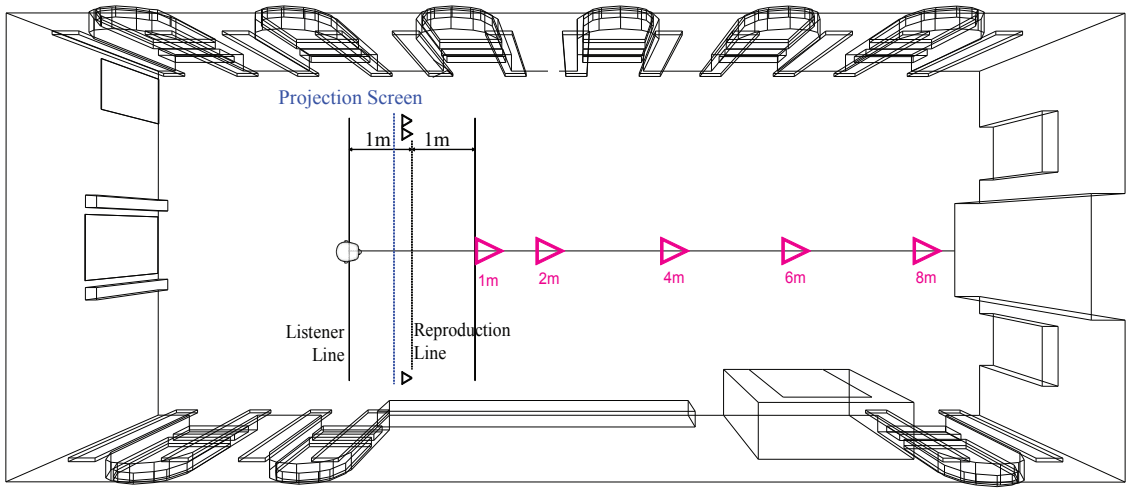
In order to assess the perception of depth of the real and virtual sources, a virtual visual display is required. The display created for these tests displayed a virtual version of Test environment 1. An important aspect of this graphical user interface (GUI) was that it should also give a plausible perception of visual depth. Several different techniques for rendering 3-D vision were investigated including the anaglyph technique, polarized glasses, shutter glasses, and chroma-depth [Fauster, 2007]. These methods are generally designed for a static viewing position and as such do not include the motion-parallax cue offered by Wavefield synthesis systems. However, recently Lee [Lee, 2007] has proposed a simple and effective technique for depth perception by head-tracking to recalculate the orientation of objects in a two-dimensional scene. The method employs the use of IR emitters mounted on a pair of see-through glasses and corresponding IR sensors mounted on the screen. As the subject moves their head, the coordinates of the infrared emitters are updated in real-time such that the perspective of the scene adjusts accordingly and a correct perception of depth is achieved.

6.2.1.1 Wavefield Measurement and Reproduction

Test environment 1 was used for this experiment due to its strong diffuse field. In this experiment, subjects were asked to gauge the depth of both real world and virtual stimuli. For this, different depths from a reference recording position were chosen pertaining to 6dB drops in the direct sound level, yielding source positions at 1m, 2m, 4m, and 8m. A 6m point was also included to gauge the accuracy over equidistant spacings between 2 and 8m as shown in Figure



(a) Wave field recording setup



(b) Subjective evaluation setup

Figure 6.7: *Recording and reproduction setups for depth perception experiment.*

6.7. In order to form the virtual sources, wavefield measurements were taken at the registration line for each of the different source positions. 32 measurements in total were taken for each source position using a Soundfield ST350 microphone and Genelec 1029a loudspeakers as the source. The spacing between each measurement was 12.5cm, the exact spacing of the loudspeakers in the reproduction array. The first 100ms of the velocity components in the Y-direction of the recorded wavefields are shown in Figure 6.8. The full B-Format responses are shown in Appendix H[1].

As was shown in Chapter 2, a consequence of finite spatial reproduction is that there is a limited frequency range over which accurate wave field reconstruction is possible, and above this

[1]Further information on this experimental setup and its use for investigating OPSI based wavefield synthesis can be found in [Gorzel, 2009]

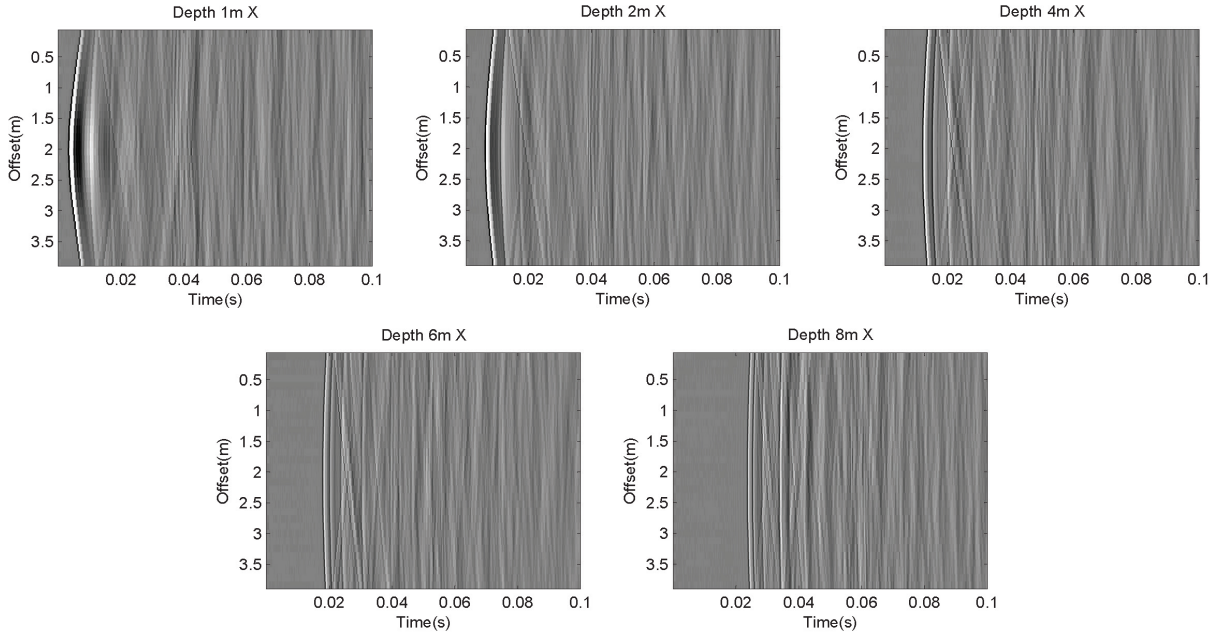


Figure 6.8: *Velocity measurements in the X (forward facing) direction for source positions of 1, 2, 4, 6, and 8m.*

range spatial aliasing effects occur. For the array used in this work, the spatial aliasing frequency (from equation 3.78) is at 1447Hz. Above this frequency the reproduced wavefield suffers audible coloration and spectral and spatial distortion. To reduce these effects, we reproduce the wavefield only below the spatial aliasing frequency. Above the aliasing frequency, f_a, we employ the previously introduced OPSI method [Wittek, 2007]. Wittek suggests the use of a vector based panning method [Pulkki, 1997] for stereophonic imaging. However, since we are using measured acoustic fields, we form recording angles from the perspective of the listener position, using selected microphone responses. We base the angles on the localization theory established by Williams, Thiele and Wittek [Williams, 1987], [Wittek and Theile, 2002]. The recording angles in this case are formed from both level and time differences at the microphones. Critical linking between recording angles is achieved by employing time or intensity bias in the RIRs (above f_a) [Williams and Le Dû, 2001]. This is a further advantage of using B-Format microphones to capture the RIRs, since 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 [Berkhout, 1988]. The array signals were also tapered using a raised cosine filter, so as to avoid any dispersion effects at the ends of the array.

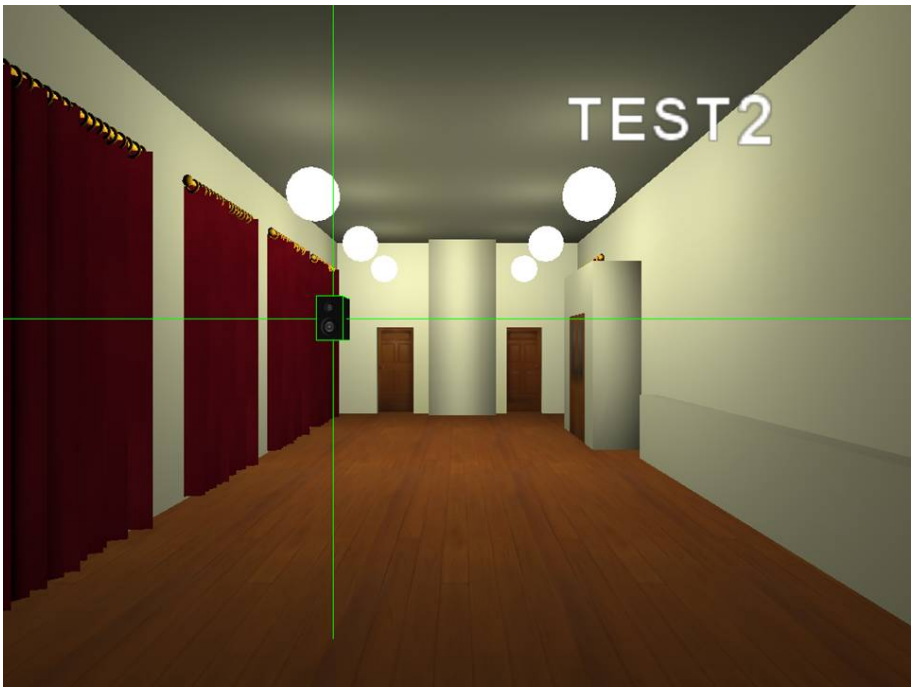
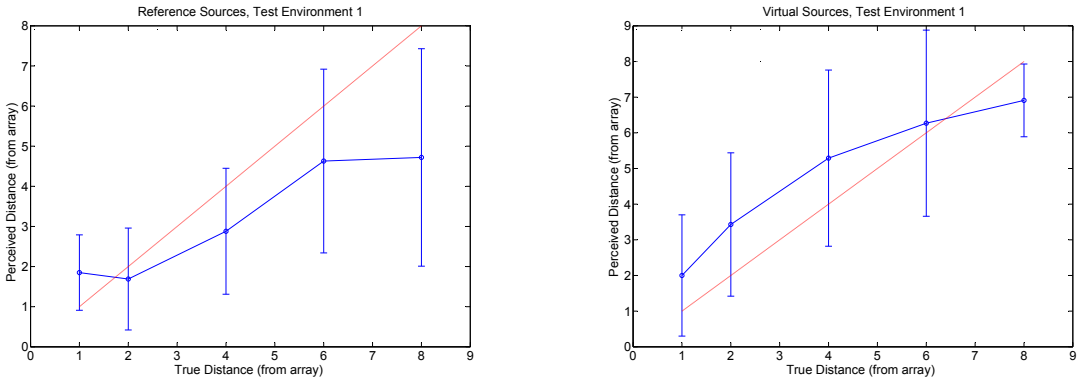


Figure 6.9: *Virtual environment created for depth perception test.*

6.2.1.2 Test Implementation

A textured virtual model of Test Environment 1 was implemented in the *Blender* [Blender, 2010] environment from the CAD drawings of the room created for the acoustic simulations in Chapter 3. Whilst the model gives a good impression of the true acoustic space, the level of detail is not photo-realistic, as can be seen in Figure 6.9. However, this is inconsequential, since, as has been shown by Thompson et al [Thompson et al., 2004], non-realistic graphics rendering does not affect distance perception. A 3-D model of a loudspeaker was also created as a cursor with which the subject could move freely around the virtual room, so as to indicate the position of the source. The perspective of the cursor also changed according to the tracked head position. A cross-hair was used which intersected the cursor and made it easier to position the loudspeaker in the virtual room.

10 experienced subjects were used for these listening tests, each subject under 35 years of age and of excellent hearing. The stimuli used were the same female speech samples used in the subjective testing of Section 4.3.2. Each depth position was assessed twice, for both the actual and virtual sources, resulting in 20 presentations in all. The order of the sample playback was completely random. For each presentation, the subject was asked to place the cursor at the point in the virtual room where they felt the source originated.



(a) (b)

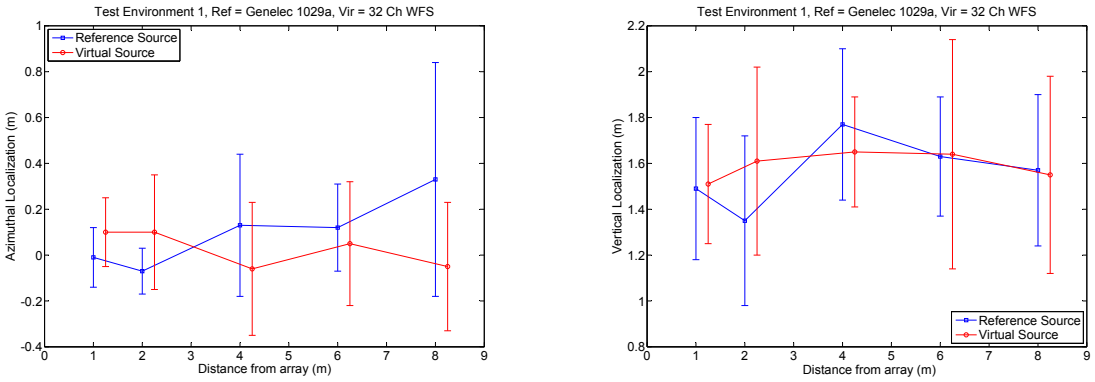
Figure 6.10: *Depth perception results (a) Reference sources (b) Virtual sources.*

6.2.1.3 Results

The results for the reference source localization are found in Figure 6.10(a). It is found that the localization accuracy is underestimated when the source is in the diffuse field. At both 6m and 8m for the real sources, the mean localization is located at approximately 4.5m. This result corroborates the results of the other aforementioned studies on depth perception in real rooms. Interestingly, in the near-field the source position is over-estimated. This also occurs with the virtual source presentations, but in this case, the over estimation continues beyond 1m until 6m. This is partly due to the fact that virtual reproduction systems such as Wavefield synthesis are deficient in terms of binaural direct sound cues, and thus the head shadowing effects of virtual sources are not the same as that of real sources [2] [Wittek et al., 2004]. Wittek also hypothesizes that reflections caused by the array itself distort the perception of localization.

The wavefield analysis of Figure 6.8 gives some indication as why the underestimation of real sources occurs. In the near-field, the level of the direct sound is far greater than that of the reverberant field, and the initial time-delay gap between the direct sound and the first arriving reflection is sufficiently large (of the order of 5ms). As we move out into the diffuse field, the level of the direct sound drops 6dB per doubling of distance, but the diffuse field level remains consistent. Furthermore, the initial time delay gap reduces significantly, and there is little difference in this regard, between the 6m and 8m wavefield. Also, we note that as the source moves further away from the listener, it becomes sufficiently planar such that near-field distortions in the ITD and ILD do not occur, i.e. from the perspective of the listener, the direct wavefronts are no longer spherical. There is therefore no major advantage to recording impulse response information beyond 6m for this room. Also, a dense grid of RIR measurements, such as that shown in Figure 6.2 is not necessary from a psychoacoustic point of view, since the real-world perception of depth is underestimated for far sources. This underestimation occurs

[2] This is also true of stereo used for high frequencies in the OPSI method, as was seen in Figure 3.16 of Chapter 3



(a) (b)

Figure 6.11: *Localization accuracy in real and virtual environments (a) Horizontal localization (b) Vertical localization.*

just beyond the critical distance point (0.96m).

Also of importance is the localization blur that occurs as real sources move away from the listener. In the case of the 8m example, the uncertainty of localization spans 5m. This error reduces progressively as we move towards the direct field. The forward facing localization blur for a source at 0° azimuth and 0° elevation is shown in Figure 6.11. For near sources, the standard deviation is less than in the reverberant field, both horizontally and vertically.

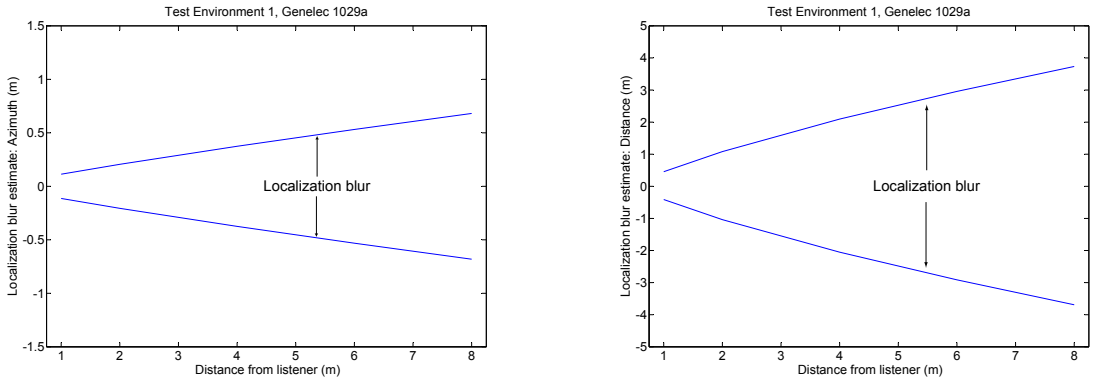
We therefore conclude in response to the experimental hypotheses that the spatial resolution of both real and virtual sound sources in the diffuse field is significantly poorer than in the direct field when correct parallax cues are used. We also recognize that for sources in or close to the near field, a large number of spatial impulse responses will be required. This verifies the current practice of high density SRIR regions around sources in computational based auralization. Based on this analysis, we will now consider a topology for SRIR measurement in Test environment 1.

6.2.2 Derivation of Perceptual-Based SRIR Topology

Distance perception can be modeled using the Stevens' power law [Stevens, 1975], which has been shown to give a good relationship between actual distance and perceived distance. This is given by

$$\sigma = k \cdot d^n \tag{6.1}$$

where d is the actual distance, σ is the judged distance, the modulus k is a unit dependent scale factor and n determines the shape of the psychophysical function. We employ the function to give an estimate of the perceptual distance blur based on the standard deviation of the results in Figures 6.10 and 6.11. Here we employ a least squares fit to determine the appropriate coefficients of equation 6.1 that form a curve matching the dispersion of localization, i.e. the power function is fit to the standard deviation at each distance. Note that we also use equation



(a) (b)

Figure 6.12: *Localization blur upper and lower bounds based on perceptual data (a) Horizontal localization (b) Distance localization.*

6.1 to fit the horizontal localization dispersion. The results are shown in Figure 6.12. The curves describe the measured perceptual localization blur as a continuous function, allowing us to derive a perceptual based topology for SRIR measurements. For example, starting with a source surrounded by a circle of receiver positions at 1m, SRIR measurements should be taken at distances 0.1m relative to each other, since this is half the localization blur at the 1m distance from Figure 6.12 (a). Since the perception of depth at 1m has a blur of approximately 1m, then the next distance measurement needs only to be taken at around 2m and so on. The resultant circular topology is shown in Figure 6.13. We see that the number of SRIR measurements is significantly reduced for far receiver points in comparison to the over-estimated topologies of Figures 6.1 and 6.2. However, a significant amount of measurement is still required close to the source (approximately every 6°), which is both tedious and prone to error. It is more practical to take sparse SRIR measurements where high SRIR density regions are required and to interpolate between them. This will now be investigated.

6.3 Interpolation of Spatial Impulse Responses

One method of interpolation of impulse responses, proposed by Haneda et al. [Haneda et al., 1999a] uses a common acoustical pole and residue model. They propose that variations in the RIR (or its frequency domain equivalent, the room transfer function) can be characterized by residue variations in the model with different source or receiver positions, and as such, interpolating between RIRs simplifies to interpolating residue functions. This approach seems to be effective for the low frequency component of the room transfer function. Wavefield based approaches to RIR interpolation typically involve extrapolation from a given dataset of RIRs using the Kirchoff-Helmholtz or Rayleigh integrals [Berkhout et al., 1993], [de Vries and Baan, 1999]. However, this still involves the time-consuming measurement of real RIR data, and is

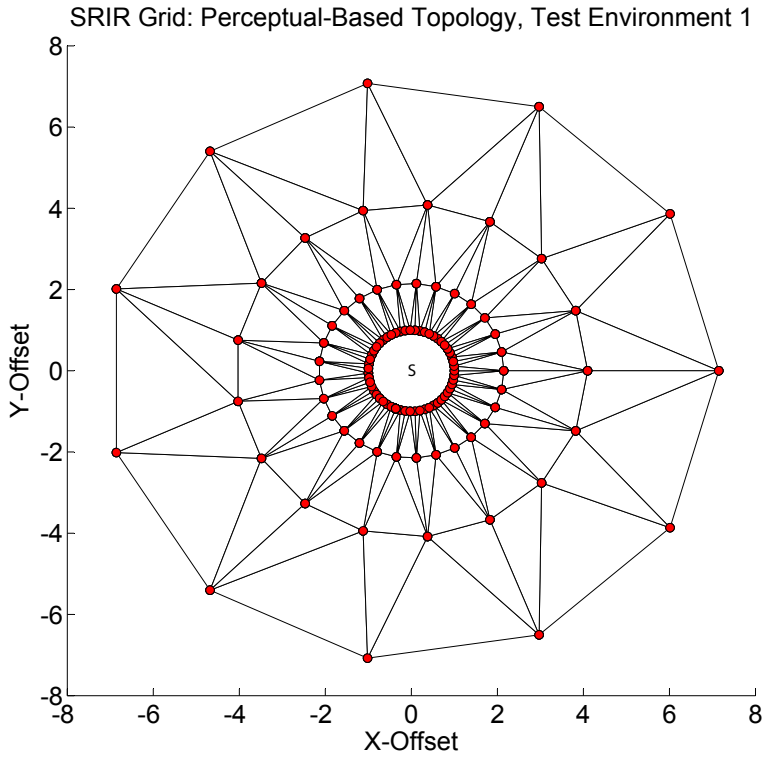


Figure 6.13: *SRIR Topology based on perceptual analysis.*

only effective below the aliasing frequency. Microphone array geometries have been proposed by De Vries et al [de Vries and Hulsebos, 2004], [de Vries and Berkhout, 1996] which allow for the extrapolation of acoustic impulse responses based on plane wave decomposition. Here, circular arrays (typically of 1m radius) can be used to record wavefields whereby the early reflections are extracted from the RIRs and modeled as point sources producing plane waves at the same incident angles as the original reflected waves. The diffuse decay sections are then further parameterized through the use of critical band filtering, where the time envelopes in each band are recorded at low sampling rates and the random fine structure of the RIR tail ignored. A more recent method, proposed by Huszty et.al [Huszty et al., 2008], uses fuzzy modeling techniques for RIR interpolation. Their work recognizes that only the early reflections can be interpolated and for this, a type of temporal mapping is necessary. They propose a manual pairing of reflections to ensure the temporal mapping is accurate. In this chapter, we too split the measured impulse responses into their early reflection and diffuse decay regions. However, in the approach presented here, automatic temporal alignment of early reflections is possible. We utilize a process known as *Dynamic Time Warping* (DTW) [Sakoe and Chiba, 1978] to temporally align early reflections from pairs of acoustic measurements for interpolation. Tail synthesis is implemented through critical band filtering, where decorrelation is achieved through randomized time shifts in each band.

6.3.1 Transition Region in Room Impulse Responses

The generation of new RIRs consists of three main parts: splitting the existing RIR measurements into their early reflections and diffuse decay sections, early reflection generation through interpolation, and diffuse decay synthesis through decorrelation.

We therefore consider an RIR to have two significant sections: the direct sound with early reflections and the diffuse decay. Let h_i denote the room impulse response measured at position i in a 1-D microphone array. h_i can be split into two components, the early reflections, h_i^e and the diffuse decay h_i^l such that

$$h_i = \left[h_i^e[1 : n_t] : h_i^l[(n_t + 1) : N]\right] \tag{6.2}$$

where n_t is the point of transition between the early and late reflections, and N is the total number of samples in the RIR. The transition time, T_t, at which n_t occurs is of significant importance here. Existent measures of T_t are typically related to room volume and the density of reflections and a good summary can be found in [Meesawat and Bajers, 2002]. Recently, new statistical measures of the transition time have been introduced by Stewart [Stewart and Sandler, 2007]. These are based on higher order cumulants such as Kurtosis and provide an effective method for detecting the transition time of individual responses.

In this work we use the geometrical method suggested by Naylor and Rindel [Naylor and Rindel, 1992] where T_t is the time of arrival of the fourth order reflections in h_i. This can be computed from the mean-free path by

$$T_{ro} = \frac{4V}{cS}\left(O_e + 1\right) \tag{6.3}$$

where V is the room volume, S is the surface area, c is the speed of sound and O_e is the reflection order [Kuttruff, 1979]. We therefore take T_t as T_{ro} when O_e is equal to 4. We justify the use of this method based on its simplicity and the fact that for a given recording framework in a VAE, the geometrical parameters of the room *must* be known.

6.3.2 Interpolation of Early Reflections using Dynamic Time Warping

Let us consider the case where we have measured two RIRs h_1^e and h_3^e at points R1 and R3 as shown in Figure 6.14. We will now attempt to create a new interpolated (direct-sound and early reflections) RIR, $\hat{h}_2^e$, as if it were measured at position R2. Figure 6.15 shows an example of the first 20ms of two such measured RIRs, taken from test environment 1.

Since the RIRs are recorded at different spatial locations their early components will contain sparse reflections occurring at different times in each impulse. Consequently, even at spatially close locations this sparseness means that linear interpolation can result in significant smearing of reflections in the interpolated result, as shown in Figure 6.16. It is therefore necessary to

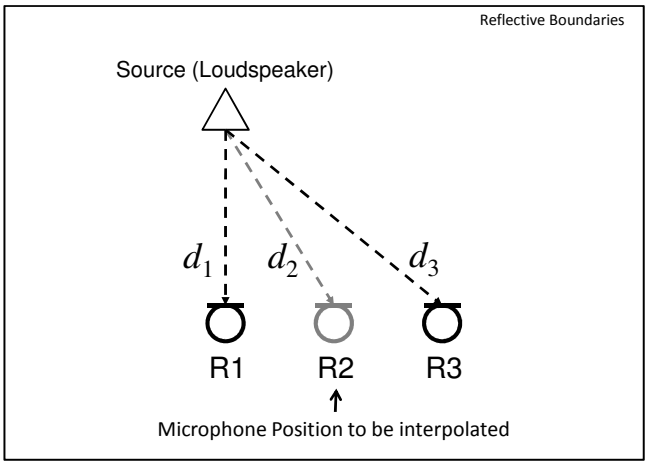


Figure 6.14: *Interpolation between two microphones.*

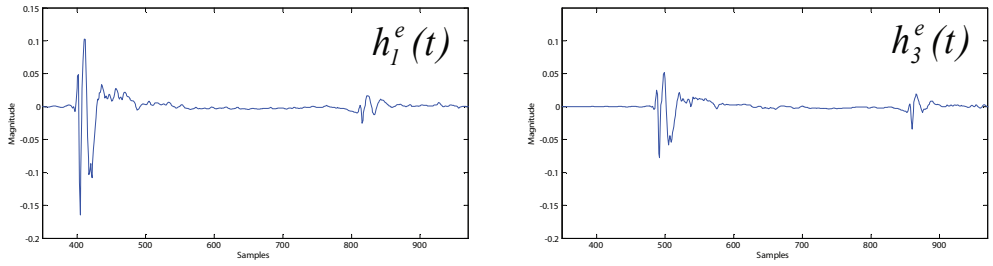


Figure 6.15: *RIRs at positions 1 and 2.*

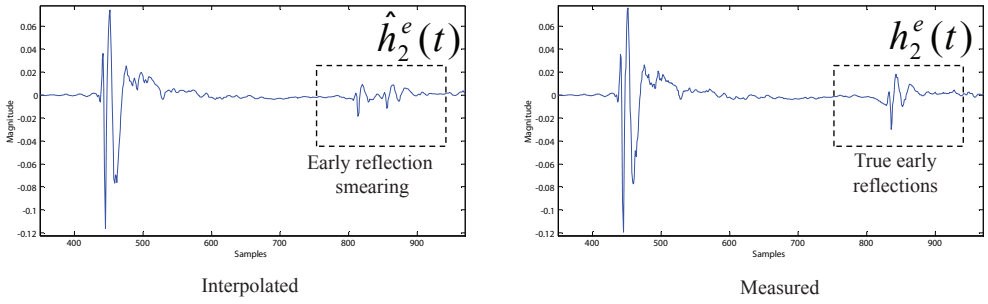


Figure 6.16: *Comparison of impulse response created from linear interpolation between two RIRs to an actual impulse response measured at the position of interpolation.*

---

align the signals in some way. One can apply a delay to one signal so that the direct paths align, but this does not guarantee that subsequent reflections will match up, due to different reflection path lengths at different positions in the room.

It is therefore necessary to temporally align the main reflections of the impulses prior to interpolation. Dynamic Time Warping (DTW) [Sakoe and Chiba, 1978] is a technique which allows us to do this. It stretches (warps) the signals non-linearly by repeating samples in each

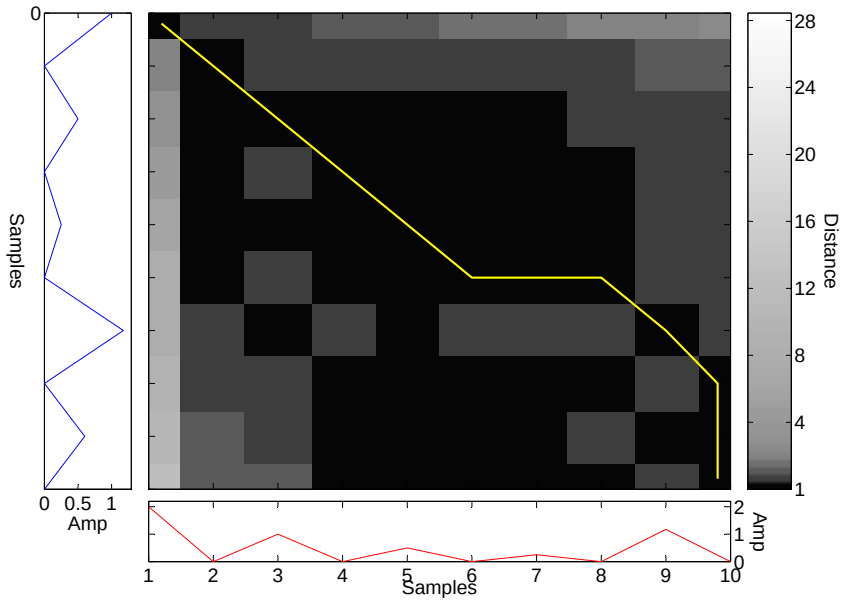


Figure 6.17: *Accumulated Distance Matrix.*

time series allowing us to 'line up' the reflections. A warp vector is created for each time series which describes how the signals are stretched. Fast implementation of DTW can be found in [Salvador and Chan, 2004].

In describing our application of DTW, consider the case where we wish to interpolate the direct sound and early reflections of two impulse responses h_1^e and h_3^e. The warp vectors are formed by calculating a minimum distance warp path through an accumulated distance matrix. The 'distance' is the Euclidean distance between data point i in one time series and data point j in the other time series. The warp path W will be of length K such that

$$n_t \leq K \leq 2n_t \tag{6.4}$$

and the k^{th} element of the warp path is given by

$$w_k = (i, j) \tag{6.5}$$

where i and j are time indexes in h_1^e and h_3^e respectively. The optimal warp path is then the path through the matrix with the minimum accumulated distance given by

$$D(W) = \sum_{k=1}^{k=K} D(w_{ki}, w_{kj}) \tag{6.6}$$

where $D(W)$ denotes the distance of the warp path and $D(w_{ki}, w_{kj})$ represent the distances between the sample indexes at the k^{th} element of the warp path. A trivial example of such a matrix with two time series (16 samples long) is shown in Figure 6.17.

The warp path is subject to several constraints. First, the path must begin at the first sample of each signal ($h_1^e(1)$ and $h_3^e(1)$) and end at the last sample of each signal ($h_1^e(n_t)$ and

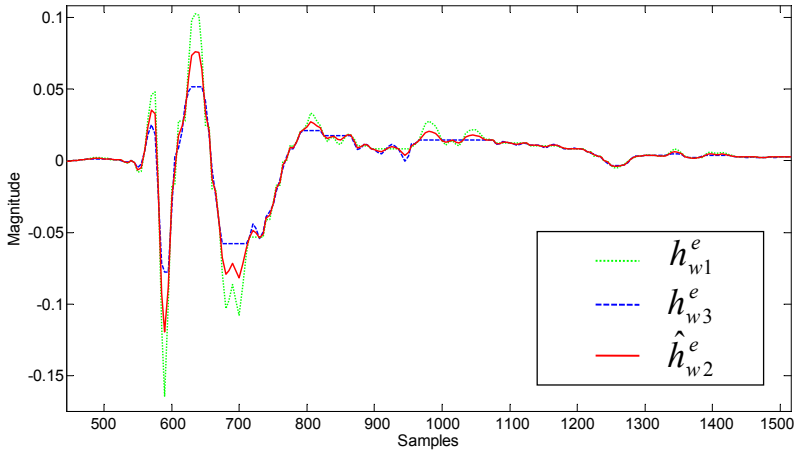


Figure 6.18: *Warped RIRs at positions 1 and 2 and the interpolated warped RIR.*

$h^e_3(n_t)$), ensuring that each sample index of each signal is used during its formation. A continuity condition is also applied which ensures that the warp path only traverses through the matrix via adjacent cells. Furthermore the path must monotonically increase, in order to ensure that it never overlaps itself. Thus, to obtain the interpolated impulse response $\hat{h}^e_2$, it is first necessary to apply DTW to h^e_1 and h^e_3, which gives their warped versions h^e_{w1} and h^e_{w3}. This aligns the main reflections of both RIRs and allows for simple linear interpolation between them to obtain the magnitude of the unknown RIR, $\hat{h}^e_{w2}$. This is shown in Figure 6.18.

The magnitude interpolation is weighted based on a ratio of the inverse distances between the source and the microphones, since sound pressure level is inversely proportional to the distance from the source.

$$\alpha = \frac{\frac{1}{d_3} - \frac{1}{d_2}}{\frac{1}{d_3} - \frac{1}{d_1}} \tag{6.7}$$

where $d_3 > d_2$ and hence,

$$\hat{h}^e_{w2} = \alpha h^e_{w1} + (1 - \alpha) h^e_{w3} \tag{6.8}$$

Now the warp vectors, w_1 and w_3 that describe how h^e_1 and h^e_3 are mapped onto h^e_{w1} and h^e_{w3} by the DTW must be interpolated to obtain w_2. Again linear interpolation is used to accomplish this and the weights, β and $1 - \beta$, are calculated based on the distances of the microphones to the source.

$$\beta = \frac{d_3 - d_2}{d_3 - d_1} \tag{6.9}$$

and hence,

$$w_{int} = \beta \; w_1 + (1 - \beta) \; w_2 \tag{6.10}$$

An example of this interpolation is shown in Figure 6.19.

The final step in the process is to map the warped interpolated vector back into the "unwarped" time domain using the interpolated warp vector. This is achieved by incrementally passing through the interpolated warp vector as described in the pseudo-code below.

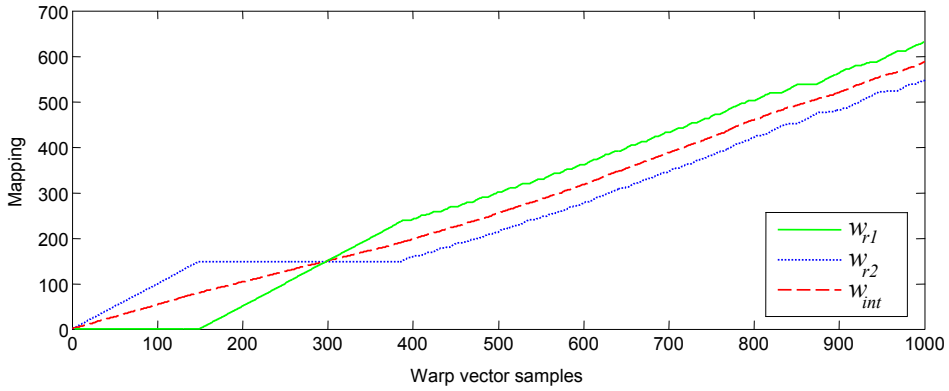


Figure 6.19: *Interpolation of Warp Vectors.*

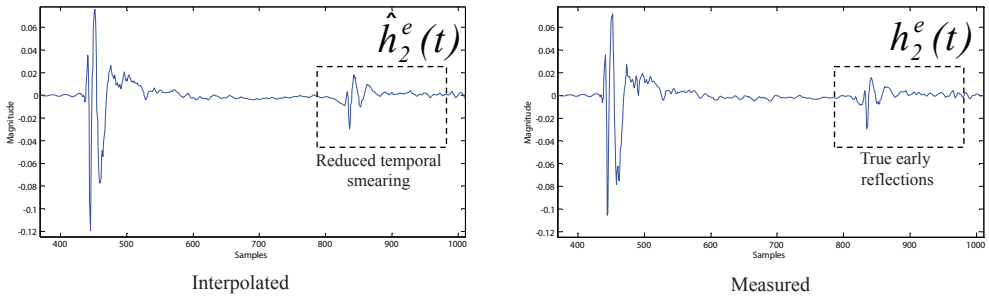


Figure 6.20: *Comparison of impulse response created from DTW interpolation between two RIRs to an actual impulse response measured at the position of interpolation.*

```
r_int[1]= r_int_w[1]
i=2
for x=2 to length(w_int)
if w_int[x] != w_int[x-1]
r_int[i]=r_int_w[x]
i++
end if
end for
```

Figure 6.20 shows a comparison of an interpolated RIR $\hat{h}_2^e(1)$ and a real measured RIR $h_2^e(1)$ taken from the desired interpolated position. We notice that the temporal distortions that were present in the linear interpolation of Figure 6.16 are no longer present.

6.4 Synthesis of Diffuse Decay

Having obtained an interpolated version of h_i^e, we now turn to the synthesis of the tail. In the following we again consider that we have measured a dataset of RIRs and wish to synthesize a new RIR in between two of them. Here we will attempt to synthesize a new version of h_i^l which

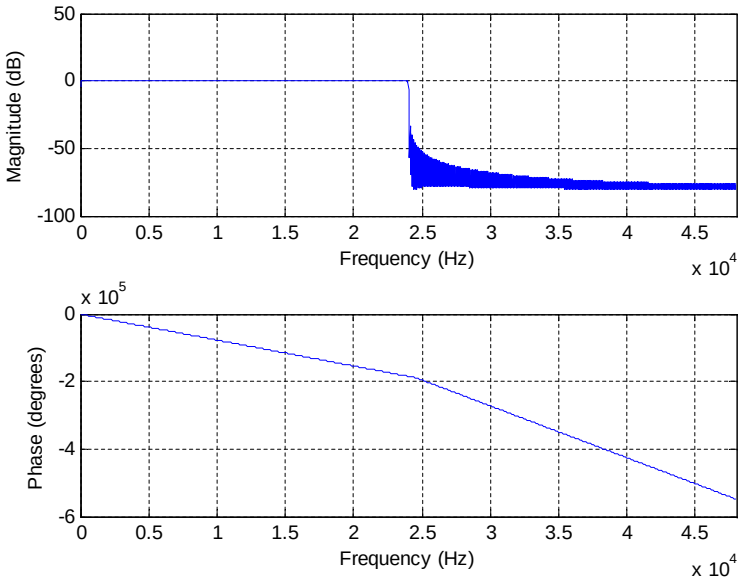


Figure 6.21: *Combined response of ERB filterbank.*

is decorrelated from the measured RIRs, but will retain their timbre.

In general, the decay of an RIR is considered a stochastic process above the Schroeder frequency, with the response having independent real and imaginary components [Kuttruff, 1979]. This has led to synthesis methods based on statistical time-frequency models characterized by the frequency dependent reverberation time and the initial amplitude spectrum [J. M. Jot and Warusfel, 1997]. Here we employ a different time-frequency approach based on the separation of the tail into perceptual critical bands. It has been shown in [Boueri and Kyirakakis, 2004] that perceptual based decorrelation can be achieved by delaying source signals after critical band filtering. We employ critical band filtering here using a bank of Equivalent Rectangular Bandwidth (ERB) FIR filters with a combined flat frequency response and piecewise linear phase.

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 [Boueri and Kyirakakis, 2004]. The resultant synthesized reverberation tail retains the timbre of its measured counterpart, but is now successfully decorrelated. The output is then summed and appended at time T_t to the early reflections h_i^e. A crossfade of 5ms is applied to ensure a smooth transition.

6.5 Analysis of DTW Interpolation Algorithm: Ambisonic Room Simulations

The DTW interpolation process does not aspire to a complete physical facsimile of measured SRIRs, but instead, to the correct perception of localization and spatialization of the true SRIR.

Thus, the perceptual attributes of interpolated SRIRs at reproduction, rather than the physical attributes are investigated here. To this end, the capture and reproduction setup shown in Figure 6.22 is simulated. The capture setup consists of an omnidirectional point source and two B-Format receivers enclosed in a reflective environment. We simulate RIR capture at the receiver points using the Image Method of Allen and Berkley [Allen and Berkley, 1979]. This auralization process allows us to create finite impulse responses that model the acoustic channel between specified source and receiver points, as well as varying the reflection characteristics of the simulated environment.

In this setup, we will keep the source position S1 and receiver position R1 stationary and 2m apart. We will then vary the position of receiver R3 from 0 to 1m away from R1 in finite steps. At each step, we capture spatial impulse responses at R1 and R3 and interpolate another spatial impulse response $\hat{\mathbf{h}}_\mathbf{2} = \left(\hat{h}_{w2}, \hat{h}_{x2}, \hat{h}_{y2}, \hat{h}_{z2}\right)$ exactly in-between the receivers (at R2). By comparing this interpolated estimate to the true response at R2 ($\mathbf{h}_2$), we can see the effect of increased microphone separation on the interpolation process. Furthermore, if we vary the reflection coefficient β of the room (where $\beta = 0$ represents free-field conditions and $\beta = 1$ represents completely reflective surfaces), we can study the effect of increasing the levels of early reflections on the interpolation process.

Also shown in Figure 6.22 is the reproduction setup, which consists of an 8-channel array and an Ambisonics 1$^{\text{st}}$ Order 'energy' decoder. Reproduction is simulated by the virtual loudspeaker approach, where head related impulse responses (HRIRs) captured from a KEMAR binaural mannequin [Algazi et al., 2001] are convolved with the decoded loudspeaker feeds to the 'virtual' array. The decoded loudspeaker feeds are formed from the pseudo-inverse decode (Ambisonic equation 3.42) of $\mathbf{h}_2$ and $\hat{\mathbf{h}}_2$. The resultant binaural signals are

$$s_l(t) = \sum_{n=1}^{N} s_i(t) \otimes k_i^l(t) \tag{6.11}$$

$$s_r(t) = \sum_{n=1}^{N} s_i(t) \otimes k_i^r(t) \tag{6.12}$$

where $s_l(t)$ and $s_r(t)$ are the resultant left and right ear signals, s_i is the decoded loudspeaker feed, and k_i^l and k_i^r are the KEMAR binaural responses to loudspeakers i.

In this analysis we will investigate the interaural time difference below 700Hz and the interaural level difference above 700Hz for both the simulated and interpolated playbacks. The early IACC ($> 80ms$) estimated in the 500, 1000 and 2000Hz bands will also be used [Beranek and Martin, 1996].

The simulation setup of Figure 6.22 was implemented for three different size rooms, with dimensions listed in Table 6.1. Errors in the interaural level difference due to the interpolation process are shown in Figure 6.23(a). It can be seen that the error is generally below 1dB, with greatest errors occurring in the small room simulation for larger reflection levels. In all cases

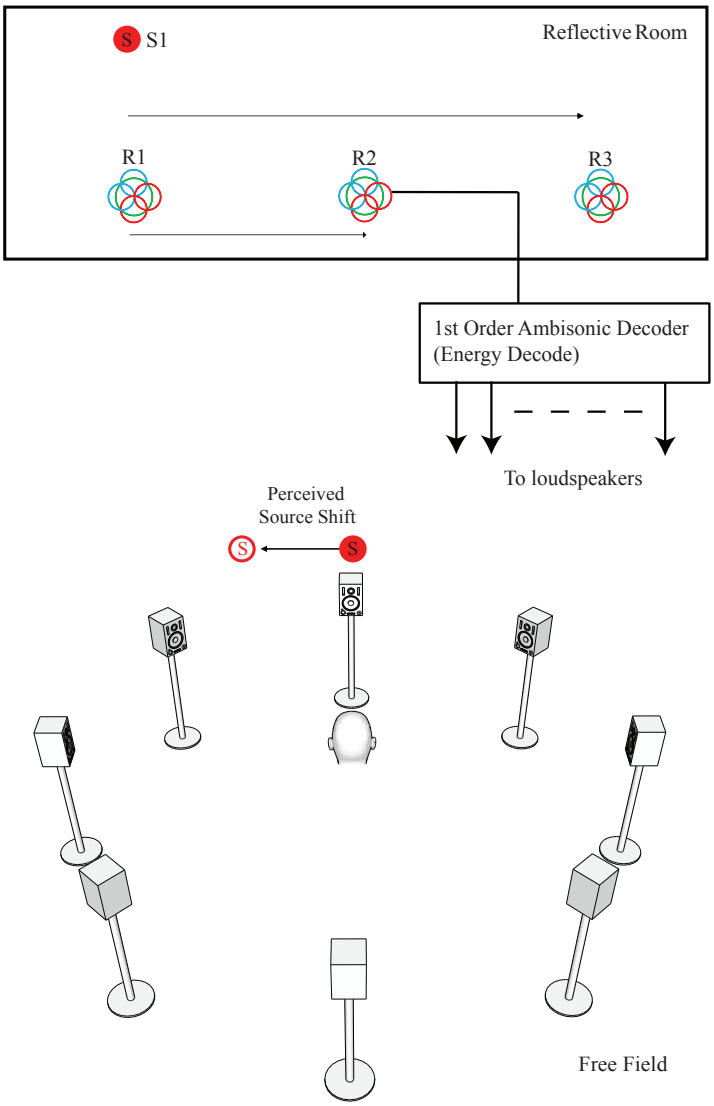


Figure 6.22: *Simulated recording and reproduction setup utilizing image source method and KEMAR HRTFs.*

---

the ILD error becomes greater the further the distance between the microphones. At almost 1m in all cases, the error goes above 1dB, which will lead to a perceptible difference. For the small room, we also note that at short distances, the error increases significantly with increasing reflection level. Although such errors are small, they will lead to high-frequency phantom source shifts above 2kHz. However in the larger rooms this error is below 1dB and is acceptable.

Errors in the ITD, shown in Figure 6.23(b), occur mainly due to the level of the reflections, and are largely independent of the spatial separation of the microphones. Again, the largest errors occur in the small room conditions. Maximum ITD errors are approximately 0.1ms,

Room	L (m)	W (m)	H(m)
Small	5.85	4	2.5
Medium	8.7375	6	3.75
Large	11.65	8	5

Table 6.1: *Room Dimensions used during simulation.*

which considering the interaural cross head delay is in the range of 0.7ms to 1ms, can lead to significant localization errors. The best results are obtained for the large room, which has the smallest region of error. Here the interpolation does not significantly affect the ITD until the reflection coefficient is very high (close to 1). Errors in the apparent source width are shown in Figure 6.23(c). These errors are mainly a function of reflection coefficient, and are very low with maximum values of approximately 0.1 in all cases. The small room again has the highest errors, with maximum values of reflection coefficient at 1m source distance causing a 0.12 error in the perceived source width.

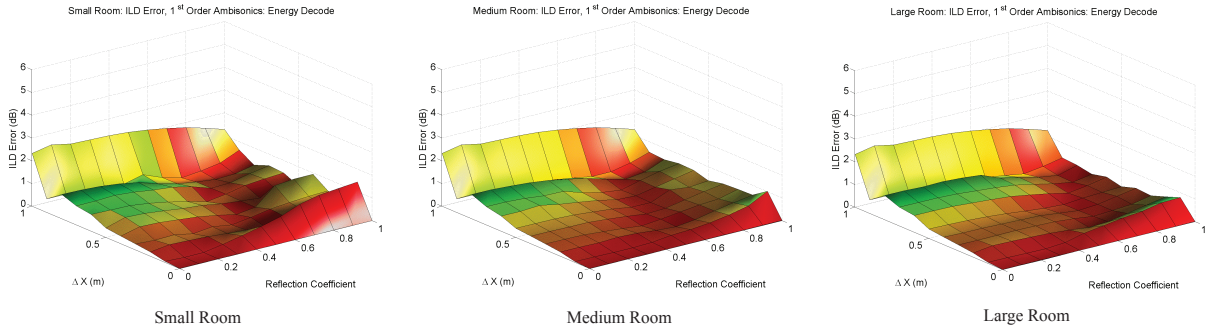
The performance of the algorithm in the small room demonstrates a necessity for the reflections to be sufficiently sparse in order for the interpolation to be effective. This is because a high density of reflections results in a feature misalignment in the algorithm. Such errors in the early reflection interpolation lead to ILDs and ITDs which differ from those of the true response. Moreover, the results demonstrate that effective interpolation is possible with the algorithm. Further studies based on real acoustic measurements and measured responses will now be implemented.

6.6 Analysis of DTW Interpolation Algorithm: Measured Soundfields

We can further gauge the effectiveness of the DTW interpolation method for real measurements by looking both objectively and subjectively at measured wavefields. This is advantageous since we can immediately see the effect of temporal distortions due to the interpolation process by performing a wavefield analysis. The main hypothesis of this study is

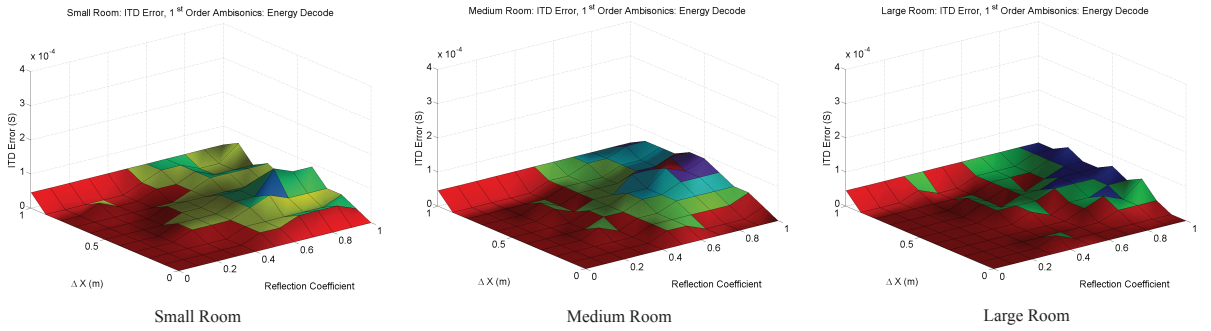
> That the localization accuracy of wavefield reproduction is not degraded using soundfields incorporating DTW interpolation.

Thus, we will test the use of our interpolation method by synthesizing 32 point WFS datasets from 4, 8 and 16 RIR measurements with interpolation, and compare these to a full 32 point measured RIR dataset for reproduction over a line array.



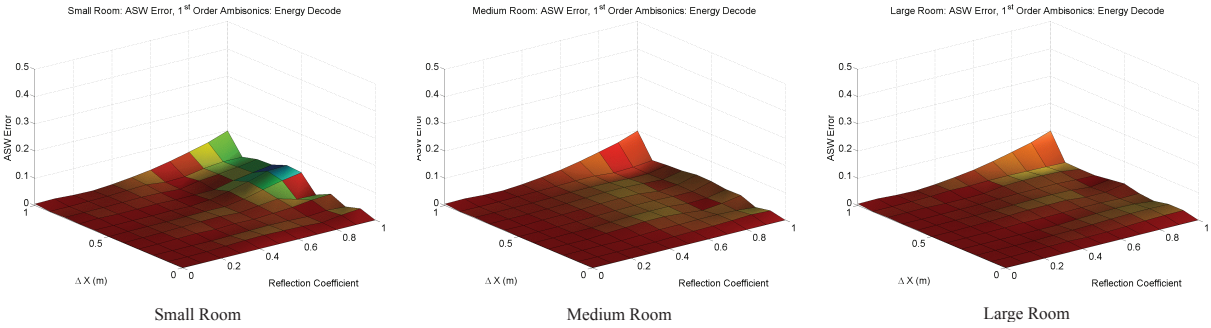
Small Room Medium Room Large Room

(a) Interaural level difference error for small, medium and large rooms.



Small Room Medium Room Large Room

(b) Interaural time difference error for small, medium and large rooms.



Small Room Medium Room Large Room

(c) Apparent Source Width error for small, medium and large rooms.

Figure 6.23: *Objective binaural analysis of B-Format reproduction in different sized rooms using DTW.*

6.6.1 Wavefield Capture and Spatial Downsampling

A dataset of 32 RIRs was captured in test environment 1 using the same methodology as in Section 6.2. RIRs from 3 different source positions, at $-30°$, $0°$ and $15°$ relative to the listening

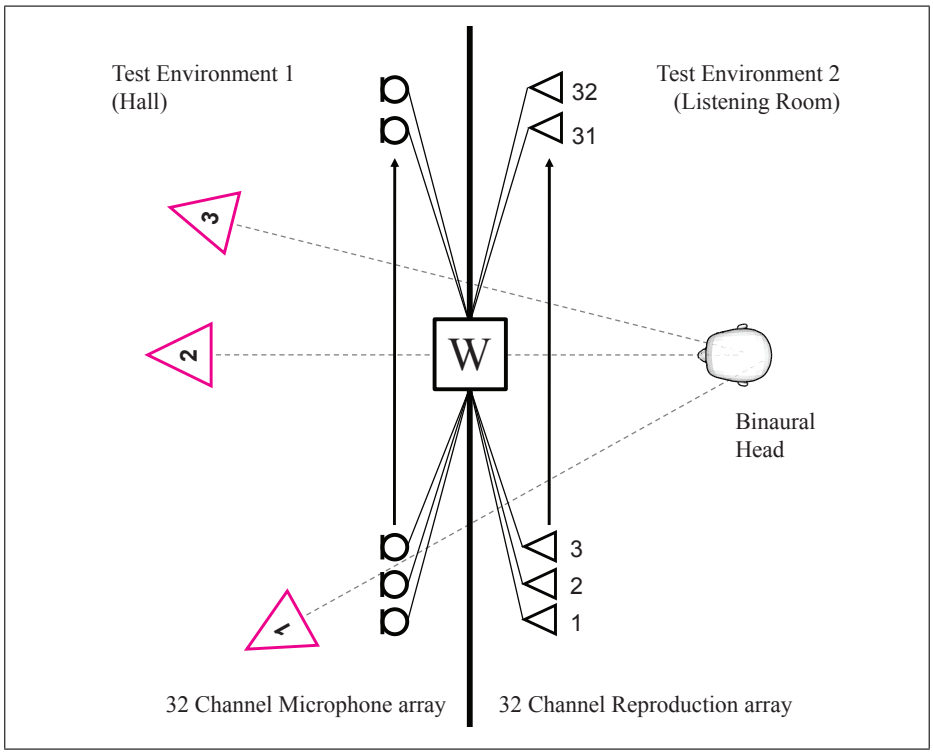


Figure 6.24: *Experimental Setup for impulse response capture and reproduction.*

S	4 RIRs	8 RIRs	16 RIRs
1	1,6,19,32	1,6,10,14,18,23,28,32	1, 4:2:32
2	1,16,17,32	1,6,11,16,17,22,27,32	1:2:29, 32
3	1,11,22,32	1,6,11,17,22,26,29,32	1, 4:2:32

Table 6.2: *Impulse response measurements utilized in interpolation (S = Source Position).*

position shown in Figure 6.24, were captured for the experiments. The radial distance from the listener position was 2.5m. A Genelec 1029A loudspeaker was used as the source, and a Soundfield MK5 system was used to record the dataset.

The transition time of the RIRs was calculated from Equation 6.3 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 are used, i.e. positions 1 and 32, and for an array with an even number of microphones such as the one here, it is also advantageous to use the microphone position closest to the source, in order to avoid singularities occurring in equations 6.7 and 6.9. Table 6.2 shows the real measurements used to create each 32-channel dataset.

The interpolation of the early reflections for the forward facing pressure gradient component can be seen in Figures 6.25, 6.26 and 6.27. The top left diagram in each figure shows the

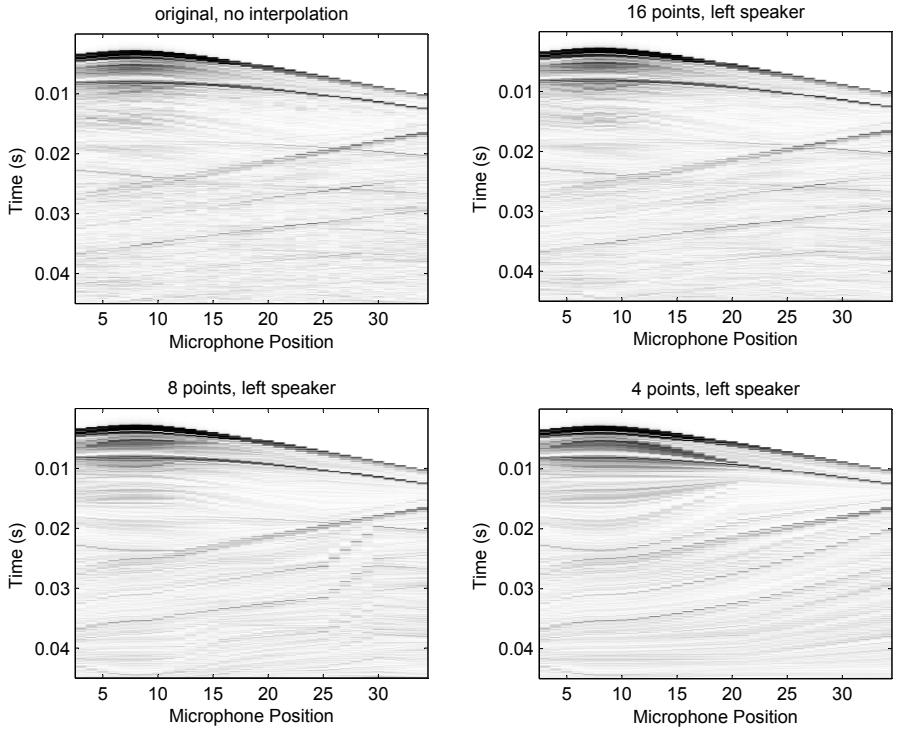


Figure 6.25: *Wave Field analysis for original and reduced datasets of source 1 (30°).*

wavefront reconstruction based on the data at all the 32 receiver positions. The plots show that the direct path and first reflections are interpolated well for as low as four microphone positions. We also note the distortion of several subsequent reflections when only 4 or 8 positions are taken.

6.6.2 Objective Analysis of the Reproduced Wavefields

WFS datasets created by the processes described in Section 6.2.1.1 and 6.6.1 were convolved with the Log-TSP signals for reproduction over the array. The listening room that houses the array has an RT60 of 0.35 seconds. A Neumann KU100 dummy head was used to capture the playback of the array at the specified listener position. The binaural signals were then analyzed in terms of the ITD-PHAT, ILD and IACC_{E3}. Table 6.3 shows the results obtained from the binaural measurements. In the case of the Interaural Time Difference, we see that virtually no compromise is made between the full and interpolated datasets and localization accuracy remains consistent. The early and late IACC measurements also show good degrees of correlation, and show that the apparent source width remains similar between full and reduced datasets. Minor fluctuations are present, attributed to early reflection distortion as data sets are reduced. $\mathrm{W_{IACC}}$ also remains consistent, again showing little change in ASW. The late IACC shows significantly lower correlation values than the $\mathrm{IACC_{E3}}$ indicating that decorrelation is achieved as expected.

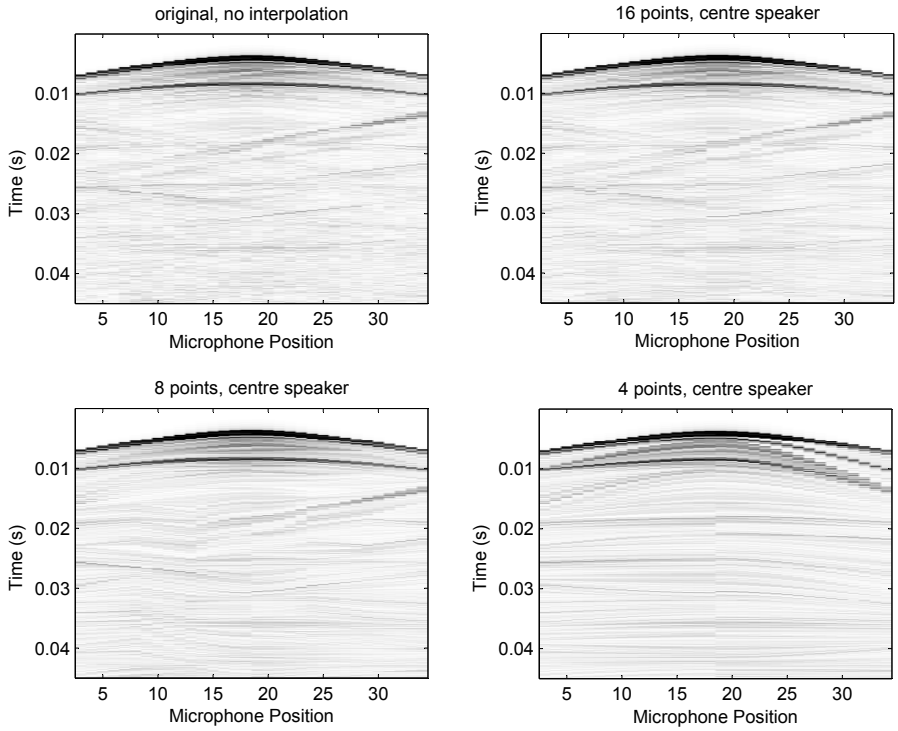


Figure 6.26: *Wave Field analysis for original and reduced datasets of source 2 (0°).*

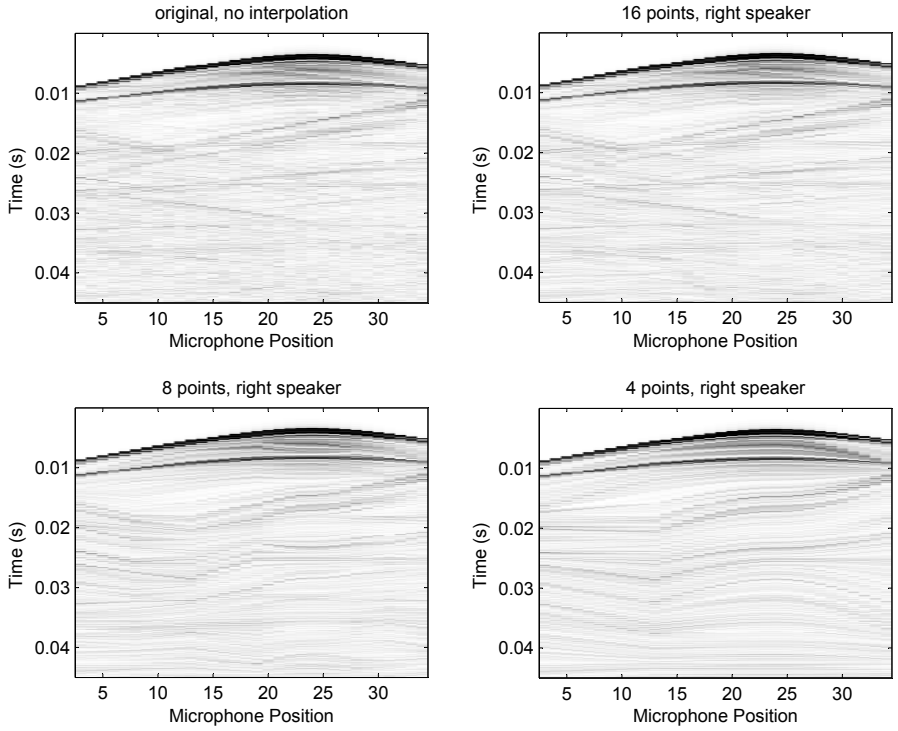


Figure 6.27: *Wave Field analysis for original and reduced datasets of source 3 (15°).*

	RIRs	$IACC_{E3}$ (ms)	ITD-PHAT (ms)	W_{IACC}	$IACC_{L3}$
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 6.3: *Binaural parameters of reproduced playback.*

6.6.3 Perceptual Localization

A series of perceptual experiments was implemented to further investigate the reproduced wavefields. The primary aim of these tests was to investigate the localization performance using the 32 channel WFS array for 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 f_a. 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 localization tests. Each participant was located at 1.5m centre-front to the array. 10 subjects were chosen, each under 35 years of age, of excellent hearing, and with good music production experience.

For each pink-noise presentation the participant was asked to identify the direction of localization using a dedicated software pointer, projected onto an acoustically transparent screen in front of the array. This setup is shown in Figure 6.28. Listeners were allowed to rotate their heads, but were asked to not move it from the centre position to accommodate the geometry of the projected pointer. 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 6.29.

We see that over all sources a reasonable azimuthal localization performance is achieved. Azimuthal deviations are greatest for the 15° source, which from the objective analysis holds the lowest $IACC_{E3}$ value (0.585 for no interpolation) as well as the largest W_{IACC} (0.194 for no

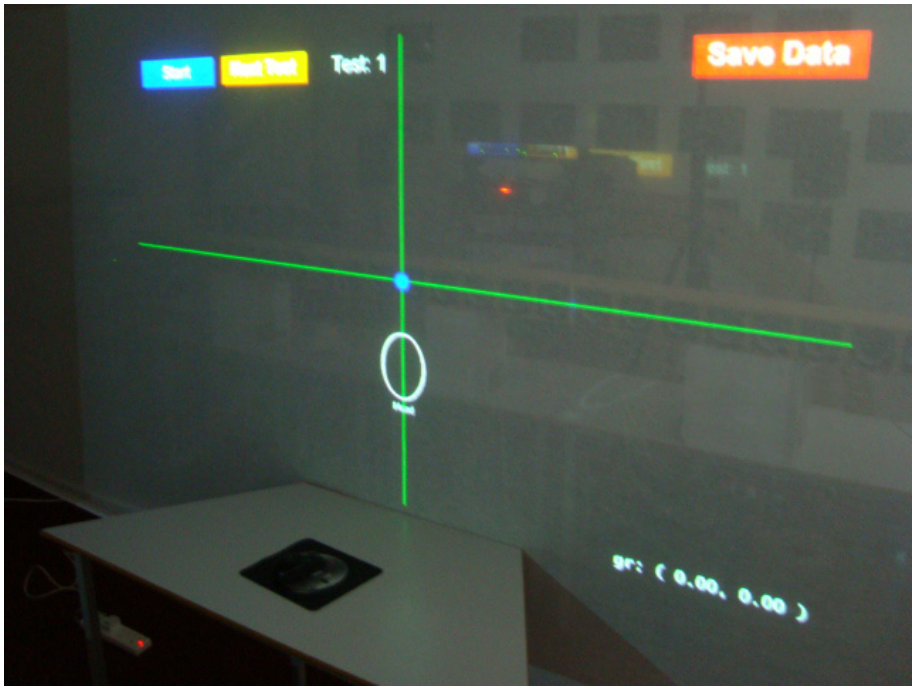


Figure 6.28: *Software Pointer Projected onto array position.*

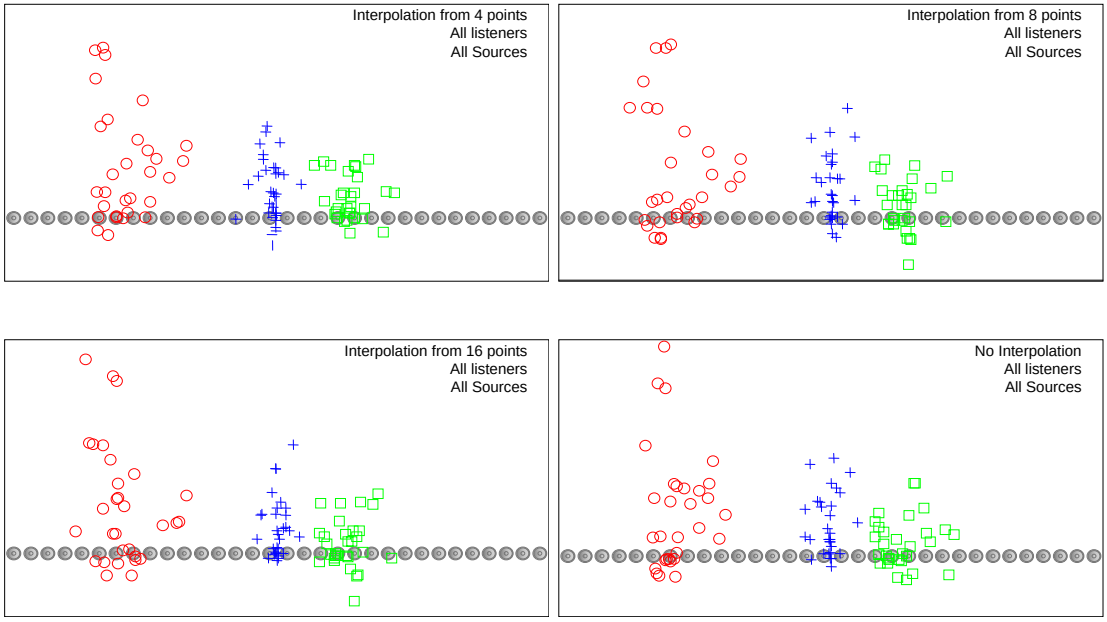


Figure 6.29: *Scatter plots of localization results over all listeners and all sources.*

interpolation). The best localization is achieved for the centre source, which holds the highest IACC_{E3} value (0.612 for no interpolation). It is difficult however to gauge the true localization performance from these plots. Instead, we must look at the statistical performance of source localization over each type of presentation. The results of the mean and standard deviation of

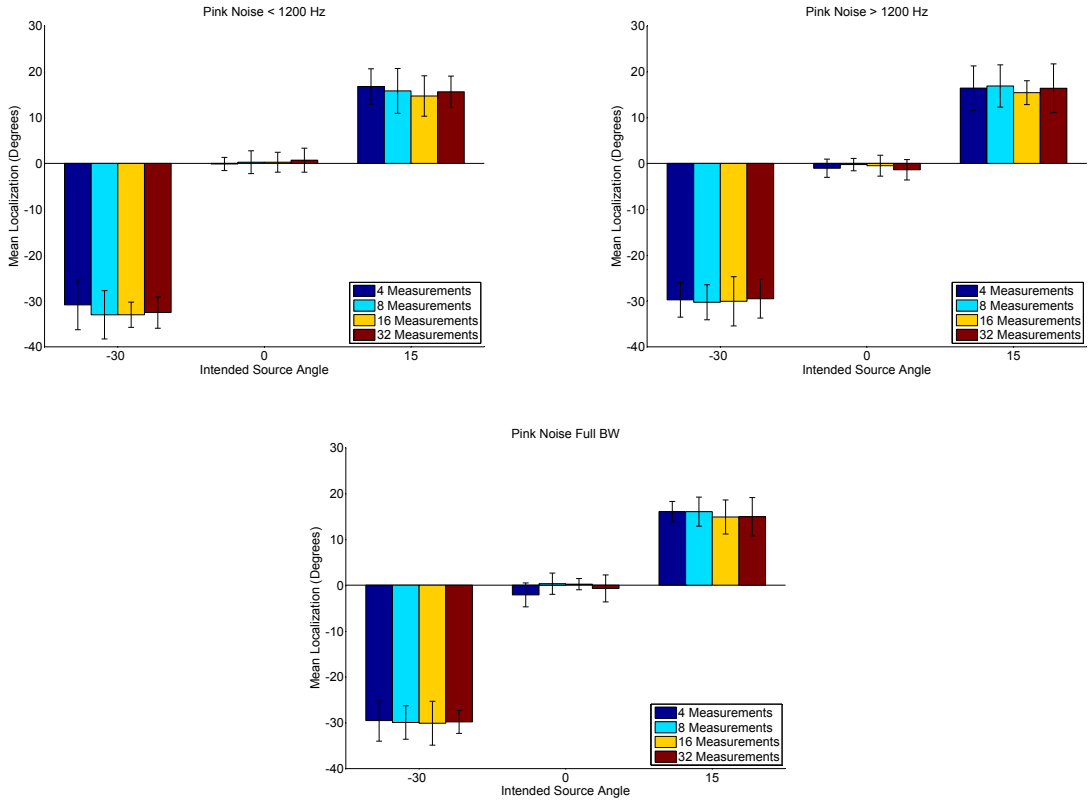


Figure 6.30: *Mean Azimuthal Localization of filtered (below and above f_a) and full bandwidth pink noise. (95% Confidence Interval)*

localization are plotted in Figure 6.30.

In the case of the presentations below f_a, the mean localization fluctuations between datasets is marginal and equivalent localization is achieved. Note that the largest deviations 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. Localization 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 localization is again comparable.

6.7 SRIR Topology Based on DTW Interpolation Algorithm

Based on the objective and subjective analysis, we see that DTW is an effective method for interpolation of acoustic responses whereby localization accuracy is maintained, and as well as the spatial properties of the reproduced soundfield. Given that interpolation can be effective up to a 1m distance for medium to large sized rooms, a topology, such as the one shown in Figure

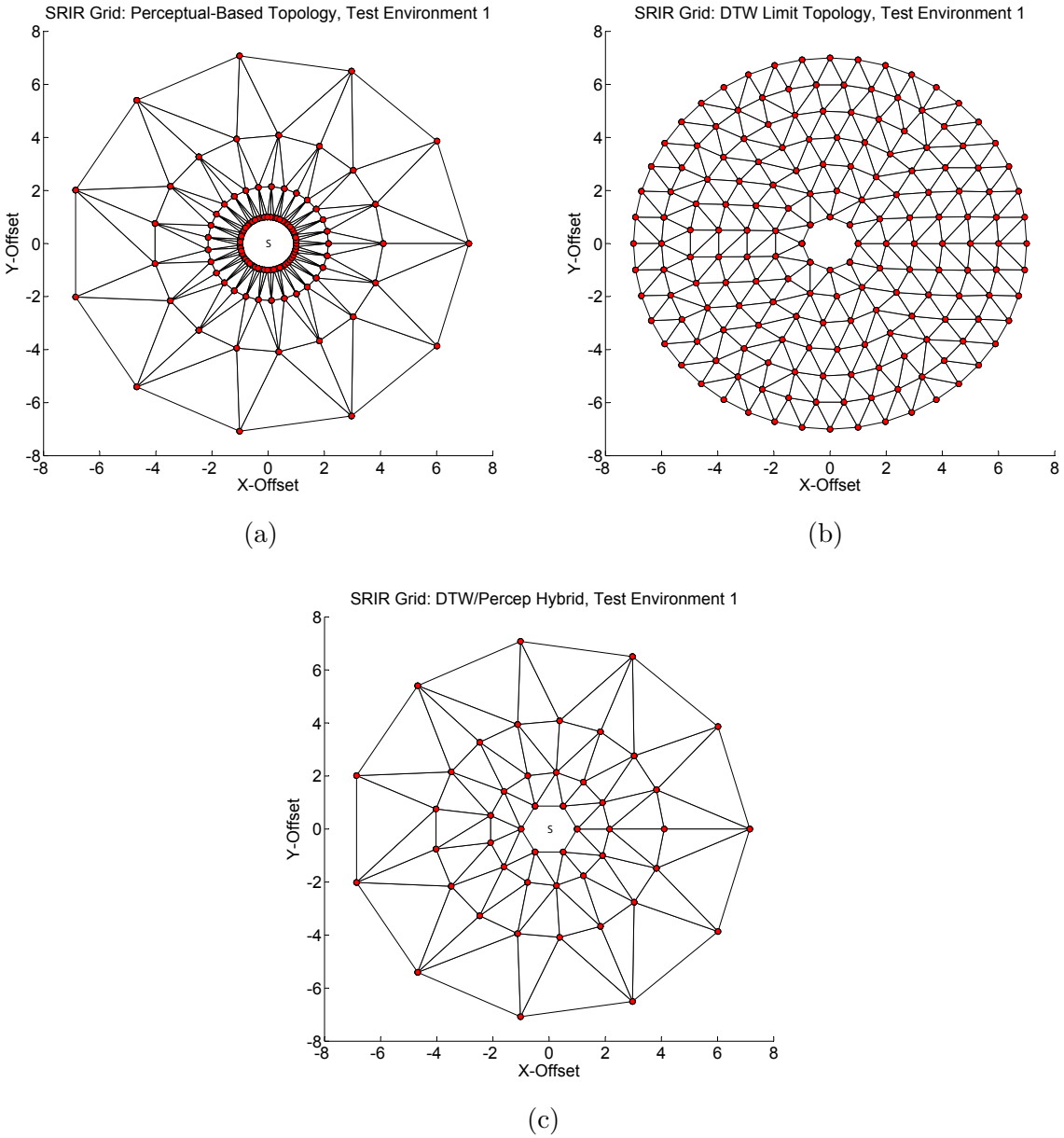


Figure 6.31: *SRIR Topologies (a) Perceptual Analysis (b) DTW interpolation (c) DTW/Perceptual hybrid.*

6.31 (b) can be derived, where each SRIR is spaced 1m apart.

This is clearly an unfeasible amount of SRIR measurements, but when combined with the perceptual derived topology of section 6.2.2 (shown again in Figure 6.31 (a)), we arrive at the sparse topology of Figure 6.31 (c). This layout is the same as the perceptual based topology, with the exception that the high density regions of SRIRs (in the 1m and 2m circles) have been replaced by sparse SRIR measurements separated close to 1m apart. A higher density of SRIRs

can then be interpolated based on the measured database.

An important aspect of this topology is the placement of the source within the acoustic environment. If the source is close to boundaries, then the topology must be truncated and constrained, such that the number of SRIR measurements reduces further. An example of this is shown in Figure 6.32 for Test environment 1. We see that only two measurements are required from the outer ring in the topology, with further SRIRs removed from the inner rings. Only the 1m ring preserves the six SRIR measurement points originally specified in the topology. There are 28 measurements in total for this particular source position. If the source is located at the centre of the room, then we can also assume acoustic symmetry and the number of SRIR measurements can be reduced further.

6.8 Conclusions

In this chapter, we have investigated virtual acoustic recording utilizing a novel method for interpolation of SRIRs. The method uses Dynamic Time Warping to align the main reflection points in SRIRs prior to interpolation and was shown to avoid the smearing distortions that occur in linear interpolation processes. A perceptual based evaluation of the interpolated impulse responses was presented, and within the scope of this study, the method was found to work well for medium to large rooms with effective interpolation up to 1m spacing under extreme reflection conditions. The studies presented here indicate that the method will work well for other rooms of larger dimensions. However, the method in its current form is not suited for interpolation in small rooms, since the early reflections are not as sparse. Perceptual studies also showed that the sparseness in spatial impulse measurements based on data pertaining to the perception of the acoustic environment. Topologies based on the DTW interpolation algorithm as well as the perceptually derived localization data were shown that allow for more practical capture of real acoustic soundfields for walk-through auralization. We must now consider how we can combine these measured B-Format responses with the source audio for both loudspeaker and headphone reproduction using a higher order Ambisonics approach. This is the focus of the following chapter.

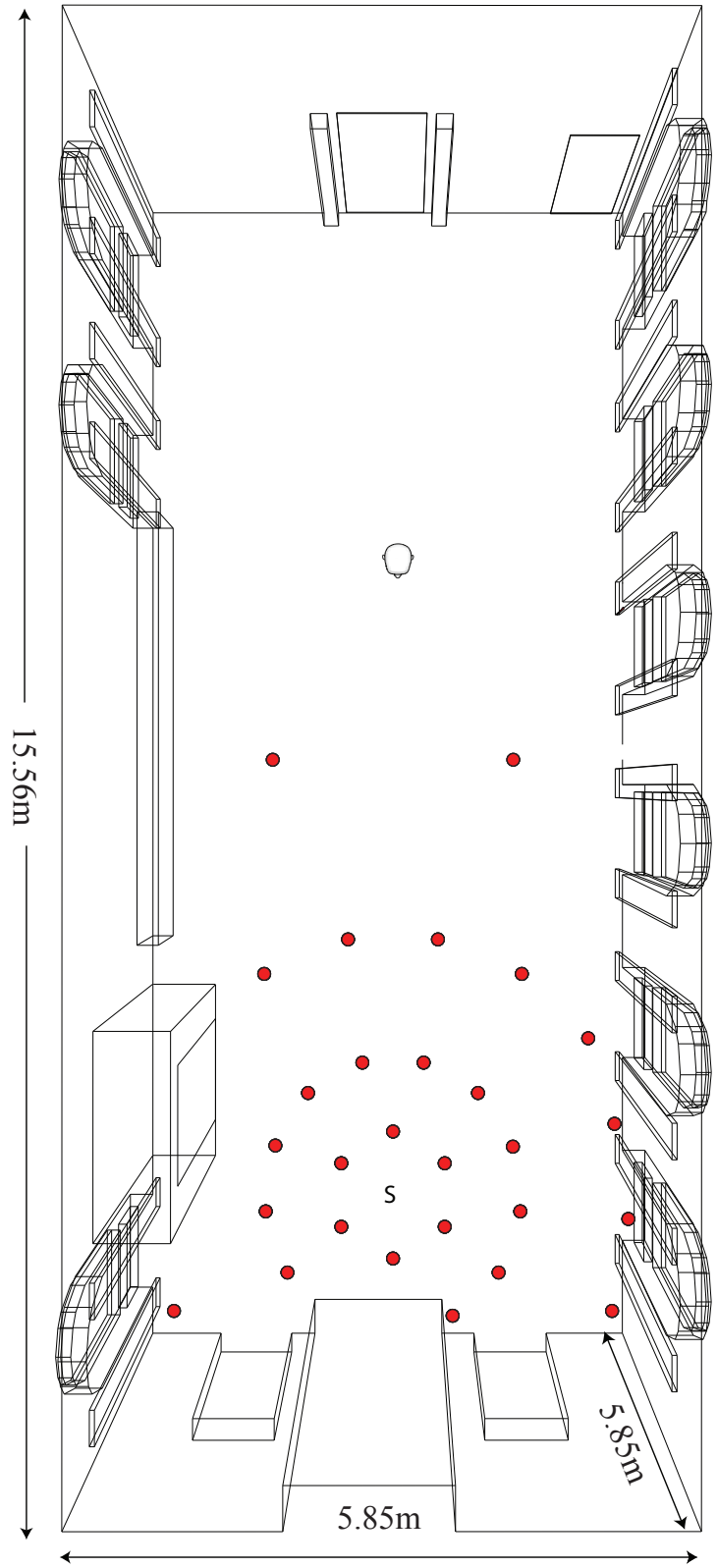


Figure 6.32: *Application of DTW/Perceptual topology to SRIR measurement in Test Environment 1.*

7

Virtual Acoustic Reproduction

7.1 Introduction

In Chapter 3 the concept of virtual loudspeaker reproduction was introduced. It was shown that this method allows for the same spatialization scheme to be adopted for both headphone and loudspeaker reproduction. In this context, binaural synthesis is regarded in itself as a reproduction method, where the listener is always in the sweet spot of a virtual loudspeaker array. Thus, for each virtual loudspeaker, we require measured HRIRs that describe the relationship between the individual transducers of a real loudspeaker array and each ear. The virtual loudspeaker feeds are derived from an Ambisonics decode of soundfield recordings, which can be transformed for head rotations according to the equations of Section 3.4. A major advantage of this method is that since the virtual loudspeakers are always static relative to the listener, the HRIR filters do not have to be switched in real-time. As a consequence, optimal reproduction of higher order Ambisonics over headphones is limited only by the computational complexity of the filter kernel. For example, a 3^{rd} order Ambisonic system requires 16 virtual loudspeakers for full periphonic surround reproduction, satisfying the diametric decoder theorem. This is achieved by convolving each of the 16 virtual loudspeaker feeds in real-time with the corresponding HRIR pair, resulting in 32 real-time filters. In systems such as games consoles, where the amount of software processing dedicated to all audio functions is small relative to real-time video processing demands, potential latencies can be avoided by shortening each HRIR filter length. However, as has been shown in [Sandvad and Hammershøi, 1994], lower order filtering can also lead to detectable artifacts. In this chapter, it is demonstrated that if the source material is predefined,

reduction of the filter order is not necessary if we factorize the HRIR into directional dependent and independent components and pre-process signals with the independent component. It is also vital that the spatial cues introduced by the HRTFs are not corrupted in a significant way by reproduction through headphones. Hence it is necessary to equalize for both the transfer function of the headphone used in reproduction and the coupling between the headphone and the ear. These two components are encapsulated by the Headphone-Pinna Impulse Response (HPIR) which is dependent both on the listener and the positioning of headphones. Finally, we must also consider how best to employ higher order Ambisonics for both virtual and real loudspeaker reproduction, given that we have only measured 1^{st} Order Ambisonic soundfields.

This chapter is therefore organized as follows: A novel approach to reduction of HRIR filters in a virtual loudspeaker setup is shown based on the work of Masterson in [Masterson and Boland, 2008] and developments in [Kearney et al., 2009a] and [Kearney et al., 2009b], and is outlined in Section 7.2. In Section 7.2.2 the algorithm is applied to HRIRs from the CIPIC database [Algazi et al., 2001] and the results presented for an 8 channel virtual loudspeaker array. The importance of the HPIR and its subsequent equalization is then addressed in Section 7.3. Finally, a method of repurposing 1^{st} Order Ambisonic soundfields for higher order Ambisonic reproduction is presented.

7.2 Headphone Reproduction: Virtual Loudspeaker Optimization

To create a virtual loudspeaker for binaural playback it is necessary to convolve the virtual loudspeaker feed, $x(t)$, with the left and right ear HRIRs, $h_l(t)$ and $h_r(t)$. Most existing research uses a block frequency domain approach to this convolution. However, given that the virtual loudspeaker feeds are controlled via head-tracking in real-time, a time-domain filtering approach can also be utilized. For *short* filter lengths, obtaining the output in a point wise manner avoids the inherent latencies introduced by block convolution in the frequency domain [Scarpaci, 2006]. In this context, it is desirable to reduce the filter length for real-time application. We therefore consider splitting the HRIR into two components, one which is static with relative changes in source-listener position (termed the Angle Independent Component or AIC) and one which is position dependent (termed the Angle Dependent Component or ADC). The convolution of the sound source with the directional independent component can be completed offline and stored. Hence a shorter filter would be applied and changed with relative listener-source movement in real time.

Many researchers have explored the topic of extracting a directional independent component from HRTF data. Several have taken the approach of taking a simple average across the data set, see for example [Kistler and Wightman, 1992]. Møller [Møller, 1992] implements "diffuse field equalization", which involves computing a diffuse reference spectrum by power averaging all

the HRTFs for one ear and finding the square root. This reference spectrum is then deconvolved from the original HRTFs to produce the equalized HRTFs. In [Middlebrooks and Green, 1990] the authors refer to splitting the HRTF into a two stage transfer function; the "directional transfer function" (DTF) and the "common transfer function" i.e. that which is directionally independent. Their method was to compute an average magnitude spectrum of the HRTF data and apply minimum phase assumptions. Freeland et al [Freeland et al., 2007] introduce a HRTF reduction and interpolation method which utilizes inter-positional transfer functions (IPTFs). These IPTFs are ratios of HRTFs from contiguous locations which are simplified using balanced order reduction and subsequently can be used to interpolate one HRTF given three. Haneda et al [Haneda et al., 1999b] 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.

However, here, it is proposed that a HRIR can be considered as the convolution of two FIR filters, a direction independent filter and a direction dependent filter. This direction dependent filter is extracted using an iterative least squares method. The removal of a common factor across the HRIR dataset is motivated by the need for adaptive real-time convolution to allow for a more interactive virtual auditory space. The convolution with the common factor could be completed off-line and stored leaving a shorter HRIR that would change with a relative movement between source and receiver.

7.2.1 Algorithm

Consider that we are given N HRIRs measured at angles ϕ over the sphere (or over a subsection of a sphere) centered on a head denoted as h^ϕ where $\phi = 1, ..., N$. The objective is to extract a common subsystem, f, assuming that

$$h^\phi = f * g^\phi \tag{7.1}$$

where g^ϕ are the residuals and * denotes convolution. A HRIR is modeled by a FIR filter with the impulse response samples as its coefficients. The algorithm used in finding this common subsystem of a HRIR dataset is equivalent to finding the approximate greatest common divisor (AGCD) of the HRIR z-domain FIR filter set. We formulate the task of finding the AGCD as a non linear optimisation problem:

$$\hat{f} = \arg\min_f \sum_{\phi=1}^{N} \|h^\phi - (f * g^\phi)\|^2 \tag{7.2}$$

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

The problem as posed here is a non convex, non linear optimisation problem with multiple local minima. There is a clear equivalence between finding a common subsystem amongst a set of HRIRs and finding a common root system amongst the equivalent z-domain HRTF set. Authors such as Zeng [Zeng and Dayton, 2004], Corless [Corless et al., 1999] and Chin [Chin et al., 1998] 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 of which there may be dozens (even hundreds) simulating a given virtual loudspeaker array, and finding a global minimum is unlikely. Our proposed iterative least squares method is equivalent to the divisor quotient method described by Chin et al [Chin et al., 1998] 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.

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 linearization 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 minimize 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.

By forming F the convolution matrix of f we can rewrite the above convolution as a matrix multiplication,

$$Fg^{\phi} = h^{\phi} \tag{7.3}$$

It is possible to solve for each residual, g^{ϕ}, given an initial guess f_0 by

$$g_{i+1}^{\phi} = F_i^{\dagger} h^{\phi} \tag{7.4}$$

where F_i is the convolution matrix formed from f_i and $\dagger$ denotes the Moore-Penrose pseudoinverse. We solve for f_{i+1} using

$$f_{i+1} = \begin{pmatrix} G_{i+1}^1 \\ \vdots \\ G_{i+1}^N \end{pmatrix}^{\dagger} \underline{h} \quad \text{where} \quad \underline{h} = \begin{pmatrix} h^1 \\ \vdots \\ h^N \end{pmatrix} \tag{7.5}$$

where G_{i+1}^{ϕ} is the convolution matrix formed from g_{i+1}^{ϕ}. Equations 7.4 and 7.5 define a recursion, where i is the iteration count, that takes an estimate f_i of the coefficients of a direction independent component (of a given length), and maps them onto a new estimate f_{i+1}. This processed is summarized as:

i=iteration count

1. Guess f_0 $(i = 0)$
2. Solve for each residual, g^{ϕ},
3. Solve for f_{i+1},
4. Set $i = i + 1$ and repeat steps 2 and 3 until there is convergence.

7.2.2 Factorization Analysis using the CIPIC Dataset

The algorithm described in Section 7.2.1 was applied to HRTF data from the KEMAR binaural mannequin (Subject 21) from the CIPIC database. For this example, an 8 channel virtual loudspeaker array was created from the HRIR measurements taken 45° apart. Figures 7.1 shows reconstructed HRIRs and their corresponding HRTFs, for each virtual loudspeaker, where a 140 sample long common subsystem has been extracted from the 8 measurement dataset, and compares them to the original unfactorized 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 140 samples of an average taken over the HRIR dataset for a given ear and the third is a 140 sample long random vector. The comparison is shown in both the time and frequency domain.

It is evident in the time domain plots that there is negligible difference between the original and the reconvolved HRIRs for each initial condition, given a 140 sample long common component. However, the frequency domain plots show a small mismatch visible for each position on the azimuth at high frequencies (>18kHz), especially when the average initial condition is used. Particular distortion occurs at the contralateral ear for the 90° and 270° HRTFs. This is because these measurements are at the point of maximum cross head delay, and an angle independent component of 60 samples is insufficient in reconstructing the information pertaining to the time-domain energy of the HRIR and its ITD (given that the maximum cross head delay for the KEMAR mannequin is approximately 28 samples at a sampling rate of 44100). This is further supported by the fact that the reconstruction is better for all initial conditions for the ipsilateral HRIRs for the the 90° and 270° HRTFs.

Whilst a length of 140 samples is used here 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 7.2 plots the mean squared error versus the length of the common subsystem for KEMAR. The mean square error measure describes the difference between the entire original HRIR dataset and the reconvolved HRIRs for each initial condition. We see that for a source at 90° the error increases in the contralateral ear as we increase the length of the angle independent component. Above 140 samples, the normalized mean square error rises close to 1 for the all-ones and random initial condition cases. The error with averaged initial condition

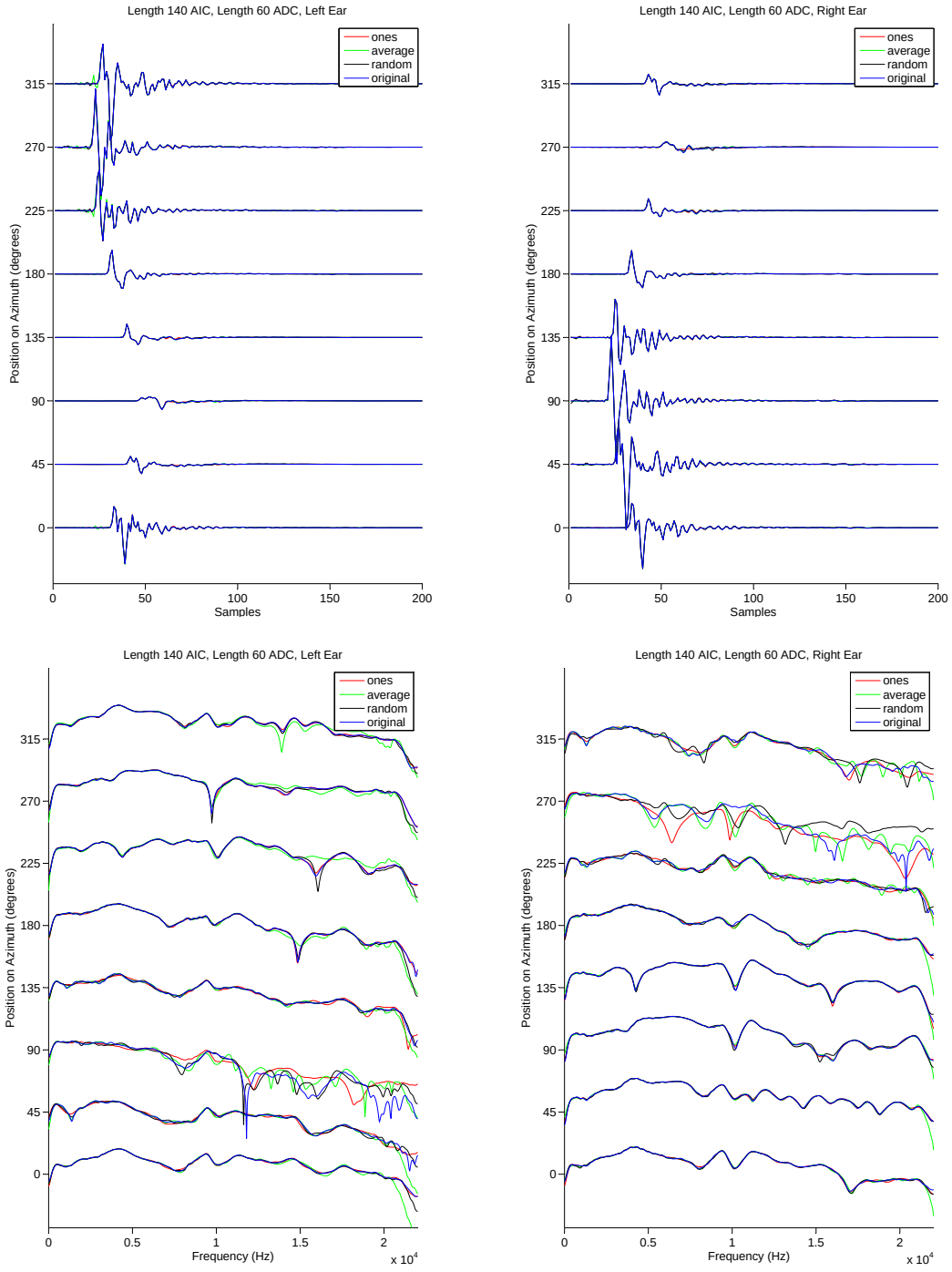
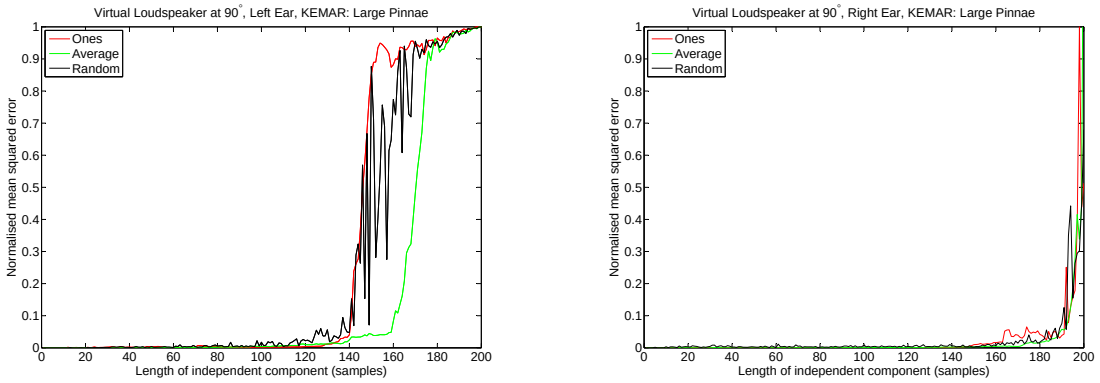


Figure 7.1: *Comparison of original HRIR/HRTFs to reconvolved HRIR/HRTFs of Subject 21 with different initial f_0 guesses, for an 8 channel virtual loudspeaker array (Length 140 ADC, Length 60 AIC).*



(a) Left ear (b) Right ear

Figure 7.2: *Normalized mean squared error of different initial f_0 conditions, for virtual loudspeaker at 90°, (a) KEMAR left ear (b) KEMAR right ear.*

becomes significant at an approximately 160 sample long directional independent component. This verifies the observations of Figure 7.1. Conversely at the right ear, the error is significantly low for all initial conditions until a directional independent component of approximately 180 samples long, after which the error increases rapidly. An example HRIR/HRTF dataset for a 180 long angle independent component is shown in Figure 7.3. The effect on the spectrum is clearly seen, and large errors occur for all initial conditions. Again the error is greatest for the 90° and 270° sources at the contralateral ear in each case.

In order to reduce the mean-squared error and to create shorter run-time filters, we will consider the minimum-phase versions of the HRIRs. This is justified by considering the work of Mehrgart and Mellart [Mehrgart, 1977] who show that HRTFs display close to minimum phase properties up to 10kHz. The minimum phase HRIR will have no time-lag and as a consequence will result in a reduced length HRIR. ITDs can then be reintroduced by a pure time lag corresponding to the ITD for the HRIR pair. The resultant error in the factorization process using minimum phase filters is therefore significantly reduced as shown in Figure 7.4. We now see that it is possibly to push the length of the angle independent component further. Figure 7.5 shows the resultant time domain and frequency domain plots of the minimum phase HRIR/HRTFs with the same conditions as before. We see that the error is significantly reduced with small spectral deviations occurring above 16kHz, even though the length of the run-time filters is only 20 samples long.

The application of the method is, of course, not limited to the simple case of an 8-channel virtual loudspeaker array. Large numbers of HRTFs can be utilized for a greater number of virtual loudspeakers. For example, Figure 7.6 compares the original HRTFs for both subject 3 and subject 21 in the CIPIC database to reconstructed HRTFs after a 140 sample directional dependent component had been extracted. The HRTFs shown are for 0° elevation and angles −45° to 45° in azimuth. The two plots are almost identical except for small discrepancies at

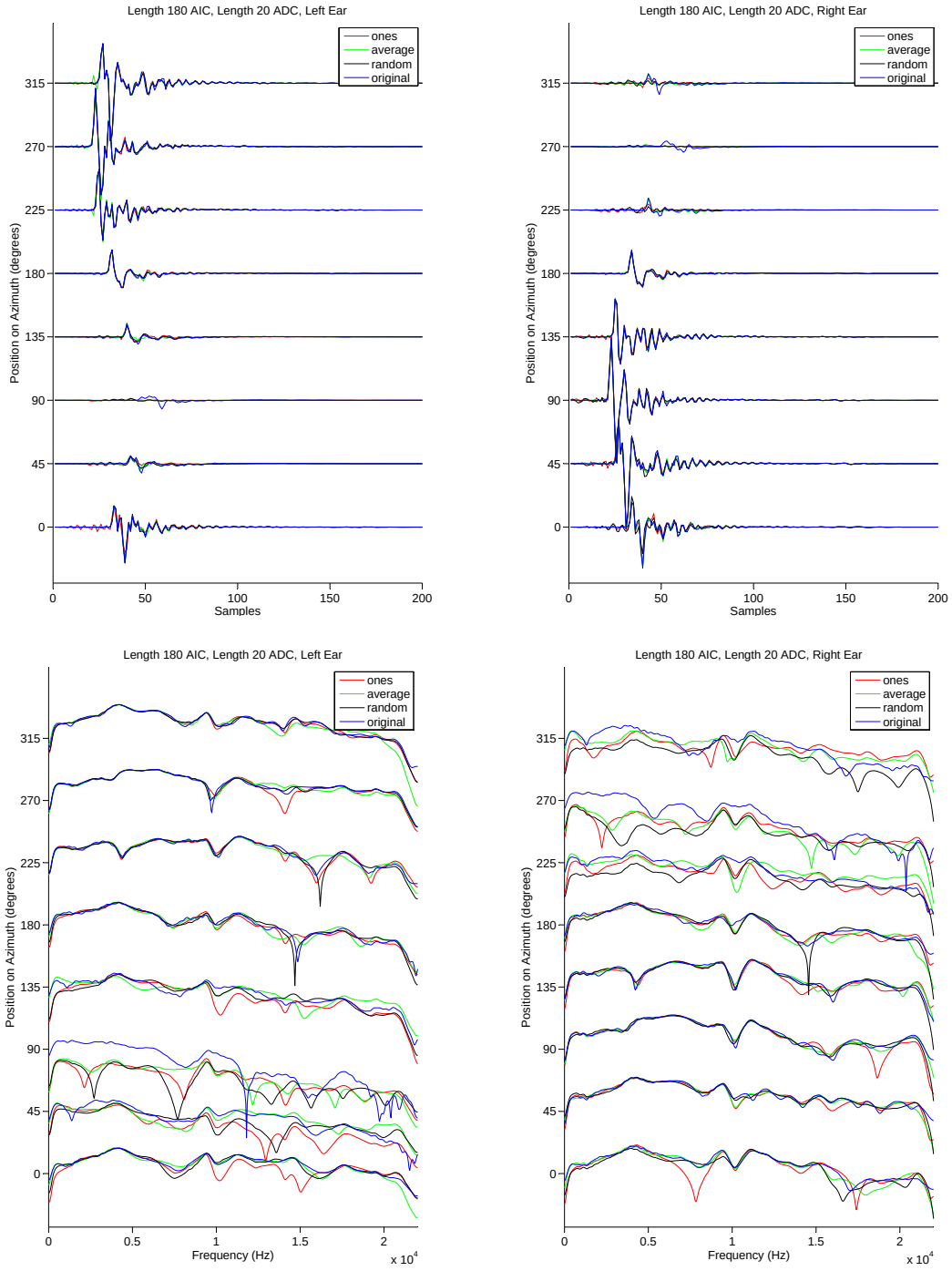
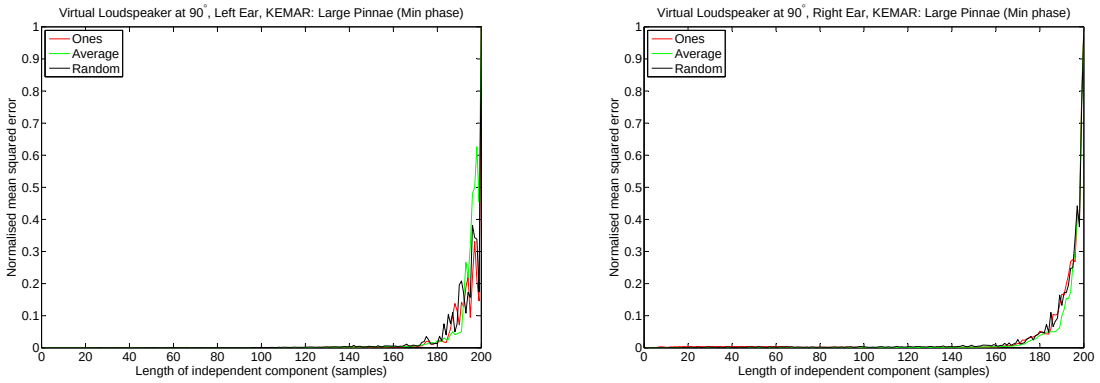


Figure 7.3: *Comparison of original HRIR/HRTFs to reconvolved HRIR/HRTFs of Subject 21 with different initial f_0 guesses, for an 8 channel virtual loudspeaker array (Length 180 ADC, Length 20 AIC).*



(a) Left ear (b) Right ear

Figure 7.4: *Normalized mean squared error of different initial f_0 conditions, for virtual loudspeaker at 90°, (a) KEMAR left ear (b) KEMAR right ear.*

very high frequencies (above 16kHz) which have little perceptual relevance for average listeners.

7.3 Headphone Equalization

In the same way that we have compensated for the coloration effects in the convolution chain in Chapter 5, it is also important to consider the possible corruption of spatial cues due to headphone reproduction. It is therefore necessary to compensate for both the headphone transfer function and the coupling of the headphone to the ear canal. These two components are characterized in Headphone-Pinna Transfer Functions (HPTFs) which are dependent both on the individual's ear physiology, as well as headphone positioning. Møller et al. [Møller et al., 1995] have demonstrated that HPTFs exhibit smooth fluctuations below 5kHz and above this, exhibit significant, high-Q variation across individuals and different headphone types. This has also been corroborated in [Pralong and Carlile, 1996].

In order to gauge the extent of the high frequency fluctuations in HPTFs with headphone positioning, a set of measurements were taken using a Nuemann KU100 dummy head with two high-grade sets of headphones, (Audio Technica ATH-ANC7 and Sennheiser HD650). Four different headphone-ear positions were examined and the transfer functions were captured using the exponential sine tone sweep method [Farina, 2000]. The resultant HPTF magnitude spectra are plotted in Figures 7.7 and 7.8[1]. It can be seen that the measurements are in agreement with the observations of previous authors. Low frequency deviations are exhibited (in particular for the ATH-ANC7 headset), as well as the high frequency notching in the range of 6-11kHz. Thus, significant ILD distortions will occur on any spatialized audio presented to the headphones. Such

[1]These magnitude spectra were measured by Stephen Adams, Trinity College Dublin and provided here via private communication with the author.

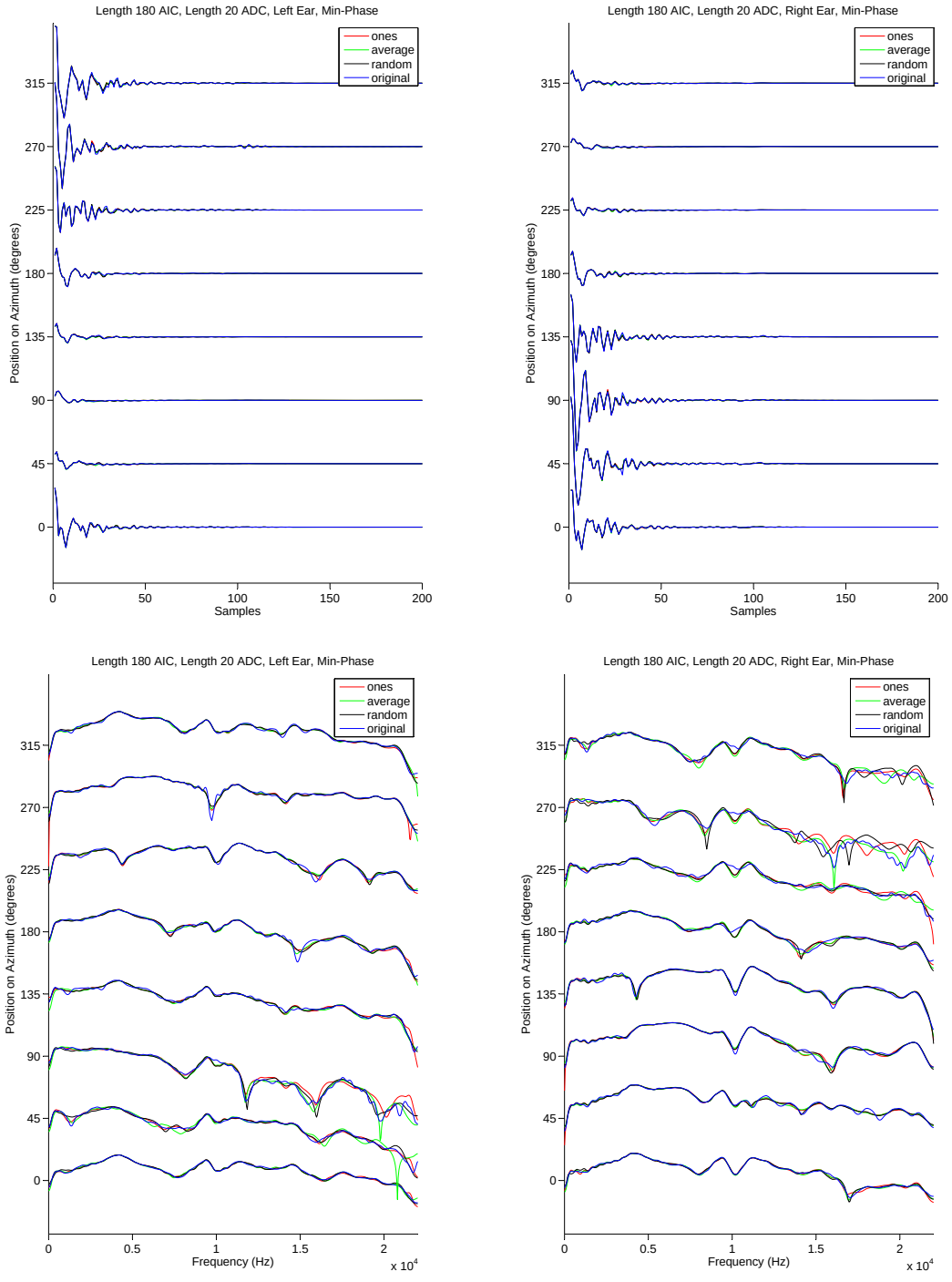


Figure 7.5: *Comparison of original minimum phase HRIR/HRTFs to reconvolved minimum phase HRIR/HRTFs of Subject 21 with different initial f_0 guesses, for an 8 channel virtual loudspeaker array (Length 180 ADC, Length 20 AIC).*

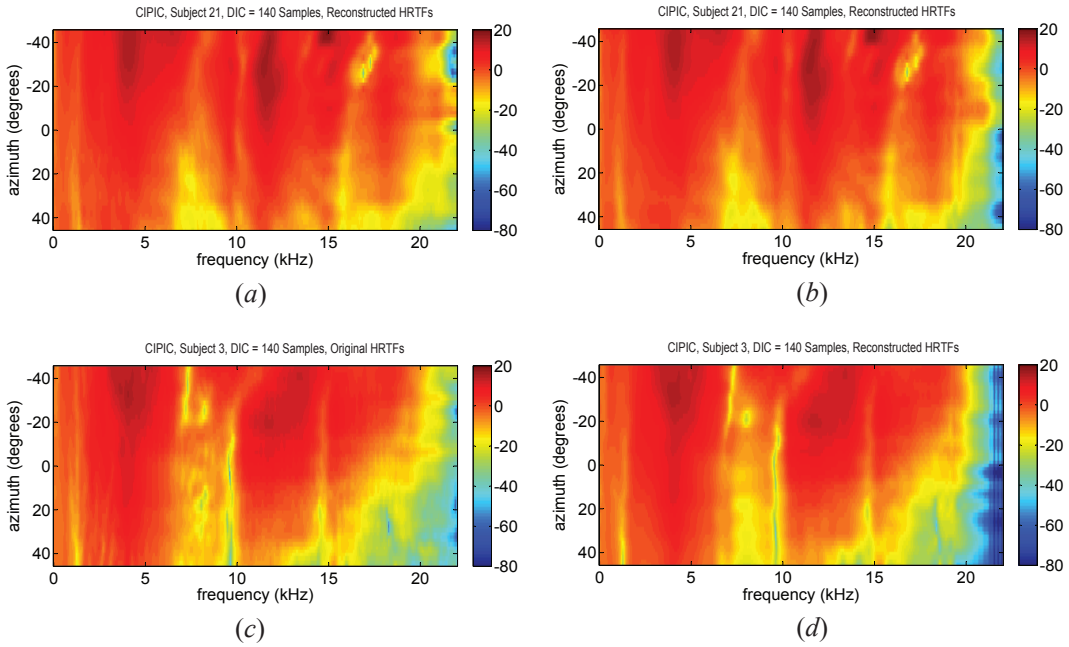


Figure 7.6: *Comparison of original HRIR to reconvolved HRIRs from CIPIC database from* $-45°$ *to* $45°$*: (a) Subject 21 (KEMAR) original HRTFs, (b) Subject 21 (KEMAR) reconvolved HRTFs, (c) Subject 3 (human) original HRTFs (d) Subject 3 (human) reconvolved HRTFs.*

an example of erroneous ILDs are shown in Figure 7.9 for the Sennheiser HD650 headset. The ILD is significant in the range of 6-11kHz, and is due to a mismatch at the notch frequencies. However, it is debatable whether this ILD is sufficient to corrupt localization of broadband stimuli.

Given that these high Q notches are dependent on the position of the headphones on the ears [Hammershøi and Møller, 2008], then equalization becomes difficult. This is due to the fact that the resonant peak measured in the HPTF will lead to an inverse filter design mainly acting on that narrow-band frequency range. However, a full inverse filter will only introduce another resonance (manifesting as unpleasant ringing artifacts in the spatial audio) the next time the headset is put on by the listener, since the position of the notch will have changed. In fact, the presence of high Q resonant peaks will be more perceptible than high Q notches. [Pralong and Carlile, 1996] has found that this phenomenon is less apparent in circumaural type headsets. Therefore, in this thesis it is suggested that 'peak compression' should be performed when equalizing a HPTF. Here, the inverse filter is restricted from creating large peaks in the response, thereby avoiding any negative artifacts. This approach is similar to that suggested in [Laitinen, 2008].

For the inverse filter design, we again use the method of Kirkeby [Kirkeby and Nelson, 1996] as outlined in Section 5.3.3.1. The use of regularization is advantageous here since pushing the

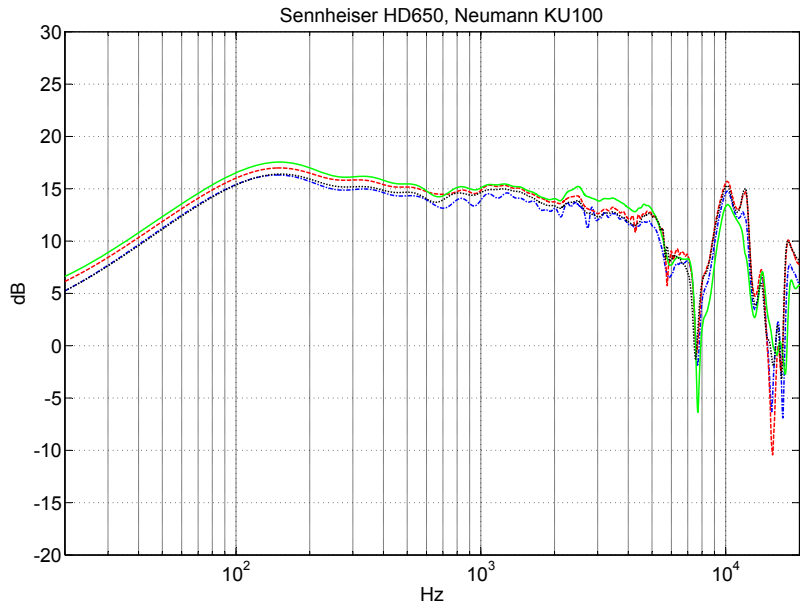


Figure 7.7: *HPTFs at the left ear for 4 random positions of the Sennheiser HD650 headphones.*

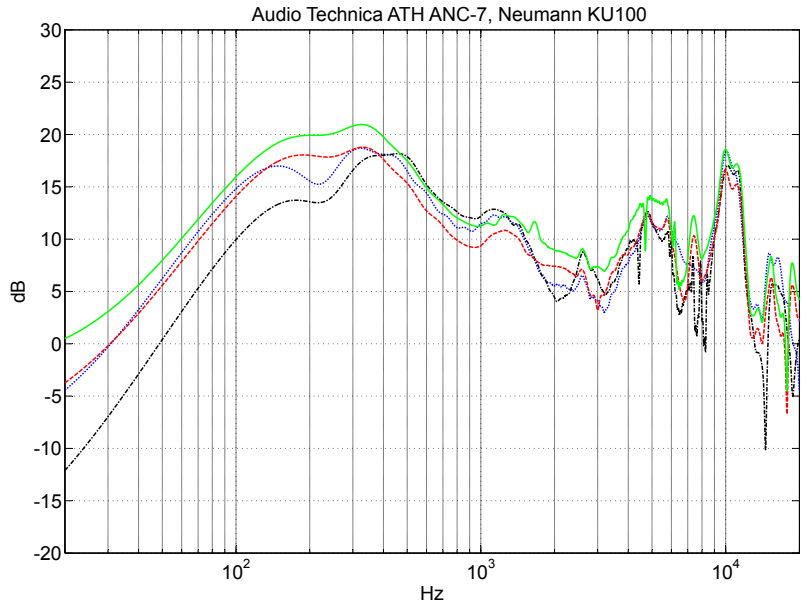


Figure 7.8: *HPTFs at the left ear for 4 random positions of the ATH-ANC7 headphones.*

poles in and away from the unit circle, allows for the required 'softening' of resonant peaks in the response we are looking for. A plot of the typical performance of the regularized least-squares approach for HPTF equalization is shown in Figure 7.10. The equalization shown here is frequency dependent, and the sharp notches shown above 10kHz are not compensated due to their sensitivity to headphone position.

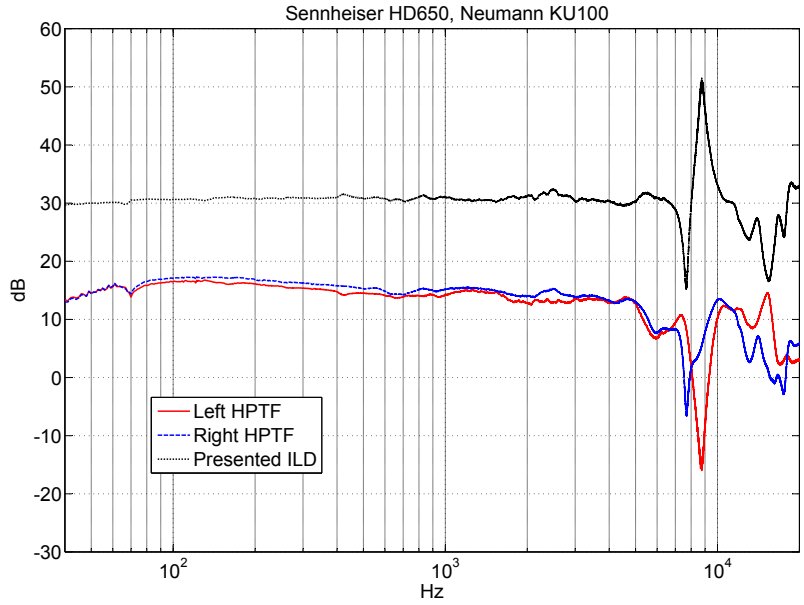


Figure 7.9: *Erroneous ILD presented through Sennheiser HD650 headphones*

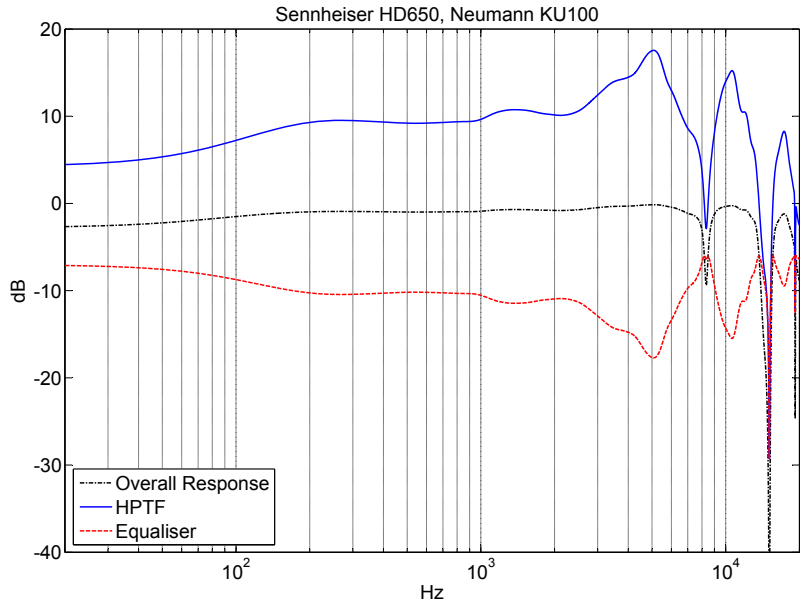


Figure 7.10: *Performance of regularized equalization for left ear HPTF from Sennheiser HD650 headphones and Neumann KU100*

7.4 Higher Order Ambisonic Reproduction over Virtual and Real Loudspeakers

7.4.1 Virtual Ambisonic Reproduction

A further optimization for headphone listening can be employed by encoding the angle dependent components of the HRIRs into spherical harmonic components. Here, any number of HRIRs

can be represented by a set of Ambisonic channels of a given order. For example, for first order Ambisonics, we can encode, for the left ear, a 16 channel periphonic virtual loudspeaker array into the four Ambisonic channels of first order by

$$W_{HRIR_L} = \sqrt{2}\sum_{i=1}^{16} h_{i_l} \tag{7.6}$$

$$X_{HRIR_L} = \sum_{i=1}^{16} \cos(\theta_i)\cos(\phi_i)h_{i_l} \tag{7.7}$$

$$Y_{HRIR_L} = \sum_{i=1}^{16} \sin(\theta_i)\cos(\phi_i)h_{i_l} \tag{7.8}$$

$$Z_{HRIR_L} = \sum_{i=1}^{16} \sin(\phi_i)h_{i_l} \tag{7.9}$$

where h_{i_l} is the i^{th} left ear HRIR. Similar relations hold for the right ear. The left and right ear signals are then given by

$$L = (W' * W_{HRIR_L}) + (X' * X_{HRIR_L}) + (Y' * Y_{HRIR_L}) + (Z' * Z_{HRIR_L}) \tag{7.10}$$

$$R = (W' * W_{HRIR_R}) + (X' * X_{HRIR_R}) + (Y' * Y_{HRIR_R}) + (Z' * Z_{HRIR_R}) \tag{7.11}$$

where W', X', Y', Z' are Ambisonic source channels preconvolved with the angle independent component. This technique, in combination with the factorization process allows for high order real-time Ambisonic renderings over virtual loudspeakers at low computational cost.

7.4.2 Higher Order Ambisonic Synthesis

However, we have thus far only considered the optimization of virtual loudspeakers for reproduction over headphones. Now we must consider the loudspeaker feeds themselves and their application to both virtual and real loudspeakers. It was shown in Chapter 4 that 3^{rd} Order Ambisonics represents a practical solution to VAE reproduction. However, we have only considered the measurement and interpolation of 1^{st} Order SRIRs. In order for us to change 1^{st} Order Ambisonic impulse responses to 3^{rd} order representations, we need to first acknowledge that physical up-conversion is an impossibility. However, a perceptual based approach will allow us to to synthesize the same spatial impression as would be experienced with a higher order Ambisonic soundfield recording. In order to achieve this we must know the following with regard to the spatial impulse responses:

1. The direction of arrival of the direct sound and early reflections
2. The diffuseness of the impulse response
3. The perceived spectral energy of the impulse response

For this parameterization, we adopt the directional analysis method of Pulkki and Merimaa, found in [Merimaa and Pulkki, 2005]. Here the B-format signals are analyzed in terms of sound intensity and energy in order to derive time-frequency based direction of arrival and diffuseness. The instantaneous energy density of the soundfield can be computed by

$$E(t) = \frac{1}{2}\rho_0 \left(\frac{p^2(t)}{Z_0^2} + \mathbf{u}^2(t) \right) \tag{7.12}$$

where $p(t)$ is the sound pressure, $\mathbf{u}(t)$ is the particle velocity vector, $Z_0 = \rho_0 c$ is the impedance of the medium and ρ_0 and c are the mean density and the speed of sound respectively. The instantaneous intensity vector is given as

$$\mathbf{I}(t) = p(t)\ \mathbf{u}(t) \tag{7.13}$$

Since we are using 1$^{\text{st}}$ Order Ambisonic SRIRs, the pressure can be approximated by

$$p(t) = w(t) \tag{7.14}$$

and the particle velocity by

$$\mathbf{u}(t) = \frac{1}{\sqrt{2}Z_0} \left(x(t)\mathbf{e}_x + y(t)\mathbf{e}_y + z(t)\mathbf{e}_z \right) \tag{7.15}$$

where $\mathbf{e}_x$, $\mathbf{e}_y$, and $\mathbf{e}_z$ represent cartesian unit vectors. The instantaneous intensity represents the direction of the energy transfer of the soundfield and the direction of arrival can be determined simply by the opposite direction of $\mathbf{I}$. For B-format Ambisonics, we can calculate the intensity for each coordinate axis, and in the frequency domain. For example, for the x-axis we have

$$I_x(\omega) = \frac{\sqrt{2}}{Z_0}\mathrm{Re}(W^*(\omega)X(\omega)) \tag{7.16}$$

Thus, the direction of arrival can be calculated by

$$\theta(\omega) = \tan^{-1}\left[\frac{-I_y(\omega)}{-I_x(\omega)} \right] \tag{7.17}$$

and for elevation by

$$\phi(\omega) = \tan^{-1}\left[\frac{-I_z(\omega)}{\sqrt{I_x^2(\omega) + I_y^2(\omega)}} \right] \tag{7.18}$$

Equation 7.18 yields two results separated by π radians and the correct result can be determined by the signs of $-I_x(\omega)$, $-I_y(\omega)$ and $-I_z(\omega)$. The direction $\mathbf{I}$ can change over time, so the net flow is determined by the time average of the instantaneous intensity. Since a portion of the energy will also oscillate locally, a diffuseness estimate can be made which is given by the ratio of the magnitude of the intensity vector to the overall energy density given as

$$\psi = 1 - \frac{||\langle \mathbf{I} \rangle||}{c\langle E \rangle} \tag{7.19}$$

where $\langle \cdot \rangle$ denotes time averaging and $|| \cdot ||$ denotes the norm of the vector. The diffuseness estimate will yield a value of zero for incident plane waves from a particular direction, but will give a value of 1 where there is no net transport of acoustic energy, such as in the cases of reverberation or standing waves. Time averaging is used since it is difficult to determine an instantaneous measure of diffuseness. For B-Format signals we can then write

$$\psi = 1 - \frac{\sqrt{2}\sqrt{\mathrm{Re}(W^*(\omega)X(\omega))^2 + \mathrm{Re}(W^*(\omega)Y(\omega))^2 + \mathrm{Re}(Z^*(\omega)X(\omega))^2}}{|W(\omega)|^2 + (|X(\omega)|^2 + |Y(\omega)|^2 + |Z(\omega)|^2)/2} \tag{7.20}$$

where * denotes the conjugate. The output of the analysis is then subject to smoothing based on the Equivalent Rectangular Bandwidth (ERB) scale, such that the resolution of the human auditory system is approximated. For this, each centre frequency of each bin in the STFT is determined, and an average of frequencies is taken within an ERB bandwidth of

$$ERB_N = 24.7 \times (4.37f + 1) \tag{7.21}$$

where f is the frequency in kHz. From this analysis the diffuse and non-diffuse portions of the soundfield can be extracted by multiplying the B-format signals by $\sqrt{1-\psi}$ and ψ respectively.

Since the frequency dependent direction of arrival of the non-diffuse portion of the soundfield can be determined, the soundfield can be re-encoded into point-like virtual sources using any spatialization method. Pulkki suggests the use of VBAP for the reproduction of the point like sources [Merimaa and Pulkki, 2005]. However, for the reasons outlined in Chapter 4, we choose to adopt Higher Order Ambisonics for this purpose. HOA reproduction can be achieved by re-encoding point like sources corresponding to the direction indicated in each temporal average and frequency band into a higher order spherical harmonic representation using the encoding and decoding equations of Section 3.4.5. The resultant HOA loudspeaker gains are then weighted in each frequency band i according to $\sqrt{1-\psi_i}$. The diffuse portion of the soundfield can be reproduced using a decorrelated version of $|W(\omega)|^2[1-\psi]$ at each loudspeaker, where decorrelation can be achieved using the method suggested in Section 6.4.

7.4.3 Examples of HOA Re-encoding using Directional Analysis

In assessing the effectiveness of the re-encoding method, consider the 1st order image source model shown in Figure 7.11. The model is of a small-sized room with dimensions 5m x 5m x 5m. The source shown is omnidirectional and results in the 1st order image source reflections shown as black circles outside the room. The receiver is a B-format receiver and its position is indicated by the Fig-8 polar pattern. The resultant B-Format impulse response is shown in Figure 7.12 (a), and its corresponding decode (infinite loudspeakers) is shown in terms of temporal-directional mapping of Figure 7.12 (a). The decode type shown here is the 'energy decode', whose anti-phase components are clearly visible (for example between 250 and 300 degrees for the direct sound).

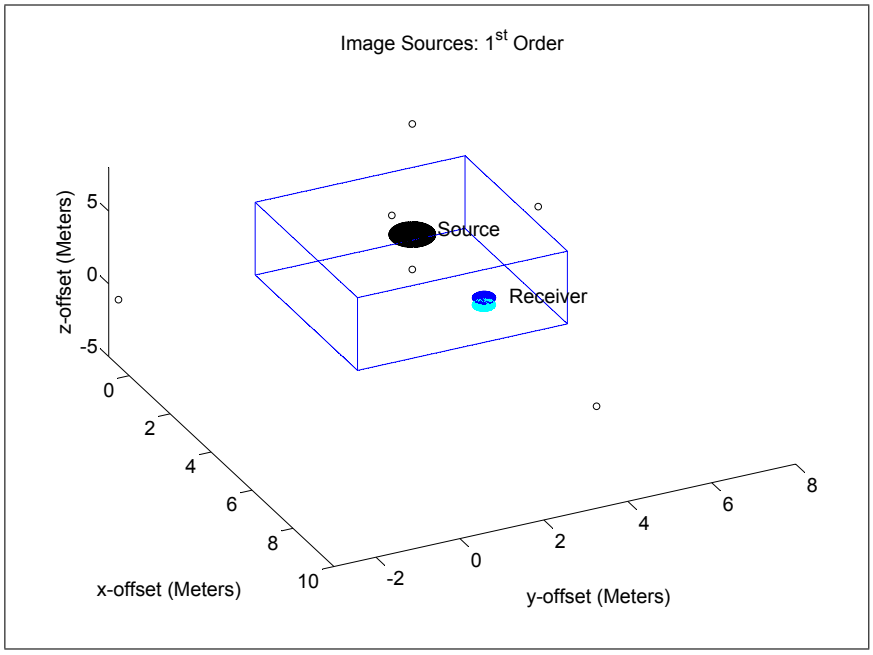


Figure 7.11: *Image source model utilizing 1^{st} Order reflections for B-Format simulation.*

Figure 7.12(c) then shows the re-encoded 3^{rd} Order Ambisonic signal. The time of arrival and direction of the direct sound and early reflections is exactly the same as in the 1^{st} Order case. However, as we would expect with a 3^{rd} order Ambisonic representation, the encoded signals now have a more fine directional resolution, seen by the reduced angular spread. This demonstrates the effectiveness of re-encoding the spatial impulse response to a higher order. Furthermore, re-encoding is possible to any order, provided the required number of loudspeakers is available. An example of a 15^{th} Order re-encode is shown in Figure 7.12(d). We see that both the direct sound and the reflections become more point like, and the overall directional accuracy is increased.

For real B-Format measurements, the performance of the re-encoding strategy works equally well. Figure 7.13 shows an example of the first 20ms of a spatial impulse response taken in Test Environment 1. Here the source was located 3m from a Soundfield ST350 microphone, and the SRIR captured using the logarithmic sine-swept tone technique [Farina, 2000]. In these plots, particular attention is drawn to the direct sound (coming from directly in front of the microphone) and a left wall reflection at approximately 14ms, which can be seen in the Y component of the SRIR. Just as was the case in the image source simulation, the directional resolution increases significantly with higher order Ambisonic representation. It should be noted, that since the A-format capsule on soundfield microphones is not truely coincident, it only displays adequate directionality up to 10kHz. This has been shown by several different authors in [Gerzon, 1975b], [Craven and Gerzon, 1977] and [Farrar, 1979]. Spatial aliasing is therefore an issue for high frequencies and as a result, the directional information above 10kHz cannot be

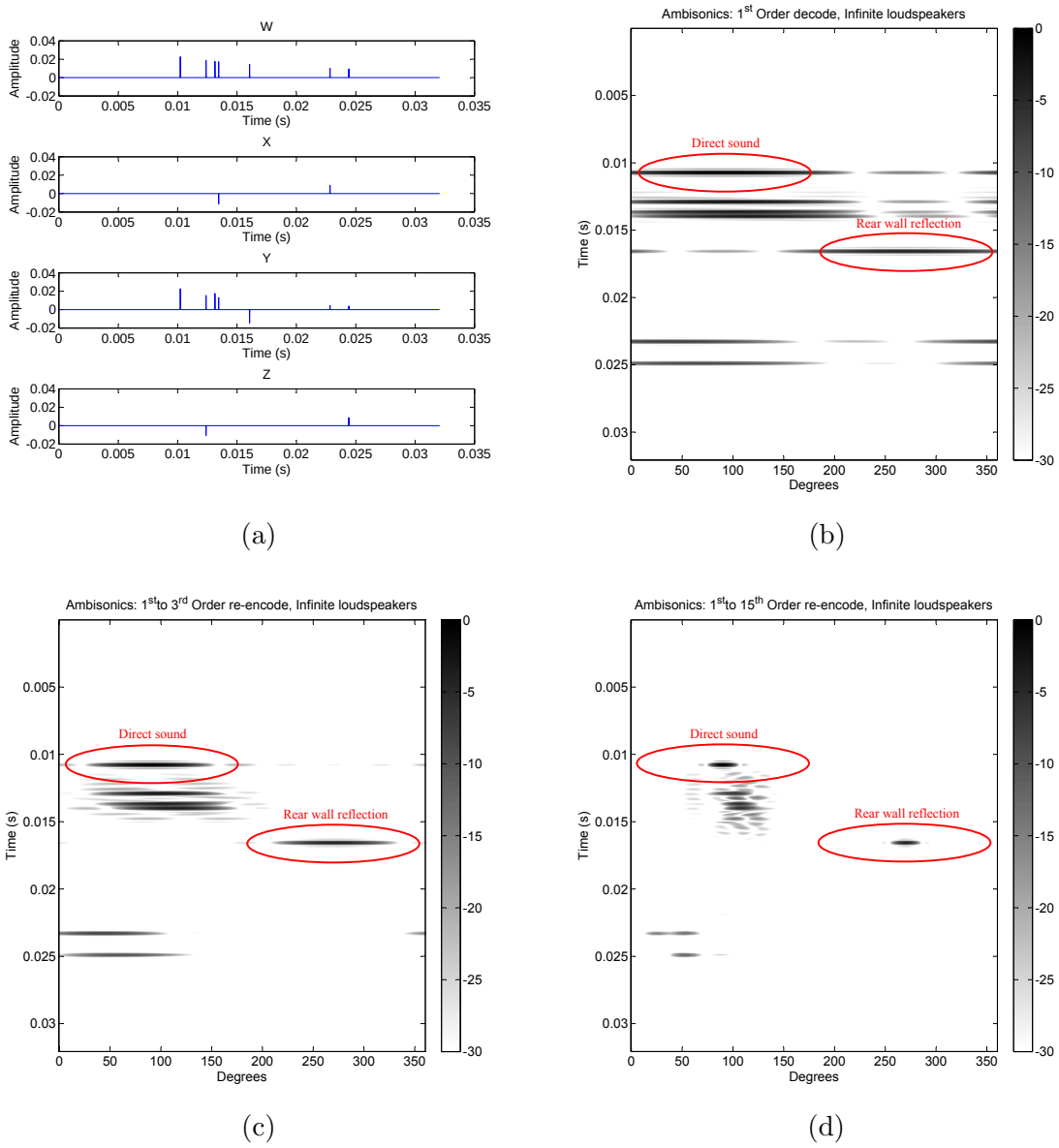


Figure 7.12: *Directional Ambisonic decode for B-Format image source simulation: (a) B-Format Impulse responses, (b) 1^{st} order decode, (c) 3^{rd} order re-encode, and (d) 15^{th} order re-encode.*

relied upon.

7.5 Conclusions

In this chapter, we have investigated different strategies for the optimisation of soundfield reproduction over both real and virtual loudspeakers. Factorization of HRIRs was implemented using

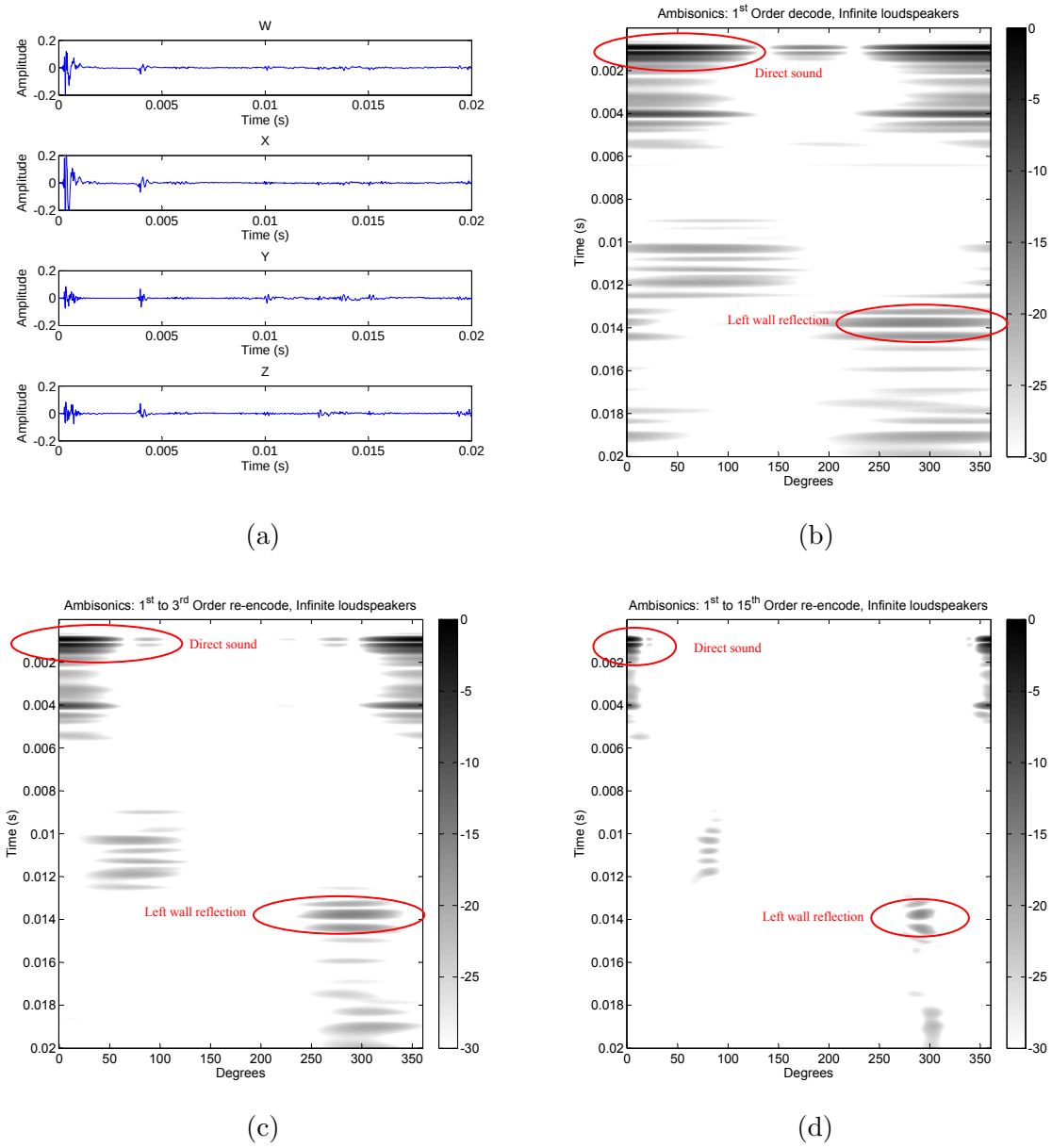


Figure 7.13: *Directional Ambisonic decode for B-Format measurement from Soundfield ST350: (a) B-Format Impulse responses, (b) 1^{st} order decode, (c) 3^{rd} order re-encode, and (d) 15^{th} order re-encode.*

an iterative least-squares based method. It was shown that good reconstruction can be achieved using directional independent components with short filter lengths. Examples of an 8-channel virtual loudspeaker array, constructed from HRIRs measurements from a KEMAR mannequin were presented. Good reconstruction was found to be possible for directional independent components of up to 140 samples in length, yielding 60 sample long run-time filters, under different initial conditions. The minimum phase representation of the HRIRs was shown to reduce the

filter length further, allowing for 20 sample long run-time filters, with low error when reconvolved with the angle independent component. The issue of headphone equalization was also addressed. It was shown that significant low frequency deviations in the magnitude response occur due to headphone placement, as well as high frequency notching, and as a consequence, care must be taken in the inverse filter design. Finally, attention was drawn to the loudspeaker feeds themselves, and a method of synthesizing higher order Ambisonics from 1^{st} Order soundfields was presented using the methodology of Merimaa and Pulkki [Merimaa and Pulkki, 2005]. The effectiveness of the method was demonstrated through image source simulation, as well as real acoustic measurements.

8 Conclusions

8.1 Summary and Conclusions

The main aim of the work presented in this thesis was to introduce novel signal processing strategies which simplify both the capture and reproduction of real acoustic musical events for auditory presentations in interactive VAEs. Particular attention was made to data-based auralization, where real acoustic spatial impulse responses can be used to form a convincing auditory scene. The main hypothesis of this work was:

> **A plausible formation of natural auditory scenes can be constructed efficiently using data-based auralization in a virtual auditory environment, and that said scenes can be rendered to either individual or multiple listeners using the same recording and reproduction paradigm.**

The research hypothesis was supported by:

1. An analysis of the psychoacoustics of source localization in both anechoic and reverberant environments.
2. An evaluation of the capabilities of spatialization techniques to cater for the localization demands of the individual listener and a distributed audience.
3. A perceptual assessment of the differences between real and virtual acoustic recordings.
4. The development of novel signal processing strategies to simplify the practical capture and reproduction of acoustic events for VAE presentations.

Throughout the investigations in this thesis, two strands of acoustic signal processing were developed based on the 3-D spatial audio capture and reproduction respectively. The strategy of direct field recording, removal of convolution chain coloration, superposition of measured source directivity and final convolution of spatial impulse responses was termed *Virtual Acoustic Recording*. The reproduction of higher order Ambisonic soundfields based on first order measurements and the corresponding optimization of virtual loudspeaker reproduction via headphones was termed *Virtual Acoustic Reproduction*.

The thesis began by investigating perceived localization accuracy in both reflective and non-reflective environments. It was found in the context of free-field localization, that the binaural localization cues of interaural time and level difference are the most significant factors in the correct perception of horizontal localization. Monaural spectral cues are also crucial in overcoming front-back confusion, as is the importance of head-movement. The perception of distance is difficult under free-field conditions however. In the case of reverberant environments, the direct to reverberant ratio has a significant effect on localization accuracy affecting the perceived spaciousness and perception of distance. Interaural cross correlation was shown to be an effective objective measure of source width when using binaural microphones. However, the localization accuracy using normalized cross correlation in reverberant rooms was shown to deviate from the human perception of localization. Use of the phase-transform weighting function was introduced, which allowed for a more perceptually accurate representation of interaural cross correlation with binaural microphones. The effectiveness of the method was then confirmed through subjective listening tests. This method was later used in the thesis for assessing the localization accuracy of VBAP and Ambisonic systems in non-ideal listening scenarios, as well as fluctuations in ITD between real and virtual acoustic recordings.

Once the relevant psychoacoustic foundations were made, a detailed review of current methods of soundfield reproduction was conducted in Chapter 3. The main aim of this chapter was to investigate potential methods of spatial audio reproduction applicable to VAE presentations. Two classes of spatial reproduction were outlined: binaural reproduction for the individual listener, and loudspeaker reproduction for the distributed audience. It was shown that the 'virtual loudspeaker' approach allows us to use the same spatialization system for both headphone and loudspeaker reproduction. Based on this premise, an analysis of spatialization schemes which satisfy the localization requirements of the individual listener and the distributed audience was presented. This analysis verified, through extensive simulations, the work of previous authors pertaining to VBAP, Ambisonics and Wave Field Synthesis. Of the spatialization schemes assessed, it was shown that Ambisonics offers the greatest potential as a practical reproduction method, since the information channel count can be much lower than the reproduction channel count, and that decoding over different loudspeaker layouts is possible. Furthermore, the higher the order of Ambisonic system, the larger the effective sweet-spot over the listening area. At high orders, Ambisonics then reaches similar limitations to WFS systems. Both VBAP and WFS on the other hand, were shown to offer good localization accuracy over the listening area. However,

WFS is not a practical solution, since it does not accommodate the use of commercially available consumer-based technology. Nevertheless, it is a useful tool in assessing the fundamental limitations of virtual source formation, equivalent to the performance of high order Ambisonics.

Based on this analysis a formal assessment of VBAP and Ambisonic systems was presented in Chapter 4. The objective was to assess each spatialization method in terms of their localization accuracy for virtual sources for a distributed audience under non-ideal listening conditions. An objective analysis was presented in the form of the vector localization theory of Gerzon, which was expanded to accommodate off-centre listening. The localization results were also corroborated through binaural simulations of Test Environment 1. Subjective listening tests were also performed to investigate the perceptual difference between the spatialization schemes. It was shown that

- VBAP provides superior localization accuracy to 1^{st} Order Ambisonics for stationary sources.
- 1^{st} Order Ambisonics is preferred over VBAP for moving sources.

A second series of experiments was implemented to assess whether the localization performance of Ambisonic systems can be improved through optimized decoders and higher order systems. It was found that

- The localization accuracy of Ambisonic systems increases under real listening conditions for all decoder types
- Optimal localization accuracy for an individual listener over an 8-channel array can be achieved using a psychoacoustically optimized decoder, with maximized r_v and r_e at low and high frequencies respectively.
- Optimal localization accuracy for a distributed audience is achieved by maximizing r_e over the entire frequency range. Decoders designed to optimize r_v result in large low frequency localization errors at off-centre positions.

It was also shown that the performance of the Ambisonic 'energy' decoder gives similar localization accuracy to VBAP in the context of a distributed audience. In both cases, localization accuracy for an 8-channel array is within the loudspeaker aperture of $\pm 22.5^{\circ}$.

A strategy for recording for VAEs using Ambisonics was then presented in Chapter 5. The strategy was based around the direct-field capture of acoustic performances and their corresponding convolution with B-Format spatial impulse responses. Particular attention was paid to the perceptual effects of coloration in the convolution chain due to loudspeaker and microphone responses. An assessment of the differences between actual acoustic recordings and virtual acoustic recordings was then presented. It was shown that

- The coloration effects due to the convolution chain are an important factor in identifying virtual recordings from actual ones.

- Source width is an important perceptual characteristic in virtual source formation from acoustic measurements.

A method of approximating source directivity utilizing measured B-Format impulse responses from a source simulation array and was then proposed. The method utilized a least squares optimization process as well as critical band filtering to achieve the synthesized SRIR. The use of virtual acoustic recording in this chapter was limited to single perspective viewing. It was recognized that for multi-perspective listening, such as in walk-through VAEs, measurements of multiple SRIRs are necessary. This was addressed in Chapter 6. It was shown that although current practice in SRIR computation uses grids of impulse response simulations, real-world measurements cannot be so easily achieved. An investigation into the reduction of SRIR measurements in real rooms was investigated based on a perceptual analysis. It was found that

- Since perception of depth in real rooms is continuously underestimated sparse SRIR sampling based on depth perception is possible.
- Since perceived source width in real rooms is dependent on the direct to reverberant ratio sparse SRIR sampling based on horizontal localization blur is possible.

Whilst such a psychoacoustical analysis allows for the reduction of the number of SRIR measurements, a high density of SRIRs is still required close to the source position. As a consequence, interpolation over sparse impulse responses is advantageous. It was shown that interpolation based on Dynamic Time Warping is an effective method for the formation of new SRIRs in between measured SRIRs. An analysis of the effectiveness of interpolation on 1^{st} Order B-format impulse responses was presented. It was shown that effective interpolation can be achieved with reasonable localization accuracy up to approximately 1m in reverberant rooms. SRIR measurement topologies were then presented for the example of Test Environment 1 based on perceived localization within the environment and the performance of the DTW algorithm.

The optimization of the virtual loudspeaker approach was then addressed in Chapter 7. It was shown that real-time HRIR filtering can be simplified by separating HRIR datasets into angle dependent and independent components. A novel factorization algorithm was then introduced, which allowed for the formation of 20-tap run-time HRIR filters based on minimum phase assumptions. The method was assessed using HRIRs from the CIPIC database, and low reconstruction errors were achieved. Finally, a method of synthesizing higher order Ambisonic SRIRs from 1^{st} Order Ambisonic measurements was presented based on a directional analysis of the measured soundfield. It was shown through both simulated and real measurements that SRIRs of any desired Ambisonic order can be constructed based on this method.

8.2 Further Work

The work presented in this thesis can be developed within three broad categories. These are:

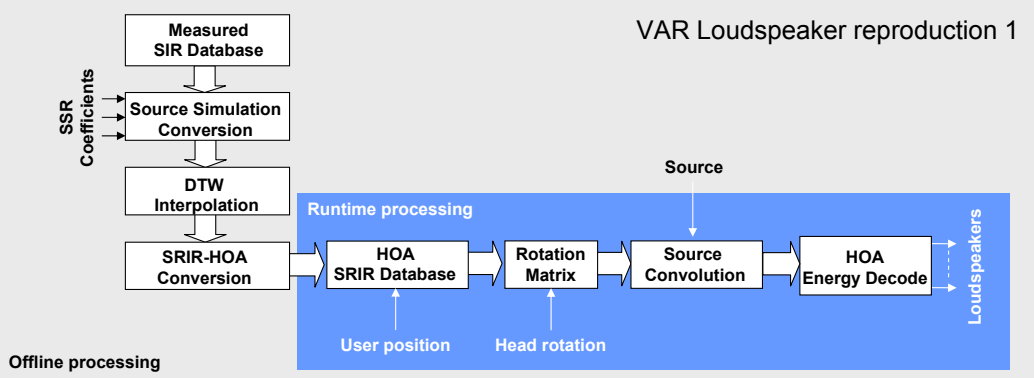


Figure 8.1: *Loudspeaker reproduction 1.*

- Integration of signal processing strategies to VAE production frameworks
- Development of virtual acoustic recording strategies for computational based auralization
- Development of optimization procedures for binaural synthesis

8.2.1 Integration of Signal Processing Strategies to VAE Production Frameworks

The signal processing strategies introduced in this thesis were designed for musical performance capture and reproduction in a VAE. Such an example of a VAE loudspeaker reproduction workflow for music presentation is shown in Figure 8.1. Here a measured SRIR database taken from a source simulation radiator is transformed to approximate a given source directivity. The density of SRIRs can then be increased using DTW interpolation and the resultant impulse responses converted to a database of higher order Ambisonic impulse responses. Based on the listener position in the virtual environment, the appropriate SRIR can be chosen from the database, and subject to a real-time rotation matrix controlled via head-tracking. The source audio can then be convolved with the SRIR and the resultant output decoded using an 'energy decode' to accommodate off-centre listening. Alternatively, the source convolution can be performed off-line and stored as shown in Figure 8.2. At run-time, audio sequences can then be cross-faded as the listener moves through the VAE. Such an application is useful where run-time computational complexity is an issue, but data storage is not. A similar approach can then be adopted for virtual loudspeaker reproduction over headphones, with the exception that the source can be pre-convolved with the angle independent component. Directional dependent filtering can then be implemented after a 'shelf-filter' Ambisonic decode. This is shown in Figure 8.3. Alternatively, the strategy oulined in Section 7.4.1 can be employed, where a set of shelf filter weighted Ambisonic HRTFs can be convolved with the Ambisonic source channels directly.

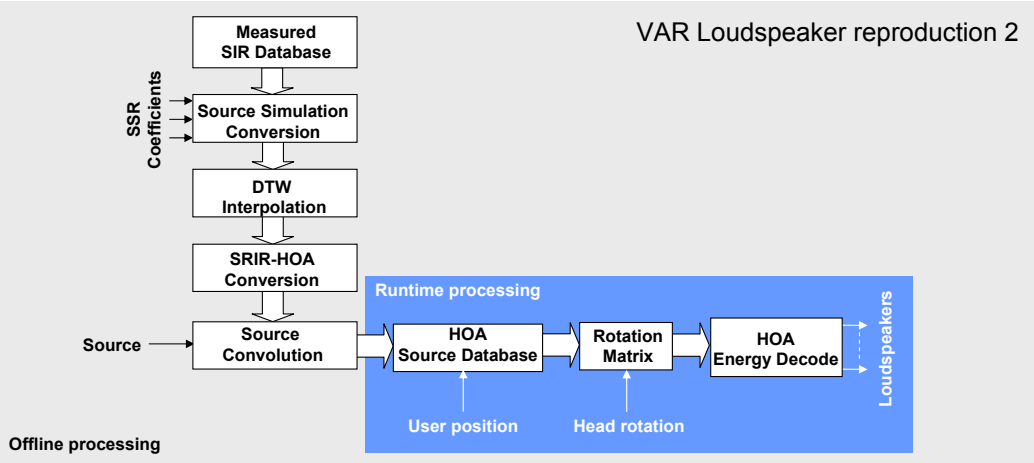


Figure 8.2: *Loudspeaker reproduction 2.*

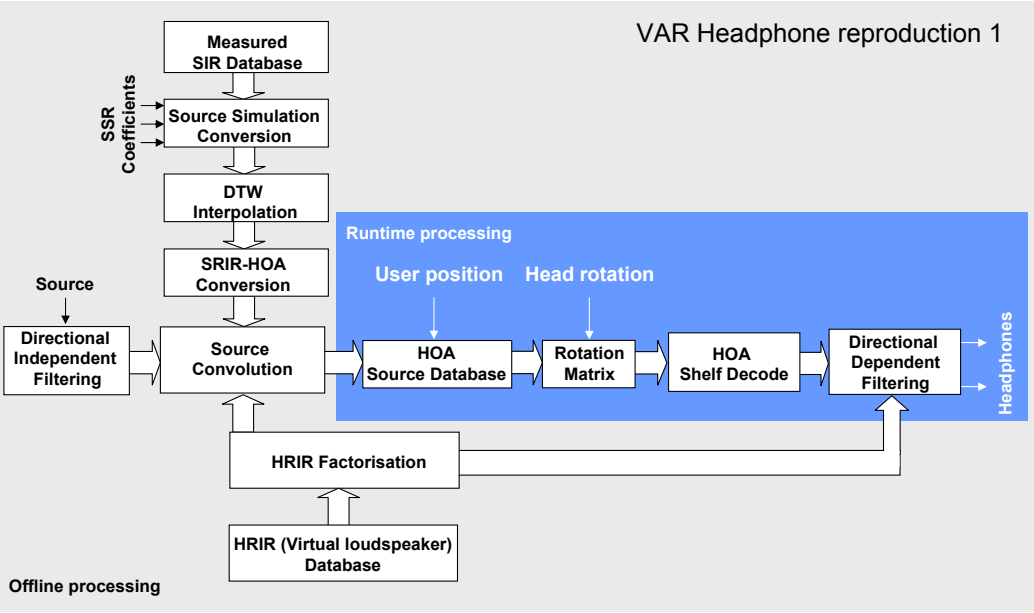


Figure 8.3: *Virtual loudspeaker headphone reproduction.*

The components of each of the reproduction strategies shown can also be used individually. For example, gaming based technology can efficiently utilize the real-time virtual loudspeaker approach by pre-convolving all game audio with the angle independent component of the virtual loudspeaker dataset. During game-time, the computational complexity of the spatial audio reproduction is therefore reduced. Also, in natural single perspective music production, the virtual acoustic recording approach can be utilized to give realistic spatial impression of reverberant spaces and the associated source directivity for studio-based recordings.

8.2.2 Development of Virtual Acoustic Recording and Reproduction Strategies for Computational Based Auralization

For all computational-based auralization processes there is a trade-off between the accuracy of the SRIR dataset and the computation time required to render the SRIRs. For example, it has been shown in [Southern et al., 2009] that the computation time for SRIRs, rendered with a 3D Digital Waveguide Mesh [Murphy et al., 2007], [Smith, 1992] on a standard desktop PC, is approximately 14 hours and 18 minutes. The advantage of this method over other auralization techniques is that the computation time does not change whether it is one receiver or 300 receivers. Instead, it is defined by the fine structure of the mesh, where each mesh point can potentially yield an RIR. If a finer resolution of RIRs is required then DTW can be used as the interpolation process, rather than recomputing the grid and increasing computation time. Furthermore, the HRIR factorization process can be applied such that more efficient run-time auralization can be achieved. This is also highly relevant to systems that utilize real-time image source calculation, such as in [Borß and Martin, 2009]. The development of perceptually-based topologies also warrants further investigation, in particular the possibility to derive topologies based on binaural ASW measurements/computation.

8.2.3 Further Optimization Procedures for Binaural Synthesis

Whilst this thesis proposes strategies for the simplification of binaural synthesis, further work is required in the perceptual attributes associated with this optimization. In particular, the effect of reducing the directional dependent component needs to be investigated through both objective binaural measurements, as well as subjective testing for minimum audible differences between the full and factorized HRIRs. The use of DTW in HRIR interpolation also warrants further investigation, which would allow for the formation of larger virtual loudspeaker arrays from a small number of HRIR measurements. Equalization of phase distortion, introduced by the reproduction headphones themselves, also requires attention, as this may lead to interaural phase-difference errors. Furthermore, the role of near-field loudspeaker directivity in the measurement of HRIRs for virtual loudspeakers, as well as its effect on near-field compensated Ambisonic rendering must be studied. Finally, a psychoacoustical basis for the initial conditions in the factorization process should also be investigated, which could potentially improve the performance of the algorithm.

8.3 Final Remarks

Plausible multi-perspective 3-D auralization, whether data-based or computational based, can only be achieved through complex rendering processes that consider the source directivity, the spatial resolution of SRIR datasets, the auditory perception in real rooms and the optimal spatialization method for reproduction. This thesis has proposed strategies that can signifi-

cantly simplify such auralization processes, particularly for measurement based audio rendering of natural music performances. It is therefore hoped by the author that the contribution of these strategies to the field of auralization will lead to more realistic and immersive auditory environments, and that their limitations can be overcome by future developments in the field.

Bibliography

[Ahnert, 1992] Ahnert, W. (1992). Ears auralization software. In *93rd Convention of the Audio Engineering Society*, San Francisco, California, USA. Paper 3388.

[Ahrens and Spors, 2007] Ahrens, J. and Spors, S. (2007). Implementation of directional sources in wave field synthesis. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 66 – 69, New Paltz, New York, USA.

[Algazi et al., 2001] Algazi, V. R., Duda, R. O., Thompson, D. M., and Avendano, C. (2001). The CIPIC HRTF database. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 99–102, New Paltz, New York, USA.

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

[ANSI, 2006] ANSI (2006). American national standard specification for a manikin for simulated in-situ airborne acoustic measurements. ANSI S3.36-1985 (R2006).

[Atal, 1966] Atal, B. (1966). Apparent sound source translator. US Patent 3236949.

[Bamford and Vanderkooy, 1995] Bamford, J. S. and Vanderkooy, J. (1995). Ambisonic sound for us. In *99th Convention of the Audio Engineering Society*, New York, New York, USA. Paper 4138.

[Barry and Kearney, 2009] Barry, D. and Kearney, G. (2009). Localisation quality assessment in source separation based upmixing algorithms. In *35th International Conference of the Audio Engineering Society*, London, UK. Paper 33.

[Barry et al., 2005] Barry, D., Lawlor, R., and Coyle, E. (2005). Comparison of signal reconstruction methods for the azimuth discrimination and resynthesis algorithm. In *118th Convention of the Audio Engineering Society*, Barcelona, Spain. Paper 6333.

[Bass et al., 1972] Bass, H. E., Bauer, H.-J., and Evans, L. B. (1972). Atmospheric absorption of sound: analytical expressions. *The Journal of the Acoustical Society of America*, 52(3B):821–825.

[Bates et al., 2007] Bates, E., Kearney, G., Furlong, D., and Boland, F. (2007). Localization accuracy of advanced spatialization techniques in small-sized concert halls. In *153rd Meeting of the Acoustical Society of America*.

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

[Benjamin et al., 2006] Benjamin, E., Lee, R., and Heller, A. J. (2006). Localization in horizontal-only Ambisonic systems. In *121st Convention of the Audio Engineering Society*, San Francisco, California, USA. Paper 6967.

[Beranek and Martin, 1996] Beranek, L. and Martin, D. W. (1996). Concert & opera halls: How they sound. *The Journal of the Acoustical Society of America*, 99(5):2637–2637.

[Berkhout, 1988] Berkhout, A. J. (1988). A Holographic Approach to Acoustic Control. *Journal of the Audio Engineering Society*, 36(12):977–995.

[Berkhout, 1998] Berkhout, A. J. (1998). A Wavefield approach to multichannel sound. In *104th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands. Paper 4749.

[Berkhout et al., 1993] Berkhout, A. J., de Vries, D., and Vogel, P. (1993). Acoustic control by wave field synthesis. *The Journal of the Acoustical Society of America*, 93(5):2764–2778.

[Blauert, 2003] Blauert, J. (2003). *Spatial Hearing*. MIT Press Cambridge MA.

[Blauert, 2008] Blauert, J. (2008). *Communication Acoustics*. Springer.

[Blender, 2010] Blender (2010). Blender 2.49b. `http://www.blender.org/`, accessed 29-03-10.

[Blumlein, 1931] Blumlein, A. (1931). Improvements in and relating to sound transmission, sound recording, and sound reproducing systems. British Patent Specification 394325.

[Boone and Verheijen, 1993] Boone, M. M. and Verheijen, E. N. G. (1993). Multi-channel sound reproduction based on Wave Field Synthesis. In *95th Convention of the Audio Engineering Society*, New York, New York, USA. Paper 3719.

[Borish and Angell, 1983] Borish, J. and Angell, J. B. (1983). An efficient algorithm for measuring the impulse response using pseudorandom noise. *Journal of the Audio Engineering Society*, 31(7/8):478–488.

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

[Borß and Martin, 2009] Borß C. and Martin, R. (2009). An improved parametric model for perception-based design of virtual acoustics. In *35th International Conference of the Audio Engineering Society*, London, UK. Paper 3.

[Boueri and Kyirakakis, 2004] Boueri, M. and Kyirakakis, C. (2004). Audio signal decorrelation based on a critical band approach. In *117th Convention of the Audio Engineering Society*, San Francisco, California, USA. Paper 6291.

[Bregman, 1994] Bregman, A. S. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press.

[Bronkhorst and Houtgast, 1999] Bronkhorst, A. W. and Houtgast, T. (1999). Auditory distance perception in rooms. *Nature*, 397:517–520.

[Brungart and Rabinowitz, 1999] Brungart, D. S. and Rabinowitz, W. M. (1999). Auditory localization of nearby sources. head-related transfer functions. *The Journal of the Acoustical Society of America*, 106(3):1465–1479.

[Chan et al., 2009] Chan, J. S., Maguinness, C., Lisiecka, D., Ennis, C., Larkin, M., O'Sullivan, C., and Newell, F. (2009). Comparing audiovisual distance perception in various real and virtual environments. In *32nd European Conference on Visual Perception*.

[Chen, 2009] Chen, T. (2009). Working in b+ format. `http://www.ambisonic.net/chenbplus.html`, accessed 24-10-2009.

[Chin et al., 1998] Chin, P., Corless, R., and Corliss, G. (1998). Optimization strategies for the approximate GCD problem. In *Proceedings of the 1998 international symposium on Symbolic and algebraic computation*, pages 228–235. ACM Press New York, NY, USA.

[Chowning, 1971] Chowning, J. M. (1971). The simulation of moving sound sources. *Journal of the Audio Engineering Society*, 19(1):2–6.

[Clark et al., 1958] Clark, H., Dutton, G., and Vanderlyn, P. (1958). The stereosonic recording and reproduction system: A two-channel system for domestic tape records. *Journal of the Audio Engineering Society*, 6:102–117.

[Cobos and Lopez, 2009] Cobos, M. and Lopez, J. (2009). Resynthesis of sound scenes on wave-field synthesis from stereo mixtures using sound source separation algorithms. *Journal of the Audio Engineering Society.*, 57(3):91–110.

[Colavita, 1974] Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics*, 16:1409–412.

[Corless et al., 1999] Corless, R., Giesbrecht, M., and Jeffrey, D. (1999). Approximate polynomial decomposition. In *Proceedings of the 1999 international symposium on Symbolic and algebraic computation*, pages 213–219. ACM New York, NY, USA.

[Corteel et al., 2007] Corteel, E., Nguyen, K., Warusfel, O., Caulkins, T., and Pellegrini, R. (2007). Objective and subjective comparison of electrodynamic and map loudspeakers for Wave Field Synthesis. In *30th International Conference of the Audio Engineering Society*. Paper 15.

[Craven, 2003] Craven, P. (2003). Continuous surround panning for 5-speaker reproduction. In *24th International Conference of the Audio Engineering Society*, Banff, Canada. Paper 9.

[Craven and Gerzon, 1977] Craven, P. G. and Gerzon, M. A. (1977). Coincident microphone simulation covering three dimensional space and yielding various directional outputs. U.S. Patent Number 4,042,779.

[Cruz-Neira, 1995] Cruz-Neira, C. (1995). *Virtual reality based on multiple projection screens: the cave and its applications to computational science and engineering*. PhD thesis, University of Illinois at Chicago, Chicago, IL, USA.

[Dalenbäck and Strömberg, 2006] Dalenbäck, B.-I. and Strömberg, M. (2006). Real time walkthrough auralization - the first year. In *Proceedings - Institute of Acoustics, vol. 28, Pt.2*.

[Damaske and Wagener, 1969] Damaske, P. and Wagener, B. (1969). Investigations of directional hearing using a dummy head. *Acustica*, 21:30–35.

[Daniel, 2003a] Daniel, J. (2003a). Further investigations of High Order Ambisonics and Wavefield Synthesis for holophonic sound imaging. In *114th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands. Paper 5788.

[Daniel, 2003b] Daniel, J. (2003b). Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new Ambisonic format. In *23rd International Conference of the Audio Engineering Society*. Paper 16.

[Daniel et al., 1998] Daniel, J., Rault, J., and Polack, J. (1998). Ambisonics encoding of other audio formats for multiple listening conditions. In *105th Convention of the Audio Engineering Society*, San Francisco, California, USA. Paper 4795.

[de Vries and Baan, 1999] de Vries, D. and Baan, J. (1999). Auralization of sound fields by Wave Field synthesis. In *106th Convention of the Audio Engineering Society*, Munich, Germany. Paper 4927.

[de Vries and Berkhout, 1996] de Vries, D. and Berkhout, A. (1996). Array technology for accurate sound measurements in enclosed spaces. In *Forum Acusticum*, Antwerpen. (published in Acta Acustica vol.82 suppl.1,S 221).

[de Vries and Hulsebos, 2004] de Vries, D. and Hulsebos, E. M. (2004). Auralization of room acoustics by Wave Field Synthesis based on array measurements of impulse responses. In *Proc. of the 12th European Signal Processing Conference (EUSIPCO)*, pages 1377 – 1380.

[de Vries et al., 1994] de Vries, D., Start, E., and Valstar, V. (1994). The wave-field synthesis concept applied to sound reinforcement restriction and solutions. In *96th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands. Paper 3812.

[Denon, 1995] Denon (1995). Anechoic orchestral music recording. Audio CD, Denon Records, ASIN: B0000034M9.

[Dolby, 1999] Dolby (1999). Surround sound, past present and future. `http://www.dolby.com/uploadedFiles/zz-_Shared_Assets/English_PDFs/Professional/2_Surround_Past.Present.pdf`, accessed 29-03-2010.

[DTS, 2007] DTS (2007). DTS-HD high definition audio. `http://www.dts.com/dts-hd/`, accessed 15-01-2009.

[Duda and Martens, 1998] Duda, R. O. and Martens, W. L. (1998). Range dependence of the response of a spherical head model. *The Journal of the Acoustical Society of America*, 104(5):3048–3058.

[Everest, 1994] Everest, F. (1994). *The Master Handbook of Acoustics: 3rd Edition*. TAB Books, McGraw-Hill.

[Farina, 2000] Farina, A. (2000). Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *108th Convention of the Audio Engineering Society*, Paris, France. Paper 5093.

[Farina and Tronchin, 2004] Farina, A. and Tronchin, L. (2004). Advanced techniques for measuring and reproducing spatial sound properties of auditoria. In *International Symposium on Room Acoustics: Design and Science*.

[Farina and Ugolotti, 1998] Farina, A. and Ugolotti, E. (1998). Spatial equalization of sound systems in cars. In *15th International Conference of the Audio Engineering Society*, Copenhagen, Denmark. Paper 15-009.

[Farrar, 1979] Farrar, K. (1979). Soundfield microphone: Design and development of microphone and control unit. Wireless World 85, 4850.

[Fauster, 2007] Fauster, L. (2007). Stereoscopic techniques in computer graphics. Thesis, TU Wien, `http://www.cg.tuwien.ac.at/research/publications/2006/Fauster-06-st/Fauster-06-st-.pdf`, acessed 29/03/2010.

[Fellgett, 1975] Fellgett, P. (1975). Ambisonics. Part One. General system description. *Studio Sound*, 17(8):20–22, 40.

[Festen and Plomp, 1981] Festen, J. M. and Plomp, R. (1981). Relations between auditory functions in normal hearing. *The Journal of the Acoustical Society of America*, 70:356–369.

[Festen and Plomp, 1983] Festen, J. M. and Plomp, R. (1983). Relations between auditory functions in impaired hearing. *The Journal of the Acoustical Society of America*, 73(2):652–662.

[Fisher et al., 1986] Fisher, W. M., Doddington, G. R., and Goudie-Marshall, K. M. (1986). The DARPA speech recognition research database: Specifications and status. In *Proceedings of DARPA Workshop on Speech Recognition*, pages 93–99. SAIC-86/1546.

[Fletcher, 1940] Fletcher, H. (1940). Auditory patterns. *Reviews of modern Physics*, 12:47–65.

[Fletcher and Munson, 1933] Fletcher, H. and Munson, W. (1933). Loudness, its definition, measurement, and calculations. *The Journal of the Acoustical Society of America*, 5:82–108.

[Frank et al., 2008] Frank, M., Zotter, F., and Sontacchi, A. (2008). Localization experiments using different 2D Ambisonics decoders. In *25th Tonmeistertagung VDT International Convention*.

[Freeland et al., 2007] Freeland, F. P., Biscainho, L. P., and Diniz, P. R. (2007). Interpositional transfer function for 3D-sound generation. *Journal of the Audio Engineering Society*, 52(9):915–930.

[Gade, 1982] Gade, A. (1982). Subjective room acoustics experiments with musicians. Ph.D. Thesis, Report No. 32, Technical University of Denmark.

[Gade, 1989] Gade, A. (1989). Investigations of musicians room acoustics conditions in concert halls. Part II: Field measurements and synthesis of results. *Acustica*, 69:249–262.

[Gardner, 1969] Gardner, M. B. (1969). Distance estimation of 0[degree] or apparent 0[degree]-oriented speech signals in anechoic space. *The Journal of the Acoustical Society of America*, 45(1):47–53.

[Gerzon, 1975a] Gerzon, M. (1975a). Ambisonics. Part Two: Studio techniques. *Studio Sound*, 17(8):24, 26, 28–30.

[Gerzon, 1975b] Gerzon, M. (1975b). The design of precisely coincident microphone arrays for stereo and surround sound. In *50th Convention of the Audio Engineering Society*, London, UK. Paper L-20.

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

[Gerzon, 1974] Gerzon, M. A. (1974). What's wrong with Quadraphonics? *Studio Sound*, 16(5):50,51,56.

[Gerzon, 1976] Gerzon, M. A. (1976). Blumlein stereo microphone technique. *Journal of the Audio Engineering Society*, 24(1):36–38.

[Gerzon, 1977a] Gerzon, M. A. (1977a). Criteria for evaluating surround-sound systems. *Journal of the Audio Engineering Society*, 25(6):400–408.

[Gerzon, 1977b] Gerzon, M. A. (1977b). Principles of Quadraphonic recording I. *Studio Sound and Tape Recording*, pages 338–339.

[Gerzon, 1977c] Gerzon, M. A. (1977c). Principles of Quadraphonic recording II. *Studio Sound and Tape Recording*, pages 380–384.

[Gerzon, 1980] Gerzon, M. A. (1980). Practical periphony: The reproduction of full sphere sound. In *65th Convention of the Audio Engineering Society*, London, UK. Paper 1571.

[Gerzon and Barton, 1992] Gerzon, M. A. and Barton, G. J. (1992). Ambisonic decoders for HDTV. In *92nd Convention of the Audio Engineering Society*, Vienna, Austria. Paper 3345.

[Giron, 1996] Giron, F. (1996). Investigations about the directivity of sound sources. PhD Thesis, Ruhr-Universitt, Bochum, Shaker Verlag.

[Glasgal, 2001] Glasgal, R. (2001). The ambiophone - derivation of a recording methodology optimized for ambiophonic reproduction. In *19th International Conference of the Audio Engineering Society*, Bavaria, Germany. Paper 1874.

[Gorzel, 2009] Gorzel, M. (2009). Investigations into OPSI with height. M.Phil Masters Thesis, Music and Media Technology, Trinity College Dublin.

[Greisenger, 1989] Greisenger, D. (1989). Practical processors and programs for digital reverberation. In *7th International Conference of the Audio Engineering Society*, Toronto, Ontario. Paper 7-027.

[Guastavino et al., 2007] Guastavino, C., Larcher, V., Catusseau, G., and Boussard, P. (2007). Spatial audio quality evaluation: Comparing Transaural, Ambisonics and Stereo. In *Proceedings of the 13th International Conference on Auditory Display.*

[Gustafsson et al., 2003] Gustafsson, T., Bhaskar, D. R., and Trivedi, M. (2003). Sound localization in reverberant environments: Modeling and statistical analysis. *Speech and Audio Processing, IEEE Transactions on*, 11(6):791–803.

[Haas, 1972] Haas, H. (1972). The influence of a single echo on the audibility of speech. *The Journal of the Acoustical Society of America*, 20(2):146–159.

[Hamasaki, 2000] Hamasaki, K. (2000). Multichannel sound in television: Technical and aesthetic approach. In *3rd International Multichannel Sound Forum*, pages 26–41.

[Hamasaki and Hiyama, 2003] Hamasaki, K. and Hiyama, K. (2003). Reproducing spatial impression with multichannel audio. In *24th International Conference of the Audio Engineering Society*, Banff, Canada. Paper 19.

[Hamasaki et al., 2004] Hamasaki, K., Komiyama, S., Hiyama, K., and Okubo, H. (2004). 5.1 and 22.2 Multichannel sound productions using an integrated surround sound panning system. In *117th Convention of the Audio Engineering Society*, San Francisco, California, USA. Paper 6226.

[Hammershøi and Møller, 2008] Hammershøi, D. and Møller, H. (2008). Binaural technique–basic methods for recording, synthesis, and reproduction. *Chapter 9 in 'Communications Acoustics' by J.Blauert.*

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

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

[Hartmann, 1983] Hartmann, W. M. (1983). Localization of sound in rooms. *The Journal of the Acoustical Society of America*, 74:1380–1391.

[Haustein and Schirmer, 1970] Haustein, B. G. and Schirmer, W. (1970). A measuring apparatus for investigation of the faculty of directional localisation. *Hochferquenztech. u. Elektroakustik*, 79:96–101.

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

[Hiekkanen et al., 2009] Hiekkanen, T., Makivirta, A., and Karjalainen, M. (2009). Virtualized listening tests for loudspeakers. *Journal of the Audio Engineering Society*, 57(4):237–251.

[Hiipakka et al., 2001] Hiipakka, J., Ilmonen, T., Lokki, T., Grohn, M., and Savioja, L. (2001). Implementation issues of 3D audio in a virtual room. In *Stereoscopic Displays and Virtual Reality Systems VIII*, pages 486–495.

[Hoeg et al., 1983] Hoeg, W., Steffen, F., and Steinke, G. (1983). Delta Stereophony: A Sound System with True Direction and Distance Perception for Large Multipurpose Halls. *Journal of the Audio Engineering Society*, 31(7/8):500–511.

[Hulsebos et al., 2001] Hulsebos, E., de Vries, D., and Baan, J. (2001). Improved microphone array configurations for auralization of sound fields by Wave Field Synthesis. In *110th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands. Paper 5337.

[Huopaniemi et al., 1997] Huopaniemi, J., Savioja, L., and Karjalainen, M. (1997). Modeling of reflections and air absorption in acoustical spaces - a digital filter design approach. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 19–22, New Paltz, New York, USA.

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

[ISO, 2009a] ISO (2009a). ISO 226:2003 acoustics - normal equal-loudness-level contours. `http://www.iso.org/`, accessed 21-09-2009.

[ISO, 2009b] ISO (2009b). ISO 3382-1:2009 acoustics - measurement of room acoustic parameters - part 1: Performance spaces. `http://www.iso.org/`, accessed 21-09-2009.

[ITU, 2006] ITU (2006). Rec. ITU-R BS. 775-2: Multichannel stereophonic sound system with and without accompanying picture. Technical report, International Telecommunications Union.

[J. M. Jot and Warusfel, 1997] J. M. Jot, L. C. and Warusfel, O. (1997). Analysis and synthesis of room reverberation based on a statistical time-frequency model. In *103rd Convention of the Audio Engineering Society*, New York, New York, USA. Paper 4629.

[Jacques et al., 2005] Jacques, R., Albrecht, B., Melchior, F., and de Vries, D. (2005). An approach for multichannel recording and reproduction of sound source directivity. In *119th convention of the Audio Engineering Society*, New York, New York, USA. Paper 6554.

[Jot et al., 1999] Jot, J., Larcher, V., and Pernaux, J. (1999). A comparative study of 3-D audio encoding and rendering techniques. In *16th International Conference of the Audio Engineering Society*, Arktikum, Rovaniemi, Finland. Paper 16-025.

[Jourjine et al., 2000] Jourjine, A. N., Rickard, S., zgr Yilmaz, and Yilmaz, O. (2000). Blind separation of disjoint orthogonal signals. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2985–2988.

[Karamustafaoglu et al., 1999] Karamustafaoglu, A., Horbach, U., Pellegrini, R., Mackensen, P., and Theile, G. (1999). Design and applications of a data-based auralisation system for surround sound. In *106th convention of the Audio Engineering Society*, Munich, Germany. Paper 4976.

[Katz, 2004] Katz, B. F. G. (2004). International round robin on room acoustical impulse response analysis software 2004. *Acoustics Research Letters Online*, 5(4):158–164.

[Kearney et al., 2007] Kearney, G., Bates, E., Furlong, D., and Boland, F. (2007). A comparative study of the performance of spatialization techniques for a distributed audience in a concert hall environment. In *31st International Conference of the Audio Engineering Society*, London, UK. Paper 13.

[Kearney and Levison, 2008] Kearney, G. and Levison, J. (2008). Actual vs. virtual multichannel acoustic recording. In *124th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands. Paper 7326.

[Kearney et al., 2009a] Kearney, G., Masterson, C., Adams, S., and Boland, F. (2009a). Approximation of binaural room impulse responses. In *IET Irish Signals and Systems Conference*, Dublin, Ireland.

[Kearney et al., 2009b] Kearney, G., Masterson, C., Adams, S., and Boland, F. (2009b). Towards efficient binaural room impulse response synthesis. In *EAA Symposium on Auralization*, Espoo, Finland.

[Keet, 1968] Keet, W. (1968). The influence of early reflections on the spatial impression. In *6th Int. Congr. on Acoustics, E-2-4*, Tokyo, Japan.

[Kessler, 2005] Kessler, R. (2005). An optimised method for capturing multidimensional acoustic fingerprints. In *118th Convention of the Audio Engineering Society*, Barcelona, Spain. Paper 6342.

[Kirkeby and Nelson, 1996] Kirkeby, O. and Nelson, P. (1996). Fast deconvolution of multichannel systems using regularisation. In *ISVR Technical Report no. 255, Southampton.*

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

[Kleiner, 1978] Kleiner, M. (1978). Problems in the design and use of dummy heads. *Acustica*, 41:183–193.

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

[Klumpp and Eady, 1956] Klumpp, R. G. and Eady, H. R. (1956). Some measurements of interaural time difference thresholds. *The Journal of the Acoustical Society of America*, 28(5):859–860.

[Knapp and Carter, 1976] Knapp, C. H. and Carter, G. C. (1976). The generalized correlation method for estimation of time delay. *IEE Transactions on Acoustics, Speech and Signal Processing*, 24(1):320–327.

[Krokstad et al., 1968] Krokstad, A., Strøm, S., and Sørsdal, S. (1968). Calculating the acoustical room response by the use of a ray tracing technique. *Journal of Sound and Vibration*, 8(1):118 – 125.

[Kuttruff, 1979] Kuttruff, H. (1979). *Room Acoustics*. Appl. Science Publ., London.

[Laitinen, 2008] Laitinen, M. (2008). Binaural Reproduction for Directional Audio Coding. Masters Thesis, Helsinki University of Technology, `http://lib.tkk.fi/Dipl/2008/urn012250.pdf`, accessed 29-3-2010.

[Lee, 2007] Lee, J. (2007). Head tracking for desktop VR displays using the Wii remote. `http://www.cs.cmu.edu/~johnny/projects/wii/`, acessed 29/03/2010.

[Leitner et al., 2000] Leitner, S., Sontacchi, A., and Holdrich, R. (2000). Multi-channel sound reproduction system for binaural signals the Ambisonic approach. In *Proceedings of the COST G-6 conference on Digital Audio Effects, DAFX-00*, Verona, Italy.

[Lipshitz, 1986] Lipshitz, S. P. (1986). Stereo microphone techniques: Are the purists wrong? *Journal of the Audio Engineering Society*, 34(9):716–744.

[Lipshitz and Vanderkooy, 1987] Lipshitz, S. P. and Vanderkooy, J. (1987). Anomalies of wavefront reconstruction in stereo and surround-sound reproduction. In *83rd Convention of the Audio Engineering Society*, New York, New York, USA. Paper 2554.

[Lokki and Savioja, 2005] Lokki, T. and Savioja, L. (2005). Evaluation of auralization results. In *Forum Acusticum*, Budapest.

[Lundeby et al., 1995] Lundeby, A., Vigran, T. E., Bietz, H., and Vorlander, M. (1995). Uncertainties of measurements in room acoustics. *Acustica*, 81:344355.

[Malham, 1992] Malham, D. G. (1992). Experience with large area 3-D Ambisonic sound systems. *Institute of Acoustics*, 8(5):209–216.

[Mason et al., 2003] Mason, R., Brookes, T., and Rumsey, F. (2003). Creation and verification of a controlled experimental stimulus for investigating selected perceived spatial attributes. In *114th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands. Paper 5771.

[Masterson and Boland, 2008] Masterson, C. and Boland, F. (2008). Head related impulse response simplication by deconvolution. In *8th IMA International Conference on Mathematics in Signal Processing*, Royal Agricultural College, Cirencester, UK.

[McKeag and McGrath, 1996] McKeag, A. and McGrath, D. (1996). Sound field format to binaural decoder with head-tracking. In *6th Australian Regional Convention of the AES*. Paper 4302.

[McKeag and McGrath, 1997] McKeag, A. and McGrath, D. (1997). Using auralisation techniques to render 5.1 surround to binaural and playback. In *102nd Convention of the Audio Engineering Society*, Munich, Germany. Paper 4458.

[Meesawat and Bajers, 2002] Meesawat, K. and Bajers, F. (2002). An investigation on the transition from early relections to a reverberation tail in a BRIR. In *Proceedings of the 2002 International Conference on Auditory Display, Kyoto, Japan.*

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

[Menzies, 2007] Menzies, D. (2007). Nearfield synthesis of complex sources with High Order Ambisonics, and binaural rendering. In *Proceedings of the 13th International Conference on Auditory Display.*

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

[Middlebrooks and Green, 1990] Middlebrooks, J. C. and Green, D. M. (1990). Directional dependence of interaural envelope delays. *The Journal of the Acoustical Society of America*, 87:2149–2162.

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

[Møller et al., 1995] Møller, H., Hammershøi, D., Jensen, C., and Sørensen, M. (1995). Transfer characteristics of headphones measured on human ears. *Journal of the Audio Engineering Society*, 43(4):203–217.

[Moore, 1995] Moore, B. (1995). *Hearing.* Academic Press.

[Moore et al., 1988] Moore, B. C. J., Glasberg, B., Plack, C., and Biswas, A. (1988). The shape of the ear's temporal window. *The Journal of the Acoustical Society of America*, 83(3):1102–1106.

[Moreau et al., 2006] Moreau, S., Daniel, J., and Bertet, S. (2006). 3D sound field recording with Higher Order Ambisonics- objective measurements and validation of a 4th order spherical microphone. In *120th Convention of the Audio Engineering Society*, Paris, France. Paper 6857.

[Mourjopoulos, 1994] Mourjopoulos, J. (1994). Digital equalization of room acoustics. *Journal of the Audio Engineering Society*, 42(11):884–900.

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

[Naylor and Rindel, 1992] Naylor, G. and Rindel, J. H. (1992). Predicting room acoustical behaviour with the odean computer model. In *124th Meeting of Acoustical Society of America*.

[Neely and Allen, 1979] Neely, S. and Allen, J. (1979). Invertibility of a room impulse response. *The Journal of the Acoustical Society of America*, 66:165–169.

[Nielsen, 1993] Nielsen, S. H. (1993). Auditory distance perception in different rooms. *Journal of the Audio Engineering Society*, 41(10):755–770.

[Noisternig et al., 2002] Noisternig, M., Sontacchi, A., Musil, T., and Holdrich, R. (2002). A 3D Ambisonic based binaural sound reproduction system. In *24th International Conference of the Audio Engineering Society*. Paper 1.

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

[Omologo and Svaizer, 1993] Omologo, M. and Svaizer, P. (1993). Talker Localization and Speech Enhancement in a Noisy Environment using a Microphone Array based Acquisition System. In *Proceedings Eurospeech*, pages 605–609, Berlin.

[Omologo and Svaizer, 1994] Omologo, M. and Svaizer, P. (1994). Acoustic Event Localization Using a Crosspower-Spectrum Phase Based Technique. In *Proc. ICASSP '94*, pages II–273–II–276, Adelaide, Austrailia.

[Otani and Hirahara, 2008] Otani, M. and Hirahara, T. (2008). Auditory artifacts due to switching head-related transfer functions of a dynamic virtual auditory display. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, E91-A(6):1320–1328.

[Otondo and Rindel, 2005] Otondo, F. and Rindel, J. H. (2005). A new method for the radiation representation of musical instruments in auralizations. *Acta Acustica united with Acustica*, 91(5):902–906.

[Patterson, 1976] Patterson, R. (1976). Auditory filtershapes derived with noise stimuli. *The Journal of the Acoustical Society of America*, 59(3):640–654.

[Patynen et al., 2008a] Patynen, J., Pulkki, V., and Lokki, T. (2008a). Anechoic recording system for symphony orchestra. *Acta Acustica united with Acustica*, 94(6):856–865.

[Patynen et al., 2008b] Patynen, J., Tervo, S., and Lokki, T. (2008b). A Loudspeaker Orchestra For Concert Hall Studies. *Proceedings of the Institute of Acoustics*, 30(3):45–52.

[Physikalisch-Technische-Bundesanstalt, 2009] Physikalisch-Technische-Bundesanstalt (2009). Directivities of musical instruments. `http://www.ptb.de/en/org/1/17/173/richtchar.htm`, accessed 1-6-2009.

[Pietrzyk, 1998] Pietrzyk, A. (1998). Computer modeling of the sound field in small rooms. In *15th International Conference of the Audio Engineering Society*, Copenhagen, Denmark. Paper 15-004.

[Plack and Moore, 1990] Plack, C. J. and Moore, B. C. J. (1990). Temporal window shape as a function of frequency and level. *The Journal of the Acoustical Society of America*, 87(5):2178–2187.

[Plack and Oxenham, 1998] Plack, C. J. and Oxenham, A. J. (1998). Basilar-membrane nonlinearity and the growth of forward masking. *The Journal of the Acoustical Society of America*, 103(3):1598–1608.

[Poletti, 2005] Poletti, M. (2005). Effect of noise and transducer variability on the performance of circular microphone arrays. *Journal of the Audio Engineering Society*, 53(5):598–610.

[Poletti, 2007] Poletti, M. (2007). Robust two-dimensional surround sound reproduction for nonuniform loudspeaker layouts. *Journal of the Audio Engineering Society*, 55(7/8):598–610.

[Pralong and Carlile, 1996] Pralong, D. and Carlile, S. (1996). The role of individualized headphone calibration for the generation of high fidelity virtual auditory space. *The Journal of the Acoustical Society of America*, 100(6):3785–3793.

[Preibish-Effenberger, 1966] Preibish-Effenberger, R. (1966). On the techniques of directional audiometry: Tests of the faculty of sound localisation by using an electronic delay line or by measurements in a free sound field? *Arch. klin. u. exp. Ohren-Nasen-Kehlkopfheilk*, 187:588–592.

[Pulkki, 1997] Pulkki, V. (1997). Virtual sound source positioning using Vector Base Amplitude Panning. *Journal of the Audio Engineering Society*, 45(6):456–466.

[Pulkki, 2001a] Pulkki, V. (2001a). Coloration of amplitude-panned virtual sources. In *110th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands. Paper 5402.

[Pulkki, 2001b] Pulkki, V. (2001b). Localization of amplitude-panned virtual sources I: Stereophonic panning. *Journal of the Audio Engineering Society*, 49(9):739–751.

[Pulkki, 2001c] Pulkki, V. (2001c). Localization of amplitude-panned virtual sources II: Two and three dimensional panning. *Journal of the Audio Engineering Society*, 49(9):753–767.

[Pulkki and Merimaa, 2006] Pulkki, V. and Merimaa, J. (2006). Spatial Impulse Response Rendering II: Reproduction of diffuse sound and listening tests. *Journal of the Audio Engineering Society*, 54(1/2):3–20.

[Rabinowitz et al., 1993] Rabinowitz, W. R., Maxwell, J., Shao, Y., and Wei, M. (1993). Sound localization cues for a magnified head: Implications from sound diffraction about a rigid sphere. *Presence*, 2:125–129.

[Rakerd and Hartmann, 1985] Rakerd, B. and Hartmann, W. M. (1985). Localization of sound in rooms, II: The effects of a single reflecting surface. *The Journal of the Acoustical Society of America*, 78:524–533.

[Ratliff, 1974] Ratliff, P. A. (1974). Properties of Hearing Related to Quadraphonic Reproduction. BBC Research Department, Report BBC RD 1974/38.

[Rayleigh, 1945] Rayleigh, J. (1945). *Theory of Sound*. Dover, N.Y.

[Renkus-Heinz, 2009] Renkus-Heinz (2009). EASE (Enhanced Acoustic Simulator for Engineers) auralization software. `http://www.renkus-heinz.com/ease/index.html, accessed24-10-09`.

[Rindel and Christensen, 2003] Rindel, J. H. and Christensen, C. L. (2003). Room acoustic simulation and auralization how close can we get to the real room? In *The Eighth Western Pacific Acoustics Conference (WESPAC8)*, Melbourne, Australia.

[Rindel et al., 1994] Rindel, J. H., Lynge, C., Naylor, G., and Rishoj, K. (1994). The use of a digital audio mainframe for room acoustical auralization. In *96th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands. Paper 3860.

[Rucci and Wray, 1999] Rucci, M. and Wray, J. (1999). Binaural cross-correlation and auditory localization in the barn owl: a theoretical study. *Neural Netw.*, 12(1):31–42.

[Rumsey, 2001] Rumsey, F. (2001). *Spatial Audio*. Focal Press.

[Rychtarikova et al., 2009] Rychtarikova, M., Bogaert, T. V. d., Vermeir, G., and Wouters, J. (2009). Binaural sound source localization in real and virtual rooms. *Journal of the Audio Engineering Society*, 57(4):205–220.

[Sabine, 1964] Sabine, W. C. (1964). *Reverberation, Collected Papers on Acoustics*. Dover Publications.

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

[Salvador and Chan, 2004] Salvador, S. and Chan, P. (2004). FastDTW: Toward accurate Dynamic Time Warping in linear time and space. Dept. of Computer Sciences, Florida Institute of Technology, `http://cs.fit.edu/~pkc/papers/tdm04.pdf`, accessed 29-3-2010.

[Sandvad and Hammershøi, 1994] Sandvad, J. and Hammershøi, D. (1994). What is the most efficient way of representing HTF filters? In *Proceedings of NORSIG'94, Alesund, Norway*, pages 174–178.

[Savioja et al., 1999] Savioja, L., Huopaniemi, J., Lokki, T., and Väänänen, R. (1999). Creating Interactive Virtual Acoustic Environments. *Journal of the Audio Engineering Society*, 47(9):675–705.

[Savioja et al., 1994] Savioja, L., Rinne, T., and Takala, T. (1994). Simulation of room acoustics with a 3-D finite difference mesh. In *International Computer Music Conference*, pages 463–466, Aarhus, Denmark.

[Scarpaci, 2006] Scarpaci, J. (2006). *Creation of a system for real time virtual auditory space and its application to dynamic sound localization.* PhD thesis, University of Cincinnati.

[Schroeder, 1965] Schroeder, M. R. (1965). New method of measuring reverberation time. *The Journal of the Acoustical Society of America*, 37(3):409–412.

[Silverman, 1986] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

[Silverman et al., 1997] Silverman, H. F., III, W. R. P., and Flanagan, J. L. (1997). The Huge Microphone Array (hma). *The Journal of the Acoustical Society of America*, 101(5):3119–3119.

[Smith, 1992] Smith, J. O. (1992). Physical modeling using digital waveguides. *Computer Music Journal*, 16(4):74–78.

[Soundfield Ltd., 2009] Soundfield Ltd. (2009). Soundfield MK5 microphone user guide 527-027.

[Southern et al., 2009] Southern, A., Wells, J., and Murphy, D. (2009). Rendering walk-through auralisations using wave-based acoustical models. In *Proc. of the 17th European Signal Processing Conference (EUSIPCO)*, pages 715 – 719.

[Start, 1997] Start, E. W. (1997). Direct sound enhancement using Wave Field Synthesis. Dissertation, Delft University of Technology, Delft, The Netherlands.

[Stevens, 1975] Stevens, S. S. (1975). *Psychophysics: Introduction to its perception, neural and social prospects.* New York: Wiley.

[Stewart and Sandler, 2007] Stewart, R. and Sandler, M. (2007). Statistical measures of early reflections of room impulse responses. In *Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07)*, Bordeaux, France.

[Streicher and Everest, 2006] Streicher, R. and Everest, F. A. (2006). *The New Stereo Soundbook, Third Edition*. Audio Engineering Associates.

[Struck and Temme, 1994] Struck, C. and Temme, S. (1994). Simulated free field measurements. *Journal of the Audio Engineering Society*, 42(6):467–482.

[Stuller, 1987] Stuller, J. (1987). Maximum-likelihood estimation of time-varying delay–Part I. *Acoustics, Speech, and Signal Processing, IEEE Transactions on,*, 35(3):300–313.

[Theile, 2001] Theile, G. (2001). Multichannel Natural Music Recording Based On Psychacoustic Principles. IRT, Extended version of the paper presented at the 19th International Conference of the Audio Engineering Society, May 2001.

[Theile and Plenge, 1977] Theile, G. and Plenge, G. (1977). Localization of lateral phantom sources. *Journal of the Audio Engineering Society*, 25(4):196–200.

[Thompson et al., 2004] Thompson, W. B., Willemsen, P., Gooch, A. A., Creem-Regehr, S. H., Loomis, J. M., and Beall, A. C. (2004). Does the quality of the computer graphics matter when judging distances in visually immersive environments. *Presence: Teleoper. Virtual Environ.*, 13(5):560–571.

[Torick, 1998] Torick, E. (1998). Highlights in the history of multichannel sound. *Journal of the Audio Engineering Society*, 46(1/2):27–31.

[UDK, 1994] UDK, B. (1994). Decca tree recording mit neumann-druckempfngern m50. `www.sengpielaudio.com/DeccaTreeRecordingM50.pdf`, acessed 29-03-10.

[Van de Veer, 1957] Van de Veer, R. (1957). Some experiments on directional hearing. Dissertation, University of Amsterdam.

[Von Bekesy, G., 1960] Von Bekesy, G. (1960). *Experiments in Hearing*. McGraw-Hill Book Company.

[Vorländer, 1995] Vorländer, M. (1995). International round-robin on room acoustical computer simulations. In *Proc 15th ICA, Trondheim, Norway*, pages 577–580.

[Waller, 1999] Waller, D. (1999). Factors affecting the perception of interobject distances in virtual environments. *Presence: Teleoper. Virtual Environ.*, 8(6):657–670.

[Wiggins, 2004] Wiggins, B. (2004). An investigation into the real-time manipulation and control of three dimensional soundfields. Ph.D Thesis, University of Derby, `http://sparg.derby.ac.uk/SPARG/PDFs/BWPhDThesis.pdf`, accessed 29-03-10.

[Wiggins, 2007] Wiggins, B. (2007). The generation of panning Laws for irregular speaker arrays using heuristic methods. In *Proceedings of the 31st International AES conference, London, UK.*

[Wilkens, 1972] Wilkens, H. (1972). Head related stereophony, an aid in comparing and judging various spatial impressions. *Acustica*, 26:213–221.

[Williams, 1987] Williams, M. (1987). Unified theory of microphone systems for stereophonic sound recording. In *82nd Convention of the Audio Engineering Society*, London, UK. Paper 2466.

[Williams, 1991] Williams, M. (1991). Microphone arrays for natural multiphony. In *91st Convention of the Audio Engineering Society*, New York, New York, USA. Paper 3157.

[Williams and Le Dû, 2001] Williams, M. and Le Dû, G. (2001). The quick reference guide to multichannel microphone arrays, part 1: Using cardioid microphones. In *110th Convention of the Audio Engineering Society*, Amsterdam, The Netherlands. Paper 6059.

[Wittek, 2007] Wittek, H. (2007). Perceptual differences between Wavefield Synthesis and Stereophony. Department of Music and Sound Recording, School of Arts, Communication and Humanities, University of Surrey.

[Wittek et al., 2004] Wittek, H., Kerber, S., Rumsey, F., and Theile, G. (2004). Spatial perception in Wave Field Synthesis rendered sound fields: Distance of real and virtual nearby sources. In *116th Convention of the Audio Engineering Society*, Berlin, Germany. Paper 6000.

[Wittek and Theile, 2002] Wittek, H. and Theile, G. (2002). The recording angle- based on the localisation curves. In *112th Convention of the Audio Engineering Society*, Munich, Germany. Paper 5568.

[Zeng and Dayton, 2004] Zeng, Z. and Dayton, B. (2004). The approximate GCD of inexact polynomials. *Proceedings of the 2004 international symposium on Symbolic and algebraic computation*, pages 320–327.

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

A

Test Environment 1

A.1 Test environment 1 measurement and simulation

Test environment 1, shown in Figure A.1 is the Printing House Hall located in Trinity College Dublin. It is the main reverberant room used throughout the course of this thesis work. The room dimensions are approximately 5.85m x 5.85m x 15m. The acoustic measurements presented in Section A.1.1 were taken using the logarithmic swept sine technique [Farina, 2000] at the measurement points shown in Figure A.2. Source and receivers are at a height of 1.2m. A B& K OmniPower Sound Source Type 4296 is used as the omnidirectional source in the measurements and a Neumann KU100 is used as the binaural receiver. The measurements conform to ISO-3382 (2009) [ISO, 2009b] standards and are presented in octave bands from 125 to 4000Hz. Measurement points are chosen which are derived from specifications by Gade [Gade, 1989], [Gade, 1982]. Equivalent acoustic parameters from ray-tracing simulations are presented in Section A.1.2. Here, 100ms BRIRs are computed from 10000 rays, and a randomized reverberation tail, appended to the each resultant IR. The results show similar trends in ITD, ILD and IACC values, and although significant deviations can be found in various parameters at different octave bands, overall the simulation gives good insight into attributes of the true environment that are difficult to capture in the real world.



Figure A.1: *Test Environment 1: Printing House Hall at Trinity College Dublin*

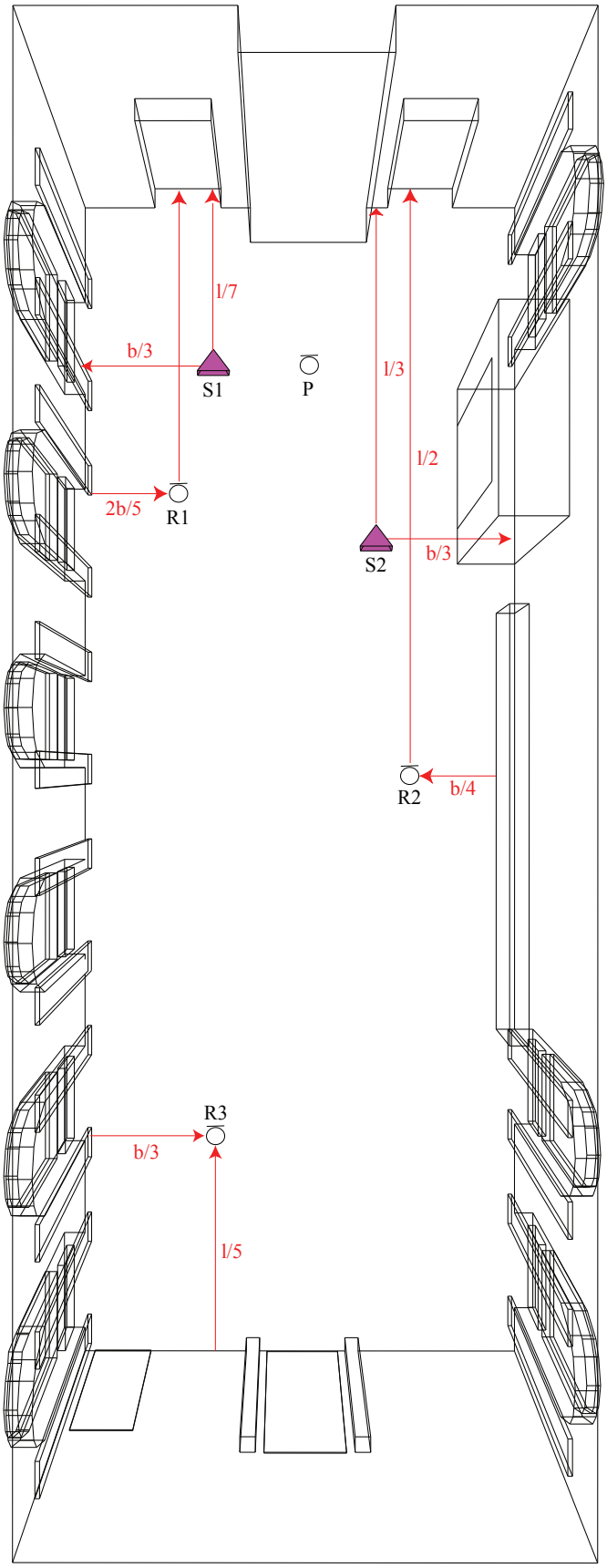


Figure A.2: *Measurement points for binaural impulse response capture and corresponding simulations*

A.1.1 Test Environment 1:Acoustic measurements

Test Environment 1, BIR Measurement, Source 1, Position P

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS
Left Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	22.53	37.399	55.861	66.82	65.453	67.226	66.487	64.961	63.062	56.066	73.208	74.162
Noise [dB]	15.084	17.693	9.87	19.517	19.054	21.787	24.583	17.479	1.111	7.514	28.64	29.279
strenGth [dB]	-46.47	-31.601	-13.139	-2.18	-3.547	-1.774	-2.513	-4.039	-5.938	-12.934	-3.792	-2.838
C50 [dB]	3.631	5.27	-5.057	1.129	-0.257	-0.529	-2.855	-1.886	-3.129	-0.984	-1.668	-1.122
C80 [dB]	6.973	7.355	0.923	5.069	4.075	3.993	1.242	2.047	5.059	7.483	2.77	3.359
D50 [%]	69.764	77.09	23.787	56.463	48.522	46.957	34.135	39.312	32.731	44.362	40.515	43.577
Ts [ms]	70.074	68.842	119.429	72.969	76.098	78.361	96.668	87.418	71.824	59.778	84.07	79.679
EDT [s]	0.676	0.645	1.292	0.859	0.869	0.941	1.069	0.913	0.579	0.387	0.94	0.933
T20 [s]	0.5	0.893	0.991	1.04	1.026	1.017	1.073	0.884	0.676	0.508	0.995	0.999
r T20	0.977	0.987	0.994	0.985	0.999	0.997	0.997	0.999	0.999	0.998	0.999	0.999
T30 [s]	0.476	0.761	1.022	0.991	0.973	1.016	1.056	0.929	0.684	0.503	1.008	1.001
r T30	0.969	0.947	0.995	0.995	0.998	0.998	0.998	0.999	0.999	0.999	0.999	1
Tuser [s]	0.423	0.791	1.035	0.981	0.975	1.048	1.114	0.918	0.672	0.487	1.038	1.016
r Tuser	0.958	0.947	0.996	0.985	0.997	0.999	0.998	0.999	0.999	0.999	0.999	0.999
IACC (Early)	0.977	0.963	0.812	0.832	0.466	0.379	0.207	0.196	0.075	0.124	0.212	0.33
Tau IACC (ms)	-0.408	-0.884	-0.59	-0.612	-0.794	-0.703	-0.748	-0.862	-0.68	-0.703	-0.68	-0.68
w IACC (ms)	0	1.882	1.02	0.544	0.295	0.159	0.113	0.045	0.045	0.045	0.091	0.113

RTU = RT User (-10. dB, -30. dB)

Right Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	22.17	35.984	56.083	67.673	67.446	69.599	68.832	67.548	70.518	63.596	76.332	76.966
Noise [dB]	9.835	17.056	12.026	18.405	16.242	19.596	22.884	17.117	0.197	6.316	26.896	27.554
strenGth [dB]	-46.83	-33.016	-12.917	-1.327	-1.554	0.599	-0.168	-1.452	1.518	-5.404	-0.668	-0.034
C50 [dB]	4.371	5.245	-0.591	3.663	2.789	7.079	3.962	4.533	10.27	11.605	5.788	5.454
C80 [dB]	5.861	8.673	2.88	6.434	6.692	8.787	6.057	7.938	14.588	16.327	8.342	8.134
D50 [%]	73.234	76.989	46.603	69.921	65.523	83.619	71.345	73.96	91.411	93.537	79.131	77.83
Ts [ms]	76.126	95.362	113.476	59.964	48.095	45.371	56.335	48.184	12.9	10.556	40.958	41.775
EDT [s]	0.952	0.518	1.429	0.937	0.789	0.709	0.912	0.664	0.192	0.201	0.681	0.73
T20 [s]	1.029	0.995	1.005	0.975	1.018	1.071	1.043	0.912	0.57	0.405	0.997	0.999
r T20	0.901	0.975	0.989	0.992	0.998	0.998	0.999	0.999	0.969	0.965	0.999	0.999
T30 [s]	0.982	0.848	0.973	0.983	0.988	1.052	1.03	0.934	0.643	0.46	1.001	0.995
r T30	0.853	0.93	0.994	0.996	0.998	0.999	0.999	0.999	0.988	0.985	1	0.999
Tuser [s] --		0.911	1.004	1.015	0.999	1.04	1.011	0.928	0.66	0.486	0.996	0.989
r Tuser --		0.884	0.994	0.991	0.997	0.998	0.999	0.999	0.996	0.996	1	0.999
IACC (Early)	0.977	0.963	0.812	0.832	0.466	0.379	0.207	0.196	0.075	0.124	0.212	0.33
Tau IACC (ms)	-0.408	-0.884	-0.59	-0.612	-0.794	-0.703	-0.748	-0.862	-0.68	-0.703	-0.68	-0.68
w IACC (ms)	0	1.882	1.02	0.544	0.295	0.159	0.113	0.045	0.045	0.045	0.091	0.113

RTU = RT User (-10. dB, -30. dB)

Table A.1: *ISO-3382 acoustical parameters: Test environment 1, source position 1, receiver position P*

Test Environment 1, BIR Measurement, Source 1, Position R1

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Left Channel Parameters

Freq. [Hz]		31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]		20.539	33.642	56.485	63.846	64.296	65.598	66.762	66.343	63.448	60.32	73.144	73.543
Noise [dB]		12.686	15.468	11.792	14.186	12.13	15.355	17.614	11.099	0.112	7.801	21.913	23.805
strenGth [dB]		-48.461	-35.358	-12.515	-5.154	-4.704	-3.402	-2.238	-2.657	-5.552	-8.68	-3.856	-3.457
C50 [dB]		-3.146	-3.333	-5.753	-7.837	-4.214	-1.765	-2.623	-0.807	-2.43	4.756	-2.084	-2.805
C80 [dB]		1.122	-0.368	-1.885	-2.003	2.313	2.756	2.164	5.064	4.394	10.906	3.164	2.483
D50 [%]		32.641	31.703	21.002	14.13	27.482	39.979	35.342	45.37	36.363	74.934	38.23	34.393
Ts [ms]		129.299	131.965	115.347	120.778	93.911	87.578	90.436	72.658	73.033	49.606	82.881	86.989
EDT [s]		0.776	1.347	0.709	1.063	0.913	0.921	0.918	0.634	0.515	0.244	0.82	0.846
T20 [s]		0.456	0.943	1.246	1.058	0.978	1.018	1.06	0.896	0.665	0.442	0.995	1.018
r T20		0.913	0.954	0.934	0.997	0.998	0.999	0.998	0.999	0.999	0.994	0.999	0.999
T30 [s]		0	0.848	1.011	1.054	0.993	1.021	1.102	0.938	0.661	0.469	1.03	1.033
r T30		101	0.962	0.954	0.996	0.999	1	0.998	0.999	1	0.997	0.999	1
Tuser [s]	--		0.692	1.118	1.078	1.004	1.027	1.101	0.926	0.659	0.48	1.036	1.033
r Tuser	--		0.99	0.942	0.995	0.998	0.999	0.999	0.999	0.999	0.998	1	1
IACC (Early)		0.98	0.946	0.895	0.556	0.258	0.419	0.379	0.347	0.12	0.171	0.312	0.332
Tau IACC (ms)		0.159	-0.068	-0.045	0.068	0.295	0.136	0.091	0.091	0.068	-0.272	0.091	0.091
w IACC (ms)		0	1.542	1.02	0.567	0.317	0.159	0.113	0.045	0.045	0.045	0.068	0.091

RTU = RT User (-10. dB, -30. dB)

Right Channel Parameters

Freq. [Hz]		31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]		18.55	32.977	56.624	63.458	64.58	63.935	66.162	64.945	62.296	58.701	72.165	72.765
Noise [dB]		10.018	13.856	11.041	15.729	9.528	14.675	18.17	9.904	-0.92	7.352	21.755	23.271
strenGth [dB]		-50.45	-36.023	-12.376	-5.542	-4.42	-5.065	-2.838	-4.055	-6.704	-10.299	-4.835	-4.235
C50 [dB]		-4.319	-2.12	-0.954	-8.683	-7.651	-5.087	-4.638	-2.197	-3.36	3.367	-4.069	-4.746
C80 [dB]		1.199	0.585	1.968	-0.77	-0.224	0.165	2.333	4.007	3.632	9.749	2.236	1.606
D50 [%]		27.001	38.035	44.528	11.927	14.658	23.661	25.581	37.616	31.567	68.467	28.152	25.109
Ts [ms]		204.447	153.821	111.724	132.066	87.959	128.438	103.835	100.262	95.218	72.021	103.85	104.06
EDT [s]		0.781	1.161	0.666	1.028	0.889	1.037	1.037	0.774	0.594	0.236	0.918	0.904
T20 [s]		0.903	0.984	1.246	1.04	0.94	1.048	0.989	0.904	0.65	0.462	0.97	0.987
r T20		0.893	0.968	0.968	0.992	0.993	0.997	0.998	0.999	0.999	0.995	0.998	0.999
T30 [s]	--		0.895	1.082	1.042	0.967	1.028	1.033	0.947	0.668	0.481	1.001	1.006
r T30	--		0.968	0.975	0.995	0.997	0.999	0.998	0.999	0.999	0.998	0.999	0.999
Tuser [s]	--		0.829	1.177	1.07	0.987	1.053	1.034	0.957	0.658	0.487	1.016	1.021
r Tuser	--		0.986	0.968	0.988	0.996	0.998	0.998	0.999	0.998	0.996	0.999	0.999
IACC (Early)		0.98	0.946	0.895	0.556	0.258	0.419	0.379	0.347	0.12	0.171	0.312	0.332
Tau IACC (ms)		0.159	-0.068	-0.045	0.068	0.295	0.136	0.091	0.091	0.068	-0.272	0.091	0.091
w IACC (ms)		0	1.542	1.02	0.567	0.317	0.159	0.113	0.045	0.045	0.045	0.068	0.091

RTU = RT User (-10. dB, -30. dB)

Table A.2: *ISO 3383 Octave band acoustical parameters: Test environment 1, source position 1, receiver position R1*

Test Environment 1, BIR Measurement, Source 1, Position R2

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Left Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	22.413	33.017	54.278	64.723	64.42	64.711	66.05	64.547	62.406	56.954	72.139	72.834
Noise [dB]	13.524	14.111	15.382	19.15	18.13	20.336	23.515	17.603	1.698	7.72	27.632	28.299
strenGth [dB]	-46.587	-35.983	-14.722	-4.277	-4.58	-4.289	-2.95	-4.453	-6.594	-12.046	-4.861	-4.166
C50 [dB]	0.46	-8.888	-10.976	-13.01	-18.242	-8.394	-11.16	-6.931	-11.85	-11.894	-9.77	-10.721
C80 [dB]	5.258	-0.536	-5.097	-1.845	-2.248	-1.405	-1.031	0.765	0.826	5.392	-0.536	-1.092
D50 [%]	52.648	11.441	7.396	4.762	1.477	12.644	7.111	16.855	6.131	6.073	9.538	7.81
Ts [ms]	120.519	127.244	163.708	116.81	111.599	111.955	108.83	98.015	90.423	72.37	104.266	106.682
EDT [s]	0.707	0.731	1.146	1.012	0.768	0.999	0.869	0.855	0.617	0.35	0.858	0.891
T20 [s]	1.573	0.954	1.007	1.014	0.943	0.963	1.109	0.902	0.635	0.458	1.002	0.991
r T20	0.925	0.987	0.992	0.988	0.996	0.998	0.997	0.999	0.999	0.998	0.999	0.999
T30 [s]	1.405	0.871	0.989	1.05	0.98	0.994	1.054	0.925	0.649	0.471	0.999	1.004
r T30	0.896	0.985	0.995	0.994	0.997	0.999	0.998	1	0.999	0.999	1	0.999
Tuser [s]	1.761	0.857	1.043	1.066	0.95	0.959	1.062	0.931	0.653	0.457	1	0.999
r Tuser	0.87	0.987	0.991	0.992	0.996	0.999	0.994	0.999	0.999	0.998	0.999	0.999
IACC (Early)	0.998	0.987	0.934	0.77	0.202	0.08	0.375	0.139	0.203	0.211	0.179	0.246
Tau IACC (ms)	-0.272	0	0.34	-0.113	0.023	0.816	0.159	-0.567	0.794	-0.317	0.136	0.136
w IACC (ms)	0	1.814	0.907	0.612	0.272	0.136	0.091	0.045	0.045	0.045	0.068	0.091

RTU = RT User (-10. dB, -30. dB)

Right Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	21.757	32.549	52.73	62.399	62.491	62.865	64.477	64.647	61.121	53.669	70.962	71.406
Noise [dB]	12.904	12.345	16.006	19.027	19.286	18.353	23.383	15.317	0.477	6.679	26.725	27.721
strenGth [dB]	-47.243	-36.451	-16.27	-6.601	-6.509	-6.135	-4.523	-4.353	-7.879	-15.331	-6.038	-5.594
C50 [dB]	-0.031	-9.738	-13.262	-8.375	-19.542	-13.445	-15.494	-14.185	-19.039	-19.017	-15.024	-14.608
C80 [dB]	5.063	-1.052	-7.888	-0.25	-3.064	-2.751	-2.06	1.848	0.481	3.484	-0.703	-0.965
D50 [%]	49.819	9.601	4.506	12.693	1.099	4.328	2.745	3.675	1.232	1.239	3.049	3.345
Ts [ms]	140.178	141.72	233.81	199.495	174.024	171.262	156.326	95.789	121.55	154.205	136.719	148.21
EDT [s]	0.612	0.765	1.178	1.195	0.852	0.957	0.929	0.783	0.577	0.477	0.862	0.915
T20 [s]	1.532	0.998	1.077	1.132	1.012	0.993	1.029	0.864	0.637	0.463	0.971	0.998
r T20	0.868	0.984	0.981	0.993	0.993	0.998	0.999	0.999	1	0.997	0.999	0.999
T30 [s]	1.475	0.936	1.048	1.073	1.002	1.002	1.051	0.887	0.659	0.482	1.002	1.014
r T30	0.873	0.981	0.993	0.995	0.997	0.999	0.999	0.999	0.999	0.998	0.999	0.999
Tuser [s] --		1.007	1.125	1.061	1.005	0.995	1.054	0.883	0.652	0.474	0.997	1.012
r Tuser --		0.988	0.983	0.992	0.993	0.999	0.999	0.999	1	0.997	1	1
IACC (Early)	0.998	0.987	0.934	0.77	0.202	0.08	0.375	0.139	0.203	0.211	0.179	0.246
Tau IACC (ms)	-0.272	0	0.34	-0.113	0.023	0.816	0.159	-0.567	0.794	-0.317	0.136	0.136
w IACC (ms)	0	1.814	0.907	0.612	0.272	0.136	0.091	0.045	0.045	0.045	0.068	0.091

RTU = RT User (-10. dB, -30. dB)

Table A.3: *ISO 3383 Octave band acoustical parameters: Test environment 1, source position 1, receiver position R2*

Test Environment 1, BIR Measurement, Source 1, Position R3

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Left Channel Parameters

Freq. [Hz]		31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]		19.208	30.829	53.936	62.259	63.405	63.689	65.225	62.842	59.605	54.037	70.799	71.405
Noise [dB]		2.436	14.009	16.412	20.93	15.93	20.916	22.694	15.406	1.721	7.873	26.892	27.919
strenGth [dB]		-49.792	-38.171	-15.064	-6.741	-5.595	-5.311	-3.775	-6.158	-9.395	-14.963	-6.201	-5.595
C50 [dB]		-5.087	-17.948	-40.689	-1.#IO	-1.#IO	-55.501	-51.742	-1.#IO	-1.#IO	-1.#IO	-55.873	-57.086
C80 [dB]		-4.126	-9.143	-9.03	-8.777	-4.929	-4.246	-3.042	-3.344	-4.831	1.859	-3.702	-4.391
D50 [%]		23.663	1.578	0.009	0	0	0	0.001	0	0	0	0	0
Ts [ms]		181.45	167.901	173.123	146.046	125.411	121.272	121.493	115.428	110.303	84.583	119.642	123.049
EDT [s]		1.271	1.169	1.155	1.142	0.829	0.832	0.926	0.829	0.596	0.413	0.866	0.909
T20 [s]	--		0.622	1.006	1.097	0.907	1.056	0.983	0.874	0.639	0.472	0.966	0.982
r T20	--		0.98	0.959	0.992	0.992	0.996	0.999	0.999	0.999	0.997	0.999	0.999
T30 [s]		0	0.566	0.987	1.052	0.936	1.051	1.045	0.894	0.66	0.483	1.007	1.01
r T30		101	0.974	0.988	0.997	0.997	0.998	0.998	0.999	0.999	0.999	0.999	0.999
Tuser [s]	--		0.54	1.068	1.112	0.903	1.053	1.042	0.89	0.655	0.486	1.001	1.02
r Tuser	--		0.991	0.973	0.993	0.993	0.996	0.998	0.999	0.999	0.998	0.999	0.999
IACC (Early)		0.978	0.959	0.971	0.763	0.601	0.53	0.255	0.385	0.226	0.543	0.354	0.398
Tau IACC (ms)		-0.363	0	-0.159	0.068	-0.159	0.023	0.045	0	0.045	0.045	0.023	0.023
w IACC (ms)		0	1.406	0.884	0.522	0.295	0.159	0.091	0.068	0.045	0.045	0.091	0.091

RTU = RT User (-10. dB, -30. dB)

Right Channel Parameters

Freq. [Hz]		31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]		17.153	29.283	52.086	62.024	62.454	62.895	64.284	62.799	59.23	52.674	70.197	70.773
Noise [dB]		0.922	13.449	16.514	20.946	15.843	21.296	22.428	14.731	0.767	6.99	26.759	27.753
strenGth [dB]		-51.847	-39.717	-16.914	-6.976	-6.546	-6.105	-4.716	-6.201	-9.77	-16.326	-6.803	-6.227
C50 [dB]		-3.691	-18.474	-39.927	-1.#IO	-1.#IO	-1.#IO	-72.006	-1.#IO	-1.#IO	-1.#IO	-1.#IO	-1.#IO
C80 [dB]		-2.261	-9.22	-9.827	-7.452	-5.505	-3.37	-3.026	-2.232	-4.631	0.827	-3.158	-3.975
D50 [%]		29.946	1.401	0.01	0	0	0	0	0	0	0	0	0
Ts [ms]		291.224	239.697	265.065	154.196	156.12	145.617	150.918	116.57	120.246	115.783	137.413	142.335
EDT [s]		1.305	1.261	1.282	1.009	0.851	0.94	0.884	0.816	0.587	0.402	0.859	0.885
T20 [s]	--		0.574	0.985	1.084	0.886	1.032	1.041	0.887	0.665	0.487	0.988	0.997
r T20	--		0.975	0.963	0.997	0.993	0.992	0.998	0.998	0.999	0.998	0.998	0.999
T30 [s]	--		0.539	0.95	1.048	0.91	1.036	1.042	0.929	0.673	0.499	1.002	1.001
r T30	--		0.977	0.989	0.999	0.996	0.997	0.999	0.998	0.999	0.999	0.999	1
Tuser [s]	--		0.474	0.945	1.036	0.872	1.07	1.057	0.946	0.663	0.485	1.019	1.009
r Tuser	--		0.989	0.973	0.997	0.993	0.996	0.999	0.998	0.999	0.998	1	1
IACC (Early)		0.978	0.959	0.971	0.763	0.601	0.53	0.255	0.385	0.226	0.543	0.354	0.398
Tau IACC (ms)		-0.363	0	-0.159	0.068	-0.159	0.023	0.045	0	0.045	0.045	0.023	0.023
w IACC (ms)		0	1.406	0.884	0.522	0.295	0.159	0.091	0.068	0.045	0.045	0.091	0.091

RTU = RT User (-10. dB, -30. dB)

Table A.4: *ISO 3383 Octave band acoustical parameters: Test environment 1, source position 1, receiver position R3*

Test Environment 1, BIR Measurement, Source 2, Position P

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Left Channel Parameters

Freq. [Hz]		31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]		26.101	43.017	60.367	69.073	72.101	72.635	72.657	71.757	72.843	66.615	79.727	80.184
Noise [dB]		6.777	18.028	11.909	14.161	12.853	17.793	23.377	13.802	0.504	8.337	26.225	26.639
strenGth [dB]		-42.899	-25.983	-8.633	0.073	3.101	3.635	3.657	2.757	3.843	-2.385	2.727	3.184
C50 [dB]		1.643	5.488	4.297	6.522	8.933	8.997	7.328	7.64	11.696	12.587	8.365	8.301
C80 [dB]		5.437	8.019	5.043	8.291	11.134	11.377	9.205	9.657	15.419	16.495	10.588	10.47
D50 [%]		59.349	77.965	72.899	81.785	88.664	88.811	84.386	85.311	93.661	94.776	87.282	87.116
Ts [ms]		105.751	68.573	67.262	54.279	41.327	38.439	43.696	40.67	31.563	30.108	38.863	39.289
EDT [s]		0.744	0.62	1.005	0.755	0.278	0.245	0.607	0.57	0.182	0.116	0.383	0.397
T20 [s]	--		1.011	0.982	1.095	0.997	0.959	1.016	0.877	0.571	0.435	0.943	0.983
r T20	--		0.995	0.989	0.983	0.992	0.994	0.998	0.995	0.965	0.97	0.996	0.997
T30 [s]		0	1.329	1.038	1.086	0.991	1.038	1.066	0.926	0.664	0.456	1	1.018
r T30		101	0.963	0.994	0.992	0.997	0.996	0.998	0.997	0.984	0.992	0.998	0.998
Tuser [s]	--		1.102	1.027	1.079	0.971	1.052	1.057	0.897	0.671	0.466	0.987	1.002
r Tuser	--		0.989	0.988	0.988	0.996	0.997	0.999	0.997	0.996	0.994	0.999	0.999
IACC (Early)		0.99	0.971	0.883	0.807	0.708	0.437	0.52	0.326	0.234	0.093	0.442	0.504
Tau IACC (ms)		0.113	0.884	0.59	0.771	0.726	0.794	0.771	0.771	0.771	0.522	0.771	0.771
w IACC (ms)		0	1.814	0.907	0.612	0.317	0.159	0.091	0.045	0.068	0.045	0.091	0.091

RTU = RT User (-10. dB, -30. dB)

Right Channel Parameters

Freq. [Hz]		31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]		23.29	41.433	58.018	67.943	68.143	67.796	67.983	64.396	62.906	55.218	74.047	75.271
Noise [dB]		6.918	16.934	9.472	17.02	13.268	14.524	21.021	11.991	-0.551	7.065	24.011	25.149
strenGth [dB]		-45.71	-27.567	-10.982	-1.057	-0.857	-1.204	-1.017	-4.604	-6.094	-13.782	-2.953	-1.729
C50 [dB]		0.686	5.412	0.815	5.421	3.464	2.767	0.648	-1.817	-0.703	-0.518	1.21	2.271
C80 [dB]		4.706	6.397	3.31	7.63	6.905	7.608	3.792	2.525	5.944	6.314	5.033	5.751
D50 [%]		53.941	77.663	54.676	77.699	68.945	65.412	53.722	39.689	45.962	47.024	56.92	62.783
Ts [ms]		202.017	98.75	115.519	70.412	102.802	117.126	128.196	221.492	311.075	415.251	143.727	121.765
EDT [s]		0.764	0.782	0.803	0.935	0.765	0.617	0.921	0.899	0.52	0.458	0.837	0.834
T20 [s]	--		0.989	0.985	1.042	1.081	1.023	0.974	0.93	0.688	0.495	0.971	0.993
r T20	--		0.995	0.99	0.991	0.995	0.996	0.998	0.999	0.997	0.996	0.999	0.999
T30 [s]	--		1.13	1.079	1.053	1.007	1.063	1.055	0.962	0.699	0.505	1.017	1.02
r T30	--		0.986	0.994	0.996	0.994	0.998	0.996	0.999	0.999	0.998	0.999	0.999
Tuser [s]	--		1.046	1.088	1.054	1.069	1.078	1.045	0.959	0.703	0.494	1.017	1.025
r Tuser	--		0.991	0.987	0.99	0.993	0.997	0.998	0.999	0.997	0.995	0.999	0.999
IACC (Early)		0.99	0.971	0.883	0.807	0.708	0.437	0.52	0.326	0.234	0.093	0.442	0.504
Tau IACC (ms)		0.113	0.884	0.59	0.771	0.726	0.794	0.771	0.771	0.771	0.522	0.771	0.771
w IACC (ms)		0	1.814	0.907	0.612	0.317	0.159	0.091	0.045	0.068	0.045	0.091	0.091

RTU = RT User (-10. dB, -30. dB)

Table A.5: *ISO-3382 acoustical parameters: Test environment 1, source position 2, receiver position P*

Test Environment 1, BIR Measurement, Source 2, Position R1

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Left Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	21.283	41.132	55.557	63.049	63.8	64.559	65.254	64.761	61.928	56.066	71.757	72.328
Noise [dB]	-2.205	4.547	8.197	13.367	17.384	18.524	16.93	10.888	-0.025	7.697	23.031	23.892
strenGth [dB]	-47.717	-27.868	-13.443	-5.951	-5.2	-4.441	-3.746	-4.239	-7.072	-12.934	-5.243	-4.672
C50 [dB]	-0.622	0.048	-0.34	-2.646	-1.41	-1.579	-3.026	-1.453	-2.438	1.619	-2.046	-2.046
C80 [dB]	4.458	4.21	1.708	1.679	2.565	2.692	0.803	2.833	3.054	6.906	2.157	2.187
D50 [%]	46.428	50.278	48.044	35.225	41.956	41.01	33.252	41.713	36.326	59.213	38.434	38.433
Ts [ms]	119.243	89.235	104.518	102.339	87.772	85.025	95.754	79.722	76.655	56.521	85.726	86.863
EDT [s]	0.766	0.55	1.179	1.208	0.92	0.914	1.026	0.813	0.679	0.441	0.911	0.941
T20 [s]	1.158	1.205	1.035	1.033	1.012	0.975	0.976	0.916	0.647	0.518	0.963	0.979
r T20	0.968	0.958	0.994	0.992	0.994	0.999	0.997	0.996	0.999	0.998	1	0.999
T30 [s]	1.24	1.378	0.971	1.04	1.034	1	1.029	0.968	0.676	0.5	1.001	1.005
r T30	0.971	0.979	0.995	0.997	0.998	0.999	0.997	0.997	0.999	0.999	0.999	0.999
Tuser [s]	1.376	1.404	0.941	1.061	1.059	0.982	0.986	0.967	0.668	0.493	0.982	0.996
r Tuser	0.967	0.972	0.987	0.995	0.998	0.999	0.996	0.997	0.998	0.997	0.999	0.999
IACC (Early)	0.99	0.982	0.958	0.663	0.482	0.225	0.231	0.143	0.191	0.201	0.201	0.282
Tau IACC (ms)	-0.59	-0.272	0.045	-0.159	-0.249	-0.181	-0.159	-0.181	-0.091	-0.113	-0.204	-0.204
w IACC (ms)	0.454	1.655	0.975	0.567	0.34	0.159	0.091	0.068	0.045	0.045	0.068	0.068

RTU = RT User (-10. dB, -30. dB)

Right Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	19.479	39.433	55.084	64.256	63.855	64.947	66.518	65.973	63.596	57.475	72.773	73.21
Noise [dB]	-7.096	4.969	8.137	12.891	13.51	15.704	17.615	10.855	-0.885	6.736	21.957	22.347
strenGth [dB]	-49.521	-29.567	-13.916	-4.744	-5.145	-4.053	-2.482	-3.027	-5.404	-11.525	-4.227	-3.79
C50 [dB]	-1.19	1.753	-1.601	-6.594	-1.584	1.555	0.787	2.41	2.586	5.55	1.109	-0.026
C80 [dB]	4.055	5.479	0.01	2.143	1.542	3.706	3.72	5.565	5.892	9.301	4.255	3.641
D50 [%]	43.19	59.958	40.885	17.972	40.98	58.855	54.516	63.526	64.464	78.211	56.352	49.851
Ts [ms]	180.654	131.971	116.546	77.503	86.679	77.759	71.58	60.31	52.216	40.862	67.846	70.898
EDT [s]	0.796	0.566	1.161	1.03	0.902	0.932	0.907	0.79	0.605	0.369	0.857	0.878
T20 [s]	1.309	1.18	0.914	1.087	0.984	0.967	1.031	0.88	0.675	0.511	0.964	0.988
r T20	0.96	0.975	0.99	0.99	0.995	0.998	0.999	1	0.999	0.997	0.999	0.999
T30 [s]	1.76	1.356	0.914	1.027	1.038	1.001	1.053	0.941	0.691	0.52	1.005	1.009
r T30	0.945	0.985	0.995	0.996	0.997	0.997	0.999	0.998	0.999	0.999	0.999	0.999
Tuser [s]	1.568	1.338	0.841	1.056	1.067	0.961	1.049	0.92	0.688	0.506	1.002	1.008
r Tuser	0.964	0.98	0.995	0.987	0.996	0.997	0.999	0.997	0.999	0.998	0.999	0.999
IACC (Early)	0.99	0.982	0.958	0.663	0.482	0.225	0.231	0.143	0.191	0.201	0.201	0.282
Tau IACC (ms)	-0.59	-0.272	0.045	-0.159	-0.249	-0.181	-0.159	-0.181	-0.091	-0.113	-0.204	-0.204
w IACC (ms)	0.454	1.655	0.975	0.567	0.34	0.159	0.091	0.068	0.045	0.045	0.068	0.068

RTU = RT User (-10. dB, -30. dB)

Table A.6: *ISO 3383 Octave band acoustical parameters: Test environment 1, source position 2, receiver position R1*

Test Environment 1, BIR Measurement, Source 2, Position R2

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Left Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	18.683	36.667	54.558	63.983	63.527	63.962	65.303	64.74	61.808	57.93	71.679	72.321
Noise [dB]	1.929	5.39	11.373	18.51	19.753	22.084	23.536	18.378	2.507	7.46	28.276	28.668
strenGth [dB]	-50.317	-32.333	-14.442	-5.017	-5.473	-5.038	-3.697	-4.26	-7.192	-11.07	-5.321	-4.679
C50 [dB]	-9.234	-3.142	-6.068	-12.458	-7.893	-7.074	-4.873	-2.988	-5.643	0.876	-5.027	-5.991
C80 [dB]	-2.112	0.335	1.79	-3.208	-1.759	-1.18	0.788	2.399	1.357	7.399	0.588	-0.093
D50 [%]	10.657	32.665	19.826	5.373	13.974	16.4	24.564	33.448	21.428	55.028	23.913	20.109
Ts [ms]	167.097	116.402	109.991	122.2	110.425	103.333	98.837	85.801	84.196	57.409	95.169	99.378
EDT [s]	1.166	0.764	0.863	0.962	0.852	0.827	0.958	0.817	0.594	0.372	0.862	0.882
T20 [s]	1.378	1.391	0.944	0.922	0.955	1.022	1.014	0.94	0.709	0.473	0.967	0.953
r T20	0.986	0.977	0.954	0.995	0.989	0.997	0.999	0.999	0.999	0.999	1	0.999
T30 [s]	1.338	1.221	0.912	0.971	1	1.055	1.049	0.961	0.71	0.495	1.009	0.997
r T30	0.982	0.983	0.985	0.997	0.995	0.998	0.999	1	1	0.999	0.999	0.999
Tuser [s]	1.407	1.252	0.83	0.981	1.025	1.063	1.021	0.968	0.711	0.496	1	0.994
r Tuser	0.975	0.958	0.981	0.995	0.997	0.998	0.999	0.999	0.999	0.999	0.999	0.999
IACC (Early)	0.978	0.966	0.985	0.904	0.51	0.501	0.576	0.539	0.457	0.547	0.54	0.584
Tau IACC (ms)	-0.091	-0.091	0.227	0.045	0.113	-0.023	0	0.023	0.023	0.023	0.023	0.023
w IACC (ms)	0	1.701	0.93	0.612	0.272	0.159	0.091	0.045	0.045	0.045	0.068	0.068

RTU = RT User (-10. dB, -30. dB)

Right Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	17.773	35.038	54.698	62.567	62.119	63.116	64.529	63.411	60.061	56.224	70.58	71.157
Noise [dB]	1.144	8.012	10.248	17.751	17.408	19.116	21.725	16.523	1.186	6.747	26.11	26.582
strenGth [dB]	-51.227	-33.962	-14.302	-6.433	-6.881	-5.884	-4.471	-5.589	-8.939	-12.776	-6.42	-5.843
C50 [dB]	-9.897	-2.141	-7.139	-10.869	-9.901	-6.236	-6.107	-2.863	-5.526	1.216	-5.319	-6.204
C80 [dB]	-2.542	0.205	3.026	-1.062	-2.878	-0.066	0.404	2.362	1.599	8.473	0.625	0.083
D50 [%]	9.289	37.92	16.195	7.568	9.281	19.218	19.684	34.094	21.886	56.957	22.712	19.335
Ts [ms]	206.076	169.385	106.49	169.31	152.702	125.576	118.107	116.527	125.891	85.023	122.583	129.944
EDT [s]	1.113	0.84	0.919	1.036	0.986	0.827	0.893	0.798	0.7	0.329	0.862	0.903
T20 [s]	1.47	1.324	0.942	0.888	0.926	1.033	1.047	0.924	0.677	0.488	0.975	0.952
r T20	0.983	0.968	0.981	0.997	0.996	0.996	0.997	0.999	0.998	0.999	0.998	0.999
T30 [s]	1.357	1.309	0.969	0.995	0.969	1.103	1.054	0.952	0.7	0.497	1.023	1.008
r T30	0.978	0.971	0.993	0.993	0.995	0.997	0.999	0.999	0.999	0.999	0.999	0.998
Tuser [s]	1.505	1.252	0.9	1.014	0.909	1.122	1.098	0.96	0.712	0.489	1.04	1.007
r Tuser	0.97	0.966	0.988	0.983	0.996	0.997	0.998	0.999	0.999	0.999	0.999	0.999
IACC (Early)	0.978	0.966	0.985	0.904	0.51	0.501	0.576	0.539	0.457	0.547	0.54	0.584
Tau IACC (ms)	-0.091	-0.091	0.227	0.045	0.113	-0.023	0	0.023	0.023	0.023	0.023	0.023
w IACC (ms)	0	1.701	0.93	0.612	0.272	0.159	0.091	0.045	0.045	0.045	0.068	0.068

RTU = RT User (-10. dB, -30. dB)

Table A.7: *ISO 3383 Octave band acoustical parameters: Test environment 1, source position 2, receiver position R2*

Test Environment 1, BIR Measurement, Source 2, Position R3

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS
Left Channel Parameters

Freq. [Hz]		31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]		17.879	34.146	52.997	63.38	61.566	61.258	63.797	62.322	59.661	53.74	69.672	70.56
Noise [dB]		16.443	13.575	16.045	18.582	18.797	21.738	22.989	17.379	1.208	7.345	27.635	28.565
strenGth [dB]		-51.121	-34.854	-16.003	-5.62	-7.434	-7.742	-5.203	-6.678	-9.339	-15.26	-7.328	-6.44
C50 [dB]		-12.649	-20.595	-32.135	-1.#IO	-1.#IO	-61.577	-1.#IO	-1.#IO	-1.#IO	-67.483	-1.#IO	-1.#IO
C80 [dB]		-11.361	-6.32	-3.193	-5.807	-6.339	-4.34	-5.088	-4.176	-4.911	0.109	-4.786	-4.939
D50 [%]		5.154	0.864	0.061	0	0	0	0	0	0	0	0	0
Ts [ms]		230.208	171.688	146.877	129.771	134.944	130.874	126.586	120.353	111.867	89.186	123.995	124.709
EDT [s]		1.294	1.042	1.23	0.838	0.94	1.042	0.877	0.827	0.625	0.448	0.869	0.877
T20 [s]	--		1.265	1.043	1.08	0.951	1.049	1.058	0.898	0.63	0.454	0.994	1.004
r T20	--		0.965	0.972	0.995	0.997	0.994	0.998	0.999	0.999	0.996	0.999	0.999
T30 [s]		0	1.168	0.998	1.065	0.948	1.057	1.051	0.935	0.66	0.47	1.013	1.014
r T30		101	0.972	0.99	0.997	0.999	0.998	0.999	0.999	0.999	0.998	0.999	1
Tuser [s]	--		1.275	0.927	1.061	0.933	1.095	1.067	0.938	0.654	0.453	1.029	1.024
r Tuser	--		0.948	0.986	0.993	0.998	0.998	0.999	0.999	0.998	0.996	1	0.999
IACC (Early)		0.961	0.953	0.948	0.954	0.726	0.784	0.356	0.494	0.338	0.327	0.458	0.604
Tau IACC (ms)		0.091	-0.272	0.113	-0.136	0	-0.091	-0.68	-0.068	-0.045	-0.136	-0.068	-0.068
w IACC (ms)		0	1.633	0.862	0.544	0.317	0.159	0.113	0.045	0.045	0.045	0.091	0.091

RTU = RT User (-10. dB, -30. dB)

Right Channel Parameters

Freq. [Hz]		31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]		15.071	34.008	53.799	62.363	62.006	61.26	63.404	62.318	59.499	53.97	69.507	70.367
Noise [dB]		14.584	13.515	15.227	18.739	16.135	21.611	22.871	16.411	0.96	6.934	27.184	27.968
strenGth [dB]		-53.929	-34.992	-15.201	-6.637	-6.994	-7.74	-5.596	-6.682	-9.501	-15.03	-7.493	-6.633
C50 [dB]		-14.055	-18.575	-33.139	-1.#IO	-1.#IO	-1.#IO	-1.#IO	-1.#IO	-1.#IO	-1.#IO	-1.#IO	-1.#IO
C80 [dB]		-11.84	-7.343	1.686	-3.151	-5.758	-3.617	-4.58	-5.528	-3.706	3.156	-4.535	-3.872
D50 [%]		3.783	1.369	0.049	0	0	0	0	0	0	0	0	0
Ts [ms]		439.487	177.221	122.111	163.979	121.952	130.79	138.583	120.468	116.113	84.581	128.778	130.359
EDT [s]		1.11	0.965	1.563	0.907	0.806	1.025	0.915	0.835	0.59	0.416	0.87	0.879
T20 [s]	--		1.123	1.053	1.107	0.952	1.029	1.054	0.878	0.68	0.493	0.985	1.002
r T20	--		0.965	0.972	0.993	0.998	0.997	0.999	0.996	0.999	0.993	0.998	0.999
T30 [s]	--		1.153	1.06	1.073	0.986	1.055	1.052	0.924	0.698	0.495	1.014	1.022
r T30	--		0.969	0.989	0.996	0.998	0.999	0.999	0.998	0.999	0.998	0.999	1
Tuser [s]	--		1.211	0.998	1.047	0.953	1.083	1.062	0.942	0.698	0.486	1.028	1.027
r Tuser	--		0.949	0.978	0.99	0.998	0.998	0.998	0.997	0.999	0.994	0.999	0.999
IACC (Early)		0.961	0.953	0.948	0.954	0.726	0.784	0.356	0.494	0.338	0.327	0.458	0.604
Tau IACC (ms)		0.091	-0.272	0.113	-0.136	0	-0.091	-0.68	-0.068	-0.045	-0.136	-0.068	-0.068
w IACC (ms)		0	1.633	0.862	0.544	0.317	0.159	0.113	0.045	0.045	0.045	0.091	0.091

RTU = RT User (-10. dB, -30. dB)

Table A.8: *ISO 3383 Octave band acoustical parameters: Test environment 1, source position 2, receiver position R3*

A.1.2 Test Environment 1: EASE Simulations

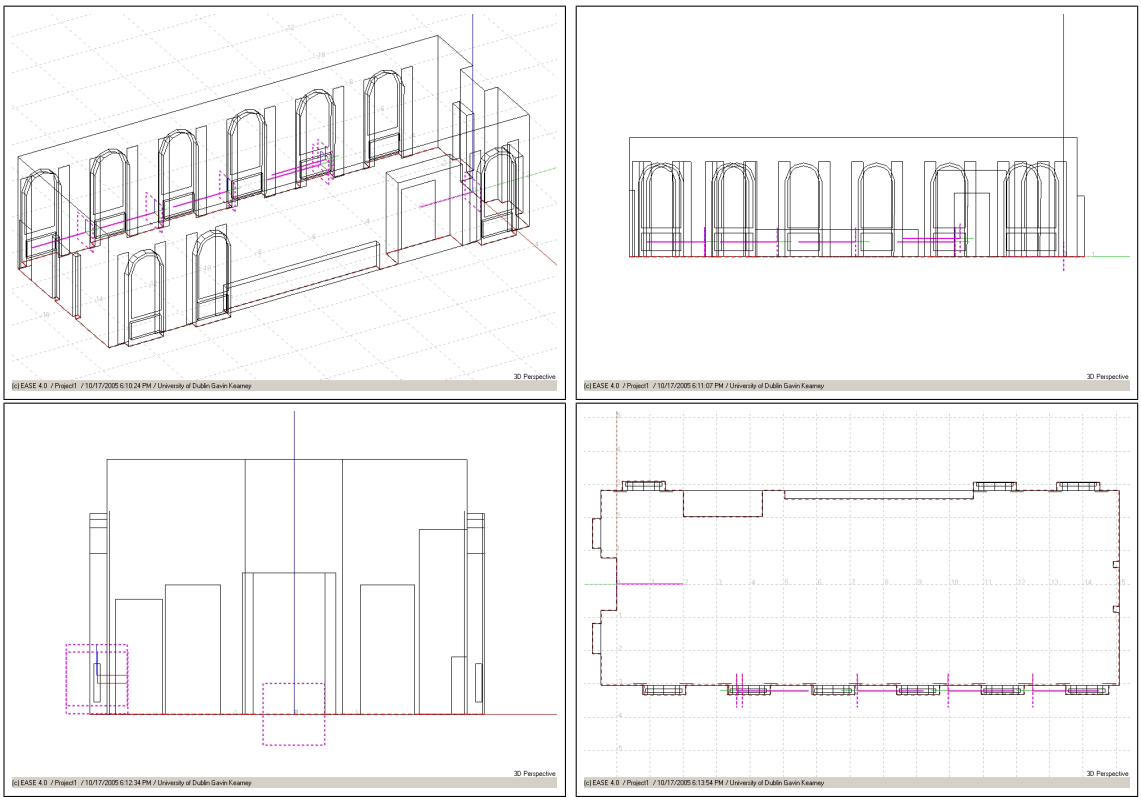


Figure A.3: *Acoustic model of test environment 1.*

Test Environment 1, BIR Simulation: Source 1, Position P

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Left Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	36.085	44.294	54.346	60.205	61.996	61.73	61.14	62.374	59.139	53.572	68.844	69.588
Noise [dB]	-3.34	-9.683	-6.725	-11.212	-8.966	-10.174	-8.276	-7.68	-10.862	-11.079	-1.386	-0.396
strenGth [dB]	-32.915	-24.706	-14.654	-8.795	-7.004	-7.27	-7.86	-6.626	-9.861	-15.428	-8.156	-7.412
C50 [dB]	0.758	0.051	2.057	5.37	2.217	3.335	2.026	1.74	-0.417	-3.139	1.981	2.239
C80 [dB]	1.447	1.23	9.299	8.658	5.838	8.072	6.092	7.015	7.982	3.782	6.903	6.992
D50 [%]	54.353	50.296	61.623	77.494	62.49	68.309	61.458	59.882	47.6	32.679	61.211	62.61
Ts [ms]	120.417	103.837	70.507	58.651	65.561	59.055	65.504	62.298	61.715	78.892	61.801	60.992
EDT [s]	1.006	1.009	0.333	0.444	0.757	0.499	0.691	0.544	0.366	0.726	0.579	0.59
T20 [s]	1.177	1.161	1.248	1.243	1.178	1.194	1.094	0.872	0.635	0.719	1.041	1.092
r T20	0.956	0.968	0.985	0.992	0.995	0.989	0.998	0.995	0.997	0.988	0.995	0.995
T30 [s]	1.235	1.243	1.338	1.272	1.181	1.199	1.076	0.875	0.647	0.705	1.065	1.128
r T30	0.984	0.979	0.993	0.996	0.997	0.996	0.999	0.998	0.998	0.996	0.998	0.998
Tuser [s]	1.239	1.147	1.348	1.323	1.224	1.215	1.077	0.911	0.63	0.749	1.076	1.137
r Tuser	0.974	0.973	0.99	0.995	0.997	0.995	0.997	0.998	0.998	0.994	0.998	0.998
IACC (Early)	0.993	0.983	0.996	0.91	0.793	0.585	0.389	0.31	0.16	0.069	0.325	0.373
Tau IACC (ms)	-0.59	-0.113	0	-0.227	-0.476	-0.544	-0.59	-0.635	-0.816	-0.658	-0.635	-0.635
w IACC (ms)	0	1.746	1.088	0.544	0.295	0.159	0.091	0.045	0.045	0.045	0.091	0.091

RTU = RT User (-10. dB, -30. dB)

Right Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	35.456	43.475	54.078	60.713	63.914	64.654	65.041	67.697	68.066	63.302	73.767	73.971
Noise [dB]	-4.607	-11.291	-8.043	-12.429	-21.95	-43.293	-57.616	-91.485	-92.181	-81.469	-61.797	-15.547
strenGth [dB]	-33.544	-25.525	-14.922	-8.287	-5.086	-4.346	-3.959	-1.303	-0.934	-5.698	-3.233	-3.029
C50 [dB]	-0.121	-2.179	1.476	6.695	5.157	7.465	6.734	7.69	12.259	11.228	7.917	7.641
C80 [dB]	0.71	-0.733	9.057	8.884	9.206	11.155	10.301	12.251	17.275	15.384	12.169	11.821
D50 [%]	49.302	37.714	58.414	82.371	76.628	84.8	82.499	85.453	94.39	92.992	86.091	85.313
Ts [ms]	139.167	125.396	74.99	52.18	42.155	30.119	26.679	18.286	7.9	8.395	19.895	22.234
EDT [s]	0.99	1.041	0.338	0.437	0.505	0.318	0.337	0.29	0.181	0.195	0.276	0.305
T20 [s]	1.234	1.162	1.267	1.241	1.245	1.08	1.068	0.805	0.522	0.664	0.951	1.004
r T20	0.955	0.973	0.987	0.995	0.989	0.985	0.992	0.985	0.919	0.956	0.984	0.984
T30 [s]	1.302	1.257	1.339	1.263	1.234	1.117	1.062	0.817	0.593	0.711	1.006	1.082
r T30	0.984	0.984	0.994	0.996	0.996	0.996	0.997	0.995	0.971	0.985	0.995	0.994
Tuser [s]	1.308	1.195	1.348	1.294	1.297	1.136	1.086	0.829	0.632	0.763	1.025	1.094
r Tuser	0.977	0.976	0.992	0.996	0.998	0.995	0.998	0.997	0.988	0.992	0.998	0.999
IACC (Early)	0.993	0.983	0.996	0.91	0.793	0.585	0.389	0.31	0.16	0.069	0.325	0.373
Tau IACC (ms)	-0.59	-0.113	0	-0.227	-0.476	-0.544	-0.59	-0.635	-0.816	-0.658	-0.635	-0.635
w IACC (ms)	0	1.746	1.088	0.544	0.295	0.159	0.091	0.045	0.045	0.045	0.091	0.091

RTU = RT User (-10. dB, -30. dB)

Table A.9: *ISO-3382 acoustical parameters: Test Environment 1, BIR simulation, source position 1, receiver position P*

Test Environment 1, BIR Simulation: Source 1, Position R1

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Left Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	39.34	45.848	57.297	60.857	64.777	64.631	65.571	68.791	66.397	63.024	73.974	74.179
Noise [dB]	3.363	1.855	10.644	13.272	13.29	12.818	15.847	17.925	20.074	25.479	25.53	26.885
strenGth [dB]	-29.66	-23.152	-11.703	-8.143	-4.223	-4.369	-3.429	-0.209	-2.603	-5.976	-3.026	-2.821
C50 [dB]	-7.55	-2.241	1.543	-2.349	3.095	5.902	5.083	6.586	6.611	8.366	5.812	4.996
C80 [dB]	3.501	3.509	6.36	2.445	6.022	9.242	8.419	11.76	12.753	12.654	10.085	9.026
D50 [%]	14.951	37.378	58.793	36.799	67.098	79.559	76.32	82.001	82.086	87.284	79.22	75.96
Ts [ms]	129.562	106.938	77.726	89.88	62.693	52.29	51.771	44.084	43.842	42.13	47.437	50.414
EDT [s]	0.632	0.86	0.548	0.815	0.708	0.412	0.523	0.312	0.261	0.218	0.352	0.452
T20 [s]	0.9	1.223	1.246	1.292	0.975	0.914	0.833	0.663	0.507	0.632	0.779	0.869
r T20	0.927	0.976	0.959	0.99	0.988	0.995	0.997	0.983	0.981	0.992	0.992	0.989
T30 [s]	1.195	1.247	1.264	1.181	1.057	0.972	0.904	0.728	0.537	0.643	0.889	1.001
r T30	0.939	0.985	0.982	0.993	0.991	0.996	0.996	0.992	0.994	0.998	0.992	0.99
Tuser [s]	1.351	1.263	1.458	1.235	1.148	0.995	0.91	0.751	0.557	0.651	0.879	1.03
r Tuser	0.915	0.974	0.99	0.99	0.997	0.995	0.998	0.997	0.996	0.997	0.996	0.991
IACC (Early)	0.997	0.993	0.999	0.913	0.766	0.441	0.384	0.312	0.161	0.214	0.332	0.403
Tau IACC (ms)	0.204	0.023	0	0.023	0.249	0.204	0.204	0.227	0.703	0.227	0.227	0.227
w IACC (ms)	0	1.973	0.975	0.522	0.317	0.159	0.068	0.068	0.045	0.045	0.068	0.068

RTU = RT User (-10. dB, -30. dB)

Right Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	39.434	45.538	57.201	61.074	63.421	61.846	63.14	65.305	62.971	59.194	71.051	71.666
Noise [dB]	0.382	3.076	9.139	15.459	14.703	13.156	16.216	19.011	20.277	25.746	26.102	27.458
strenGth [dB]	-29.566	-23.462	-11.799	-7.926	-5.579	-7.154	-5.86	-3.695	-6.029	-9.806	-5.949	-5.334
C50 [dB]	-7.772	-3.229	1.292	-3.734	-0.398	0.803	0.451	1.192	0.28	2.225	0.565	0.261
C80 [dB]	3.653	2.697	6.203	1.892	3.629	5.609	5.166	7.782	7.825	7.582	6.316	5.557
D50 [%]	14.312	32.225	57.385	29.737	47.711	54.612	52.595	56.821	51.613	62.537	53.245	51.503
Ts [ms]	126.768	114.86	79.467	85.508	85.661	99.29	90.606	98.36	96.49	101.778	93	89.937
EDT [s]	0.631	0.886	0.561	0.805	0.713	0.571	0.696	0.401	0.364	0.452	0.539	0.592
T20 [s]	0.884	1.196	1.239	1.251	1.073	0.986	0.862	0.715	0.527	0.61	0.857	0.958
r T20	0.934	0.974	0.959	0.992	0.989	0.996	0.997	0.996	0.996	0.995	0.995	0.991
T30 [s]	1.221	1.293	1.264	1.166	1.044	1.011	0.926	0.762	0.53	0.649	0.931	1.022
r T30	0.938	0.982	0.982	0.995	0.993	0.998	0.997	0.997	0.998	0.996	0.996	0.995
Tuser [s]	1.345	1.342	1.454	1.226	1.152	1.047	0.935	0.751	0.518	0.648	0.943	1.08
r Tuser	0.911	0.967	0.99	0.992	0.996	0.998	0.997	0.997	0.997	0.994	0.997	0.998
IACC (Early)	0.997	0.993	0.999	0.913	0.766	0.441	0.384	0.312	0.161	0.214	0.332	0.403
Tau IACC (ms)	0.204	0.023	0	0.023	0.249	0.204	0.204	0.227	0.703	0.227	0.227	0.227
w IACC (ms)	0	1.973	0.975	0.522	0.317	0.159	0.068	0.068	0.045	0.045	0.068	0.068

RTU = RT User (-10. dB, -30. dB)

Table A.10: *ISO 3383 Octave band acoustical parameters: Test Environment 1, BIR simulation, source position 1, receiver position R1*

Test Environment 1, BIR Simulation: Source 1, Position R2

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Left Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	45.745	54.18	60.43	68.38	70.164	67.459	69.217	71.208	68.686	65.685	77.04	77.761
Noise [dB]	5.718	9.752	21.72	15.178	11.301	11.685	12.539	13.127	14.015	19.122	21.122	23.105
strenGth [dB]	-23.255	-14.82	-8.57	-0.62	1.164	-1.541	0.217	2.208	-0.314	-3.315	0.04	0.761
C50 [dB]	1.179	-5.338	0.309	0.25	-0.351	-1.593	-0.16	1.929	1.745	2.349	0.685	0.488
C80 [dB]	4.301	-0.445	4.464	4.358	3.998	2.591	3.006	6.121	6.838	5.748	4.496	4.342
D50 [%]	56.743	22.632	51.778	51.441	47.979	40.932	49.078	60.923	59.912	63.2	53.936	52.808
Ts [ms]	114.32	110.766	87.882	75.491	79.12	82.612	78.734	61.956	56.394	59.709	68.953	69.792
EDT [s]	0.861	0.614	0.946	0.57	0.844	0.939	0.813	0.534	0.474	0.607	0.66	0.671
T20 [s]	0.984	1.502	1.409	1.2	1.093	1.1	1.005	0.85	0.609	0.694	0.969	1.025
r T20	0.945	0.955	0.991	0.992	0.997	0.996	0.999	0.998	0.999	0.996	0.999	0.998
T30 [s]	1.297	1.469	1.328	1.275	1.147	1.151	1.032	0.834	0.607	0.685	1.017	1.094
r T30	0.954	0.979	0.992	0.996	0.998	0.998	0.999	0.999	0.999	0.998	0.998	0.997
Tuser [s]	1.238	1.555	1.403	1.279	1.157	1.12	1.009	0.839	0.619	0.693	0.989	1.074
r Tuser	0.978	0.983	0.99	0.995	0.996	0.996	0.999	0.998	0.999	0.995	0.999	0.998
IACC (Early)	1	0.999	0.995	0.917	0.443	0.32	0.252	0.121	0.194	0.132	0.157	0.27
Tau IACC (ms)	0.136	0	0	-0.091	-0.159	-0.227	-0.181	0.884	-0.975	-0.227	-0.227	-0.227
w IACC (ms)	0	0.998	0.907	0.567	0.317	0.136	0.091	0.068	0.045	0.045	0.068	0.068

RTU = RT User (-10. dB, -30. dB)

Right Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	45.856	53.94	60.649	68.508	69.808	68.885	70.909	73.637	70.727	68.061	78.854	79.19
Noise [dB]	5.287	11.739	21.414	14.599	13.309	10.74	13.048	13.046	14.026	18.539	21.058	22.906
strenGth [dB]	-23.144	-15.06	-8.351	-0.492	0.808	-0.115	1.909	4.637	1.727	-0.939	1.854	2.19
C50 [dB]	1.547	-5.607	0.73	-1.336	-4.948	0.319	0.987	3.797	2.725	4.572	1.968	1.174
C80 [dB]	4.637	-1.01	4.288	4.455	2.346	3.509	4.75	7.804	8.295	8.393	5.948	5.417
D50 [%]	58.812	21.569	54.194	42.367	24.243	51.835	55.658	70.565	65.189	74.131	61.136	56.717
Ts [ms]	111.434	117.06	83.563	73.29	85.876	59.488	53.332	35.418	35.244	34.547	45.414	50.227
EDT [s]	0.839	0.632	0.946	0.515	0.708	0.784	0.666	0.423	0.376	0.392	0.501	0.521
T20 [s]	0.997	1.508	1.396	1.215	1.127	1.166	0.981	0.764	0.553	0.666	0.923	0.993
r T20	0.941	0.96	0.988	0.988	0.998	0.996	0.995	0.995	0.995	0.997	0.995	0.993
T30 [s]	1.265	1.461	1.336	1.27	1.193	1.159	1.022	0.786	0.574	0.679	1.01	1.087
r T30	0.957	0.98	0.99	0.996	0.998	0.999	0.998	0.998	0.998	0.999	0.995	0.995
Tuser [s]	1.198	1.538	1.437	1.284	1.208	1.187	0.998	0.791	0.586	0.68	0.998	1.08
r Tuser	0.982	0.983	0.989	0.996	0.995	0.998	0.997	0.998	0.997	0.999	0.998	0.998
IACC (Early)	1	0.999	0.995	0.917	0.443	0.32	0.252	0.121	0.194	0.132	0.157	0.27
Tau IACC (ms)	0.136	0	0	-0.091	-0.159	-0.227	-0.181	0.884	-0.975	-0.227	-0.227	-0.227
w IACC (ms)	0	0.998	0.907	0.567	0.317	0.136	0.091	0.068	0.045	0.045	0.068	0.068

RTU = RT User (-10. dB, -30. dB)

Table A.11: *ISO 3383 Octave band acoustical parameters: Test Environment 1, BIR simulation, source position 1, receiver position R2*

Test Environment 1, BIR Simulation: Source 1, Position R3

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Left Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	37.558	46.365	55.562	63.918	64.357	62.299	64.378	66.117	62.359	60.31	71.736	72.507
Noise [dB]	8.699	14.517	19.313	18.968	18.656	16.767	18.223	18.872	21.046	25.82	27.169	29.197
strenGth [dB]	-31.442	-22.635	-13.438	-5.082	-4.643	-6.701	-4.622	-2.883	-6.641	-8.69	-5.264	-4.493
C50 [dB]	-13.48	-12.433	-30.668	-	-	-	-	-	-	-	-	-
C80 [dB]	-11.439	-1.973	-2.783	-3.73	-4.07	-4.391	-1.318	-1.047	-2.816	0.472	-2.09	-2.513
D50 [%]	4.295	5.402	0.086	0	0	0	0	0	0	-0.002	0	0
Ts [ms]	185.251	141.111	123.482	110.056	116.843	113.032	104.856	95.935	97.916	91.194	102.192	103.867
EDT [s]	0.978	1.207	0.729	0.427	0.673	0.608	0.592	0.444	0.453	0.458	0.517	0.514
T20 [s]	1.251	1.267	1.176	1.185	1.191	1.096	0.999	0.795	0.58	0.644	0.961	1.038
r T20	0.976	0.987	0.991	0.97	0.998	0.996	0.998	0.995	0.994	0.997	0.996	0.995
T30 [s]	1.235	1.153	1.333	1.176	1.089	1.077	1.004	0.807	0.584	0.669	0.989	1.047
r T30	0.97	0.99	0.991	0.991	0.994	0.998	0.999	0.999	0.998	0.998	0.998	0.998
Tuser [s]	1.321	1.184	1.265	1.244	1.119	1.1	1.024	0.822	0.607	0.675	1.016	1.076
r Tuser	0.955	0.97	0.987	0.995	0.994	0.998	0.999	0.999	0.998	0.998	0.999	0.999
IACC (Early)	0.956	1	0.999	0.982	0.827	0.765	0.436	0.357	0.404	0.524	0.454	0.564
Tau IACC (ms)	-0.181	0.091	0	-0.023	0	0	0.023	0	0	0	0	0
w IACC (ms)	0	1.769	0.862	0.544	0.34	0.136	0.091	0.068	0.045	0.045	0.068	0.068

RTU = RT User (-10. dB, -30. dB)

Right Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	37.771	46.464	55.519	64.177	65.609	62.706	64.583	66.906	63.136	60.595	72.359	73.137
Noise [dB]	12.568	15.205	19.372	17.834	19.368	14.523	18.641	19.408	20.497	25.944	27.138	29.156
strenGth [dB]	-31.229	-22.536	-13.481	-4.823	-3.391	-6.294	-4.417	-2.094	-5.864	-8.405	-4.641	-3.863
C50 [dB]	-12.857	-12.749	-30.533	-	-	-	-	-	-	-	-	-
C80 [dB]	-10.221	-2.123	-3.231	-3.856	-7.131	-2.977	-2.485	-1.47	-1.752	0.275	-2.611	-3.226
D50 [%]	4.924	5.042	0.088	0	0	0	0	0	0	-0.002	0	0
Ts [ms]	176.402	137.933	124.731	103.672	87.587	102.943	100.026	79.999	81.883	85.406	88.543	89.835
EDT [s]	0.997	1.143	0.736	0.389	0.493	0.698	0.68	0.379	0.429	0.417	0.495	0.481
T20 [s]	1.109	1.262	1.172	1.182	1.125	1.013	0.955	0.784	0.541	0.655	0.921	0.988
r T20	0.98	0.99	0.992	0.972	0.998	0.997	0.997	0.993	0.997	0.998	0.997	0.996
T30 [s]	1.144	1.145	1.336	1.173	1.104	1.057	0.946	0.801	0.567	0.66	0.958	1.025
r T30	0.98	0.99	0.991	0.992	0.999	0.998	0.999	0.998	0.997	0.999	0.998	0.998
Tuser [s]	1.161	1.148	1.27	1.239	1.129	1.081	0.909	0.814	0.574	0.668	0.947	1.023
r Tuser	0.96	0.966	0.986	0.994	0.999	0.996	0.998	0.998	0.994	0.998	0.999	0.999
IACC (Early)	0.956	1	0.999	0.982	0.827	0.765	0.436	0.357	0.404	0.524	0.454	0.564
Tau IACC (ms)	-0.181	0.091	0	-0.023	0	0	0.023	0	0	0	0	0
w IACC (ms)	0	1.769	0.862	0.544	0.34	0.136	0.091	0.068	0.045	0.045	0.068	0.068

RTU = RT User (-10. dB, -30. dB)

Table A.12: *ISO 3383 Octave band acoustical parameters: Test Environment 1, BIR simulation, source position 1, receiver position R3*

Test Environment 1, BIR Simulation: Source 2, Position P

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Left Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	39.368	46.938	55.637	63.058	65.345	64.938	67.026	69.407	69.339	64.625	75.269	75.473
Noise [dB]	2.416	-0.054	5.783	4.074	2.796	-0.988	0.02	0.581	-3.963	-3.601	7.1	8.332
strenGth [dB]	-29.632	-22.062	-13.363	-5.942	-3.655	-4.062	-1.974	0.407	0.339	-4.375	-1.731	-1.527
C50 [dB]	0.09	4.062	-2.466	6.357	5.716	5.65	7.081	8.695	11.508	10.494	8.234	7.773
C80 [dB]	0.677	6.154	5.658	9.818	8.648	9.568	11.032	12.901	16.369	14.363	12.218	11.724
D50 [%]	50.52	71.815	36.176	81.209	78.855	78.6	83.625	88.102	93.399	91.807	86.943	85.691
Ts [ms]	121.689	82.228	89.41	54.529	53.346	48.736	44.922	40.078	34.921	36.083	40.88	42.131
EDT [s]	1.097	0.702	0.684	0.368	0.56	0.48	0.339	0.264	0.124	0.108	0.326	0.348
T20 [s]	0.881	1.054	1.04	1.175	1.228	1.087	1.027	0.796	0.495	0.686	0.925	0.993
r T20	0.962	0.965	0.963	0.986	0.992	0.986	0.981	0.973	0.951	0.976	0.973	0.976
T30 [s]	0.85	1.115	1.11	1.164	1.184	1.143	1.036	0.831	0.571	0.716	1.003	1.048
r T30	0.988	0.987	0.984	0.995	0.996	0.994	0.995	0.992	0.98	0.992	0.991	0.993
Tuser [s]	0.889	1.183	0.995	1.152	1.246	1.191	1.09	0.88	0.573	0.718	1.04	1.089
r Tuser	0.978	0.984	0.976	0.991	0.997	0.997	0.998	0.995	0.979	0.989	0.997	0.998
IACC (Early)	0.998	0.992	0.996	0.939	0.774	0.56	0.222	0.269	0.121	0.052	0.278	0.345
Tau IACC (ms)	0.431	-0.136	0.023	0.295	0.454	0.567	0.635	0.612	0.816	0.93	0.635	0.635
w IACC (ms)	0	0	1.066	0.522	0.295	0.181	0.091	0.045	0.045	0.045	0.068	0.091

RTU = RT User (-10. dB, -30. dB)

Right Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	40.007	47.394	55.652	61.887	63.653	62.22	63.299	63.985	61.852	56.753	70.557	71.323
Noise [dB]	2.565	-2.904	4.048	-1.945	-2.75	-13.881	-14.148	-12.597	-12.942	-11.329	-6.345	-2.852
strenGth [dB]	-28.993	-21.606	-13.348	-7.113	-5.347	-6.78	-5.701	-5.015	-7.148	-12.247	-6.443	-5.677
C50 [dB]	1.246	4.635	-2.312	4.844	3.086	1.774	-1.272	0.054	-0.841	-0.738	0.276	1.034
C80 [dB]	1.807	6.639	5.868	8.707	6.567	6.4	6.318	7.154	9.045	6.562	6.923	7.132
D50 [%]	57.125	74.406	36.995	75.313	67.055	60.073	42.73	50.308	45.174	45.764	51.587	55.927
Ts [ms]	105.042	74.033	89.107	71.397	78.759	91.137	105.966	139.657	195.813	221.037	120.999	109.531
EDT [s]	1.189	0.699	0.637	0.504	0.717	0.659	0.61	0.5	0.278	0.516	0.555	0.567
T20 [s]	0.843	1.065	1.039	1.214	1.219	1.172	1.07	0.85	0.642	0.724	1.024	1.072
r T20	0.956	0.957	0.965	0.987	0.997	0.995	0.997	0.996	0.993	0.989	0.996	0.997
T30 [s]	0.82	1.107	1.122	1.171	1.175	1.14	1.06	0.868	0.622	0.726	1.039	1.078
r T30	0.984	0.988	0.984	0.996	0.998	0.998	0.999	0.998	0.998	0.997	0.999	0.999
Tuser [s]	0.885	1.21	1.013	1.151	1.201	1.159	1.094	0.891	0.631	0.754	1.071	1.095
r Tuser	0.977	0.983	0.976	0.992	0.997	0.996	0.998	0.999	0.997	0.996	0.999	0.999
IACC (Early)	0.998	0.992	0.996	0.939	0.774	0.56	0.222	0.269	0.121	0.052	0.278	0.345
Tau IACC (ms)	0.431	-0.136	0.023	0.295	0.454	0.567	0.635	0.612	0.816	0.93	0.635	0.635
w IACC (ms)	0	0	1.066	0.522	0.295	0.181	0.091	0.045	0.045	0.045	0.068	0.091

RTU = RT User (-10. dB, -30. dB)

Table A.13: *ISO-3382 acoustical parameters: Test Environment 1, BIR simulation, source position 2, receiver position P*

Test Environment 1, BIR Simulation: Source 2, Position R1

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Left Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	38.267	44.284	55.521	58.287	62	60.638	61.932	64.1	62.01	59.652	69.874	70.358
Noise [dB]	-1.417	3.591	4.261	4.64	10.583	10.243	11.843	13.782	14.008	19.355	20.791	21.859
strenGth [dB]	-30.733	-24.716	-13.479	-10.713	-7	-8.362	-7.068	-4.9	-6.99	-9.348	-7.126	-6.642
C50 [dB]	-3.72	-1.535	1.208	-2.001	-0.028	2.676	1.504	2.992	3.028	4.316	2.374	1.954
C80 [dB]	5.253	2.859	6.41	2.304	3.695	5.926	5.423	8.012	8.143	8.539	6.658	6.039
D50 [%]	29.804	41.254	56.911	38.68	49.839	64.936	58.573	66.572	66.759	72.984	63.338	61.064
Ts [ms]	123.368	106.887	73.858	87.608	71.329	62.606	64.552	54.439	53.585	51.512	58.442	60.729
EDT [s]	0.635	0.884	0.57	0.801	0.707	0.571	0.697	0.4	0.357	0.376	0.531	0.576
T20 [s]	0.885	1.189	1.233	1.243	1.069	0.986	0.862	0.715	0.526	0.595	0.85	0.936
r T20	0.935	0.976	0.958	0.991	0.989	0.996	0.997	0.996	0.996	0.994	0.996	0.991
T30 [s]	1.28	1.319	1.254	1.157	1.044	1.01	0.925	0.762	0.53	0.638	0.924	1.004
r T30	0.942	0.984	0.984	0.994	0.993	0.998	0.997	0.997	0.998	0.996	0.996	0.995
Tuser [s]	1.425	1.369	1.436	1.221	1.15	1.046	0.935	0.751	0.519	0.623	0.93	1.056
r Tuser	0.905	0.969	0.99	0.991	0.996	0.999	0.997	0.997	0.997	0.995	0.998	0.998
IACC (Early)	0.997	0.993	0.999	0.912	0.76	0.437	0.38	0.309	0.158	0.242	0.327	0.387
Tau IACC (ms)	0.091	0	0	-0.023	-0.249	-0.204	-0.204	-0.227	-0.703	-0.227	-0.227	-0.227
w IACC (ms)	0	1.995	0.998	0.522	0.317	0.159	0.068	0.068	0.045	0.045	0.068	0.068

RTU = RT User (-10. dB, -30. dB)

Right Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	38.126	44.591	55.621	58.124	63.389	63.419	64.363	67.586	65.372	62.88	72.8	72.941
Noise [dB]	-4.22	1.508	1.639	9.095	10.601	9.095	11.133	12.766	13.187	18.597	20.017	21.267
strenGth [dB]	-30.874	-24.409	-13.379	-10.876	-5.611	-5.581	-4.637	-1.414	-3.628	-6.12	-4.2	-4.059
C50 [dB]	-3.603	-0.584	1.472	-0.741	3.807	6.758	5.895	7.517	7.728	9.043	6.773	6.105
C80 [dB]	4.988	3.65	6.589	2.901	6.133	9.437	8.628	11.989	13.145	12.899	10.372	9.533
D50 [%]	30.374	46.644	58.393	45.746	70.613	82.58	79.532	84.952	85.564	88.917	82.629	80.308
Ts [ms]	127.443	99.58	72.173	90.949	51.802	32.994	36.883	24.395	24.709	24.493	29.796	33.509
EDT [s]	0.638	0.858	0.559	0.803	0.708	0.411	0.523	0.312	0.261	0.213	0.346	0.412
T20 [s]	0.913	1.218	1.243	1.294	0.97	0.914	0.833	0.663	0.507	0.62	0.775	0.843
r T20	0.923	0.976	0.958	0.988	0.988	0.995	0.997	0.983	0.981	0.992	0.992	0.99
T30 [s]	1.217	1.257	1.265	1.181	1.055	0.972	0.904	0.728	0.535	0.63	0.88	0.98
r T30	0.949	0.984	0.983	0.992	0.991	0.996	0.996	0.992	0.994	0.998	0.992	0.99
Tuser [s]	1.362	1.29	1.457	1.247	1.147	0.995	0.91	0.752	0.554	0.633	0.87	0.999
r Tuser	0.917	0.972	0.99	0.989	0.996	0.995	0.998	0.997	0.996	0.997	0.997	0.992
IACC (Early)	0.997	0.993	0.999	0.912	0.76	0.437	0.38	0.309	0.158	0.242	0.327	0.387
Tau IACC (ms)	0.091	0	0	-0.023	-0.249	-0.204	-0.204	-0.227	-0.703	-0.227	-0.227	-0.227
w IACC (ms)	0	1.995	0.998	0.522	0.317	0.159	0.068	0.068	0.045	0.045	0.068	0.068

RTU = RT User (-10. dB, -30. dB)

Table A.14: *ISO 3383 Octave band acoustical parameters: Test Environment 1, BIR simulation, source position 2, receiver position R1*

Test Environment 1, BIR Simulation: Source 2, Position R2

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Left Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	40.756	45.815	57.538	62.571	63.673	60.777	63.973	65.982	61.856	60.544	71.354	72.019
Noise [dB]	2.025	12.437	20.602	21.557	24.226	21.329	20.451	19.325	21.973	26.938	29.315	31.448
strenGth [dB]	-28.244	-23.185	-11.462	-6.429	-5.327	-8.223	-5.027	-3.018	-7.144	-8.456	-5.646	-4.981
C50 [dB]	-1.885	1.428	-1.026	-5.536	-9.298	-4.819	-1.165	-3.366	-3.284	-0.56	-3.112	-3.73
C80 [dB]	4.135	3.103	5.542	2.609	0.861	1.533	3.632	5.191	4.31	6.024	3.852	3.4
D50 [%]	39.314	58.15	44.122	21.846	10.518	24.796	43.335	31.539	31.949	46.78	32.817	29.758
Ts [ms]	119.722	105.979	84.384	88.877	95.374	95.535	78.925	70.911	70.998	64.309	76.764	78.905
EDT [s]	0.598	1.104	0.654	0.603	0.71	0.964	0.814	0.435	0.475	0.451	0.601	0.609
T20 [s]	1.106	1.32	1.188	1.356	1.153	1.127	0.99	0.824	0.608	0.756	0.975	1.059
r T20	0.954	0.984	0.992	0.988	0.989	0.997	0.998	0.995	0.997	0.993	0.997	0.996
T30 [s]	1.213	1.258	1.217	1.302	1.18	1.128	1.025	0.814	0.598	0.7	1.026	1.114
r T30	0.974	0.99	0.997	0.993	0.995	0.999	0.999	0.998	0.999	0.995	0.998	0.997
Tuser [s]	1.374	1.24	1.226	1.393	1.203	1.106	0.998	0.823	0.613	0.746	1.002	1.106
r Tuser	0.978	0.986	0.992	0.995	0.989	0.998	0.998	0.997	0.999	0.994	0.999	0.999
IACC (Early)	0.999	0.998	0.999	0.963	0.758	0.55	0.45	0.346	0.277	0.406	0.374	0.495
Tau IACC (ms)	-0.113	0.068	0	-0.023	0	0.023	-0.045	0.249	0	0	0	0
w IACC (ms)	0	1.837	0.952	0.544	0.317	0.159	0.068	0.045	0.045	0.045	0.068	0.068

RTU = RT User (-10. dB, -30. dB)

Right Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	40.782	45.766	57.813	62.276	64.369	62.541	63.375	66.821	62.529	62.013	71.881	72.526
Noise [dB]	6.195	12.531	20.866	21.83	23.838	20.67	19.779	19.38	21.81	27.456	29.27	31.549
strenGth [dB]	-28.218	-23.234	-11.187	-6.724	-4.631	-6.459	-5.625	-2.179	-6.471	-6.987	-5.119	-4.474
C50 [dB]	-1.867	1.811	-1.107	-4.537	-8.022	-2.057	-4.07	-0.481	-2.501	1.797	-2.141	-2.749
C80 [dB]	3.984	3.429	5.859	2.561	0.192	3.212	1.728	4.896	4.654	6.96	3.567	3.216
D50 [%]	39.413	60.276	43.663	26.026	13.621	38.375	28.146	47.236	35.987	60.201	37.921	34.684
Ts [ms]	119.013	107.195	79.202	95.109	81.259	63.645	90.583	58.454	60.808	45.846	67.993	70.208
EDT [s]	0.609	1.04	0.6	0.729	0.73	0.754	0.859	0.488	0.467	0.481	0.636	0.644
T20 [s]	1.131	1.32	1.171	1.344	1.129	1.161	0.986	0.78	0.602	0.683	0.962	1.043
r T20	0.948	0.986	0.992	0.992	0.995	0.997	0.999	0.993	0.997	0.997	0.997	0.997
T30 [s]	1.193	1.276	1.206	1.275	1.135	1.166	1.028	0.798	0.585	0.69	1.023	1.101
r T30	0.972	0.992	0.996	0.993	0.998	0.999	0.999	0.998	0.998	0.999	0.997	0.998
Tuser [s]	1.389	1.264	1.222	1.382	1.142	1.184	1.015	0.8	0.563	0.695	1.008	1.108
r Tuser	0.981	0.986	0.993	0.996	0.995	0.998	0.998	0.997	0.997	0.998	0.999	0.999
IACC (Early)	0.999	0.998	0.999	0.963	0.758	0.55	0.45	0.346	0.277	0.406	0.374	0.495
Tau IACC (ms)	-0.113	0.068	0	-0.023	0	0.023	-0.045	0.249	0	0	0	0
w IACC (ms)	0	1.837	0.952	0.544	0.317	0.159	0.068	0.045	0.045	0.045	0.068	0.068

RTU = RT User (-10. dB, -30. dB)

Table A.15: *ISO 3383 Octave band acoustical parameters: Test Environment 1, BIR simulation, source position 2, receiver position R2*

Test Environment 1, BIR Simulation: Source 2, Position R3

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Left Channel Parameters

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	43.497	47.886	57.873	67.372	66.684	64.255	66.865	68.181	64.082	61.623	73.936	74.828
Noise [dB]	10.23	18.385	26.433	32.121	26.285	21.229	21.805	20.129	20.513	25.717	30.772	35.232
strenGth [dB]	-25.503	-21.114	-11.127	-1.628	-2.316	-4.745	-2.135	-0.819	-4.918	-7.377	-3.064	-2.172
C50 [dB]	-8.924	-15.591	-42.425	-	-	-	-	-	-	-	-	-
C80 [dB]	1.042	-3.862	-4.042	-4.964	-11.233	-5.356	-5.814	-1.947	-4.486	-1.253	-4.204	-4.976
D50 [%]	11.357	2.686	0.006	-0.002	0	0	0	0	0	-0.001	0	-0.001
Ts [ms]	145.1	166.628	144.734	125.121	128.052	126.55	121.168	104.629	109.57	104.344	114.555	117.145
EDT [s]	0.717	1.059	1.348	0.823	0.814	1.011	0.879	0.692	0.664	0.746	0.786	0.799
T20 [s]	1.733	1.464	1.336	1.081	1.232	1.072	0.949	0.821	0.577	0.702	0.961	1.045
r T20	0.921	0.97	0.989	0.992	0.994	0.998	0.998	0.999	0.995	0.996	0.999	0.999
T30 [s]	1.439	1.287	1.296	1.167	1.139	1.081	0.961	0.819	0.586	0.669	0.985	1.066
r T30	0.953	0.983	0.995	0.994	0.995	0.999	0.999	0.999	0.998	0.997	0.999	1
Tuser [s]	1.882	1.304	1.279	1.126	1.129	1.093	0.988	0.81	0.577	0.673	0.995	1.066
r Tuser	0.926	0.965	0.993	0.988	0.995	0.998	0.998	0.999	0.995	0.997	0.999	0.999
IACC (Early)	1	0.999	0.999	0.994	0.692	0.63	0.632	0.492	0.254	0.357	0.518	0.576
Tau IACC (ms)	0.159	0.136	0.068	-0.045	0.045	-0.136	-0.136	0.522	0.068	-0.136	-0.113	-0.113
w IACC (ms)	0	1.746	0.952	0.544	0.363	0.136	0.091	0.068	0.045	0.045	0.068	0.068

RTU = RT User (-10. dB, -30. dB)

Right Channel Parameters

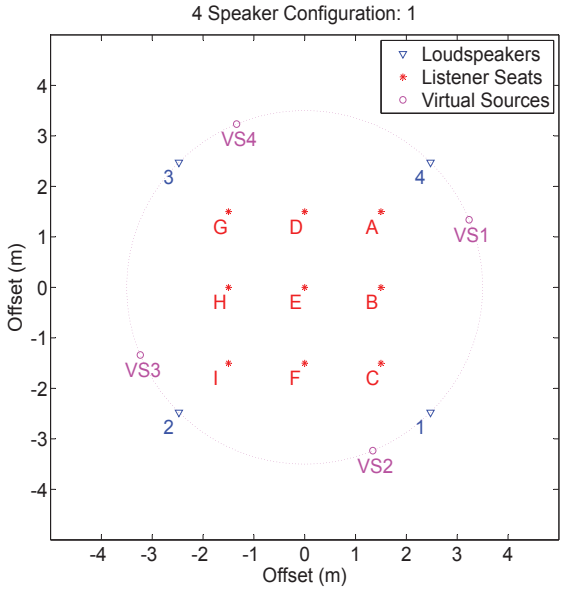
Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB]	43.44	47.95	57.702	67.493	67.269	65.244	67.961	70.141	65.485	64.668	75.366	75.932
Noise [dB]	10.713	17.859	26.094	33.079	27.7	23.638	22.793	19.526	20.812	26.09	31.441	35.705
strenGth [dB]	-25.56	-21.05	-11.298	-1.507	-1.731	-3.756	-1.039	1.141	-3.515	-4.332	-1.634	-1.068
C50 [dB]	-8.888	-15.094	-39.9	-	-	-	-	-	-	-	-1.#IO	-1.#IO
C80 [dB]	0.985	-4.14	-3.508	-3.5	-9.645	-3.544	-6.309	-0.435	-2.744	-0.125	-2.873	-3.508
D50 [%]	11.441	3.002	0.01	-0.002	-0.001	0	0	0	0	-0.001	0	0
Ts [ms]	147.04	164.161	150.561	121.67	111.935	100.775	94.13	66.623	79.321	51.757	82.407	90.84
EDT [s]	0.731	1.024	1.405	0.871	1.014	0.838	0.689	0.508	0.537	0.405	0.614	0.688
T20 [s]	1.696	1.505	1.318	1.117	1.149	1.106	0.98	0.802	0.598	0.705	0.95	1.036
r T20	0.94	0.973	0.988	0.995	0.995	0.998	0.996	0.998	0.999	0.99	0.998	0.999
T30 [s]	1.376	1.327	1.285	1.151	1.121	1.088	0.975	0.792	0.593	0.682	0.981	1.062
r T30	0.95	0.984	0.995	0.998	0.998	0.999	0.999	0.999	0.999	0.996	0.999	0.999
Tuser [s]	1.796	1.343	1.265	1.152	1.113	1.117	0.961	0.786	0.61	0.704	0.972	1.047
r Tuser	0.937	0.968	0.992	0.994	0.998	0.998	0.998	0.999	0.999	0.993	0.999	0.999
IACC (Early)	1	0.999	0.999	0.994	0.692	0.63	0.632	0.492	0.254	0.357	0.518	0.576
Tau IACC (ms)	0.159	0.136	0.068	-0.045	0.045	-0.136	-0.136	0.522	0.068	-0.136	-0.113	-0.113
w IACC (ms)	0	1.746	0.952	0.544	0.363	0.136	0.091	0.068	0.045	0.045	0.068	0.068

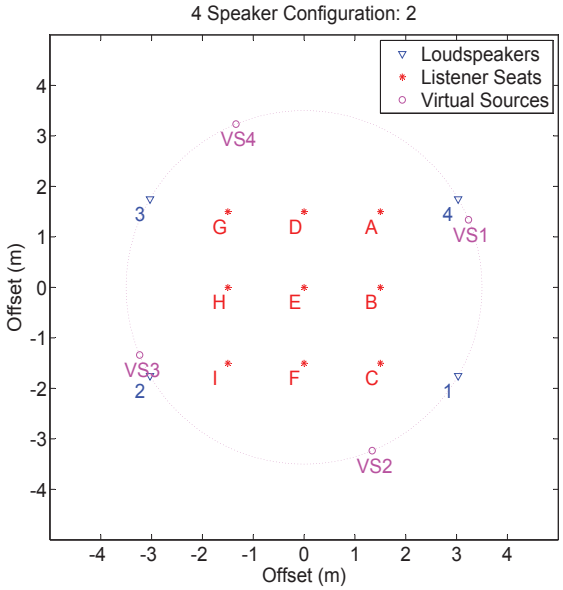
RTU = RT User (-10. dB, -30. dB)

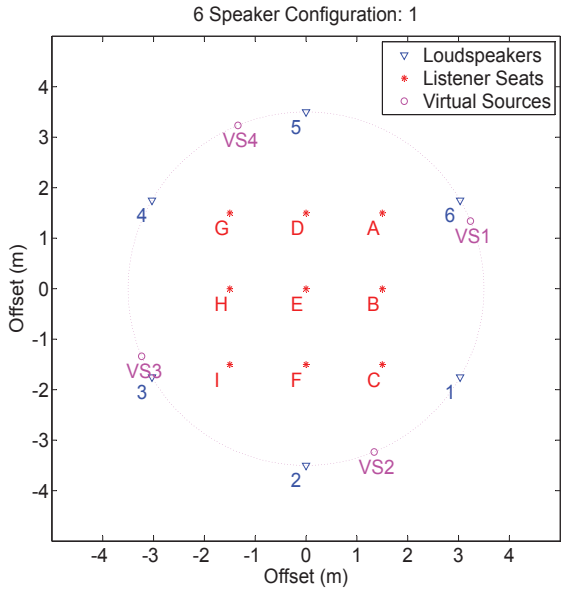
Table A.16: *ISO 3383 Octave band acoustical parameters: Test Environment 1, BIR simulation, source position 2, receiver position R3*

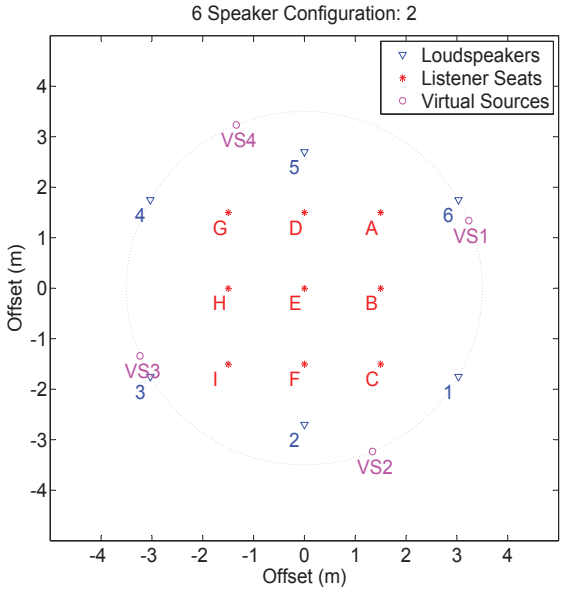
B

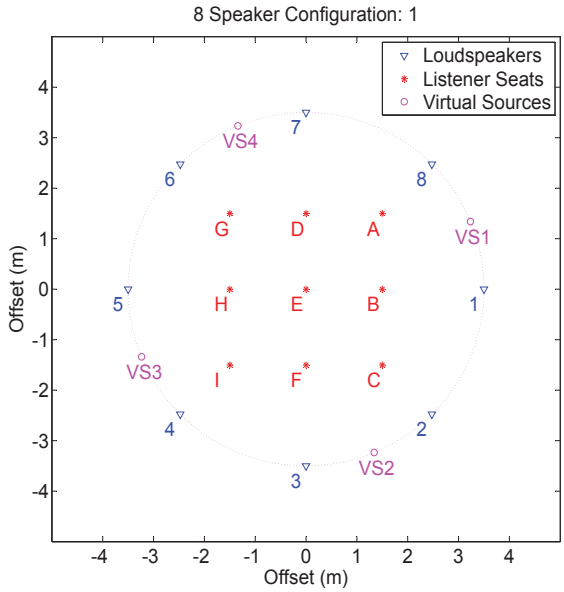
Loudspeaker Layouts

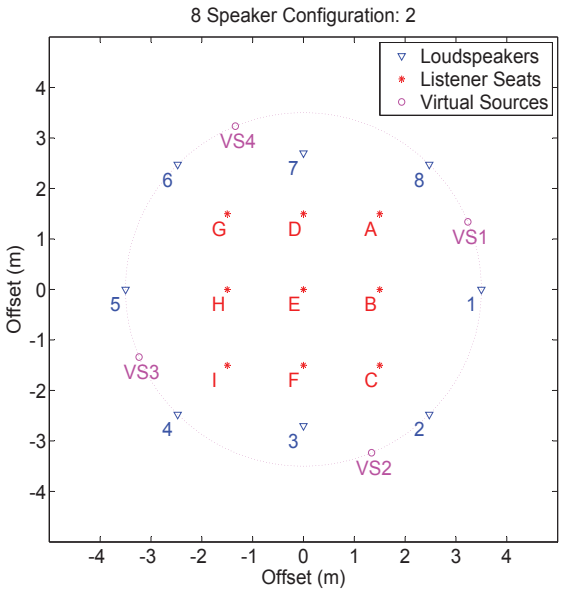

4 Speaker Configuration: 1
Loudspeakers
Listener Seats
Virtual Sources
VS4
VS1
VS3
VS2
G
D
A
H
E
B
I
F
C
1
2
3
4
Offset (m)
Offset (m)


4 Speaker Configuration: 2
Loudspeakers
Listener Seats
Virtual Sources
VS4
VS1
VS3
VS2
G
D
A
H
E
B
I
F
C
1
2
3
4
Offset (m)
Offset (m)


6 Speaker Configuration: 1
Loudspeakers
Listener Seats
Virtual Sources
VS4
VS1
VS3
VS2
G
D
A
H
E
B
I
F
C
1
2
3
4
5
6
Offset (m)
Offset (m)


6 Speaker Configuration: 2
Loudspeakers
Listener Seats
Virtual Sources
VS4
VS1
VS3
VS2
G
D
A
H
E
B
I
F
C
1
2
3
4
5
6
Offset (m)
Offset (m)


8 Speaker Configuration: 1
Loudspeakers
Listener Seats
Virtual Sources
VS4
VS1
VS3
VS2
G
D
A
H
E
B
I
F
C
1
2
3
4
5
6
7
8
Offset (m)
Offset (m)


8 Speaker Configuration: 2
Loudspeakers
Listener Seats
Virtual Sources
VS4
VS1
VS3
VS2
G
D
A
H
E
B
I
F
C
1
2
3
4
5
6
7
8
Offset (m)
Offset (m)

C

VBAP Localization Vectors

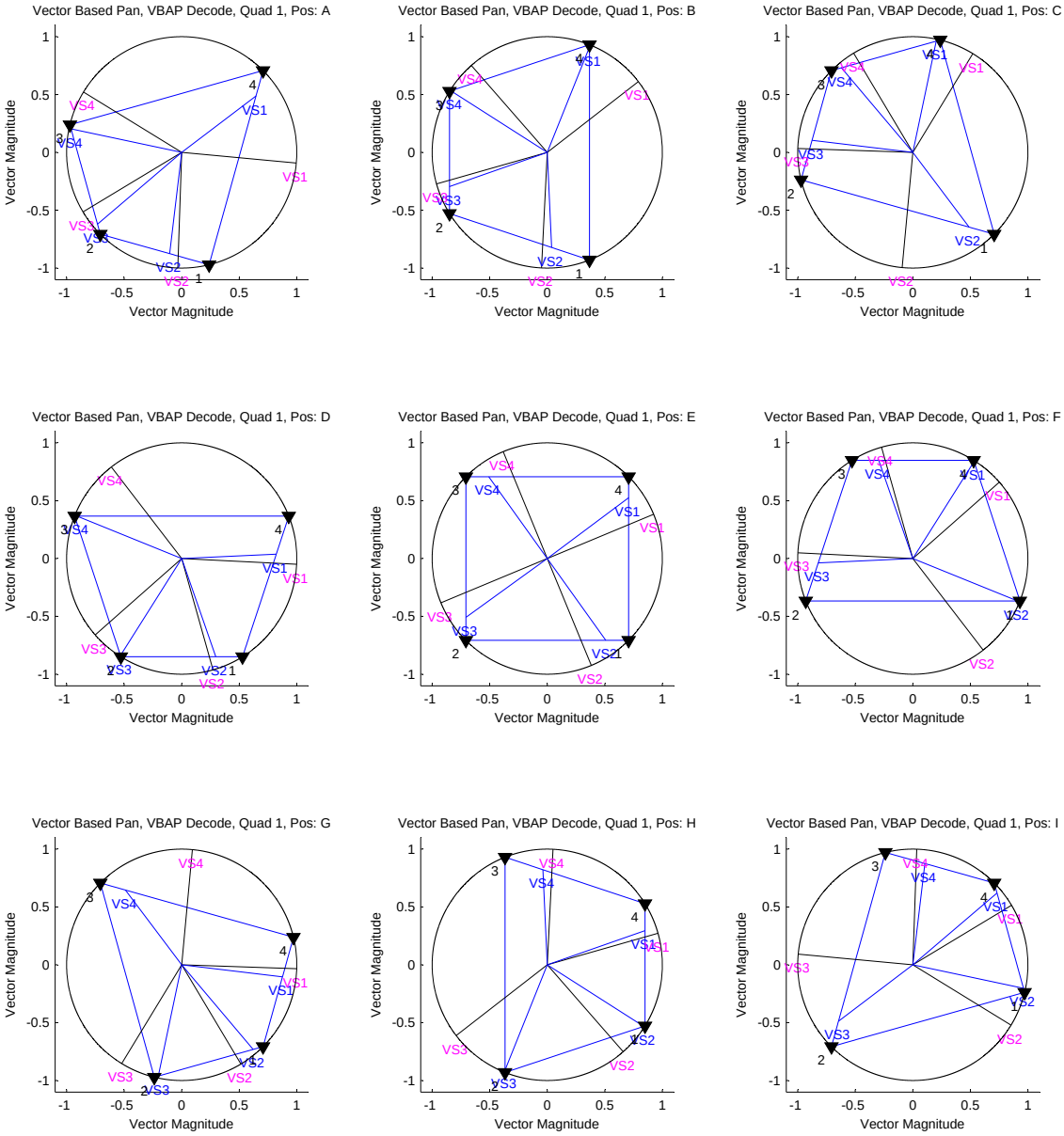


Figure C.1: *VBAP energy vectors: Quad setup 1*

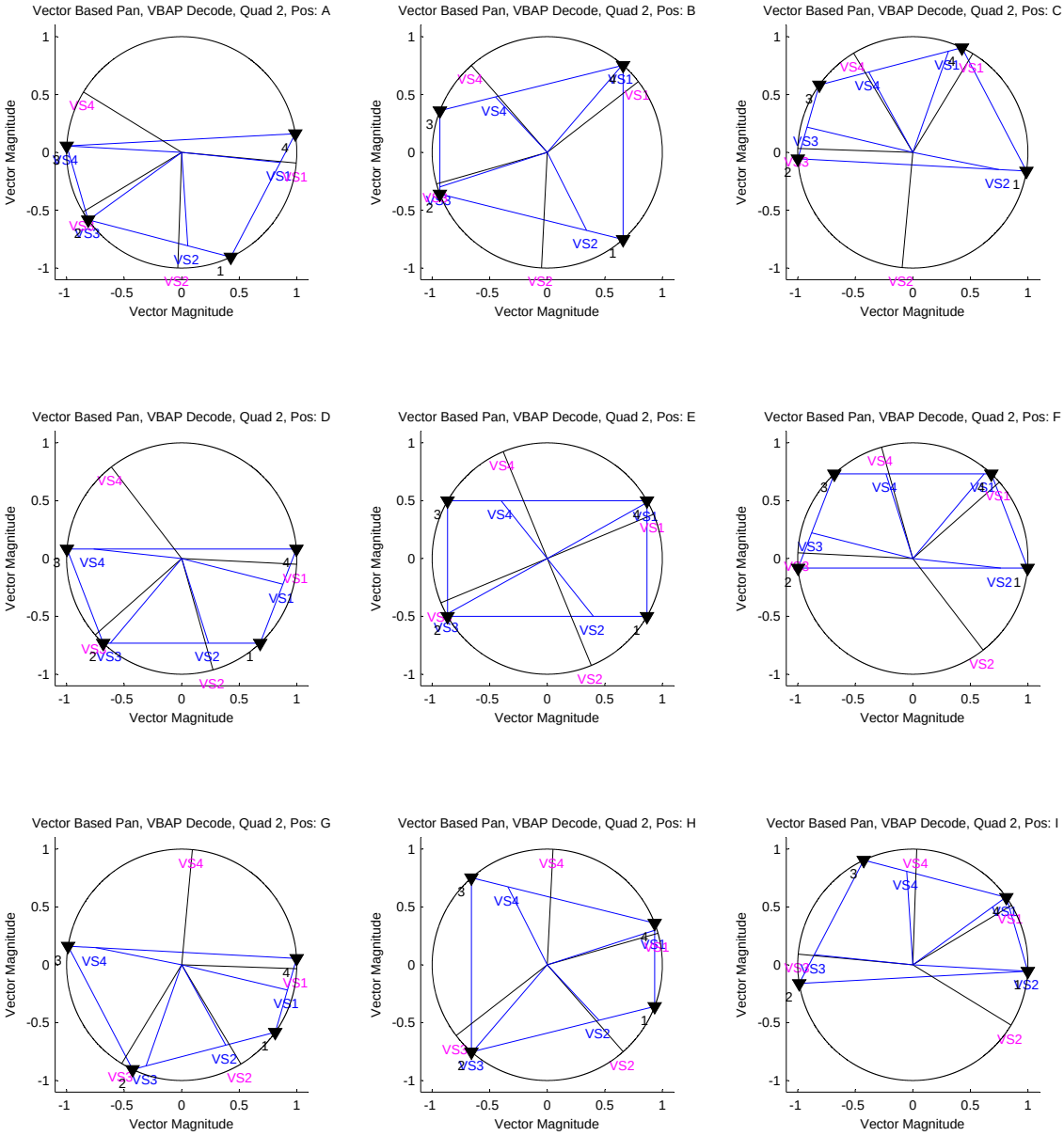


Figure C.2: *VBAP energy vectors: Quad setup 2*

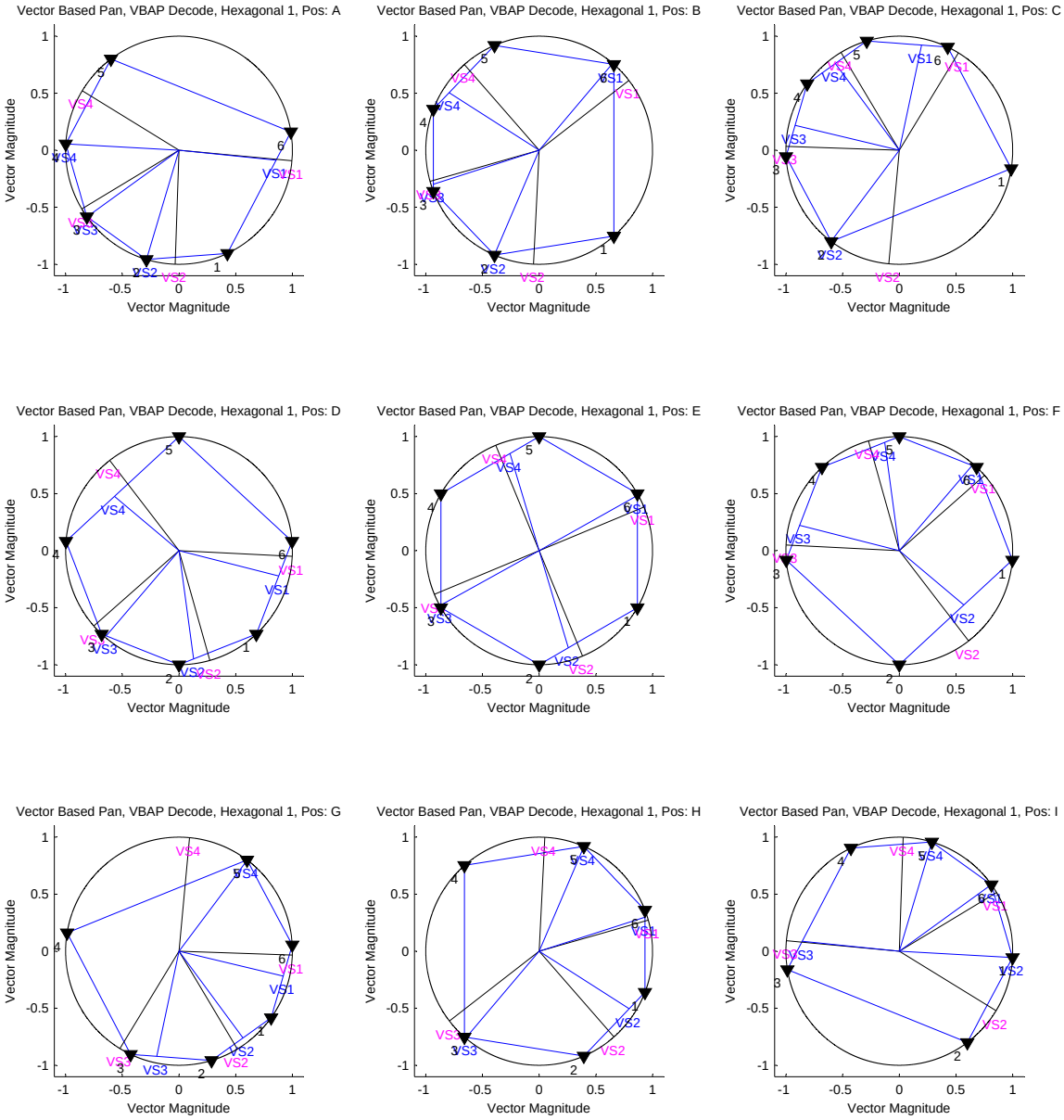


Figure C.3: *VBAP energy vectors: Hexagonal setup 1*

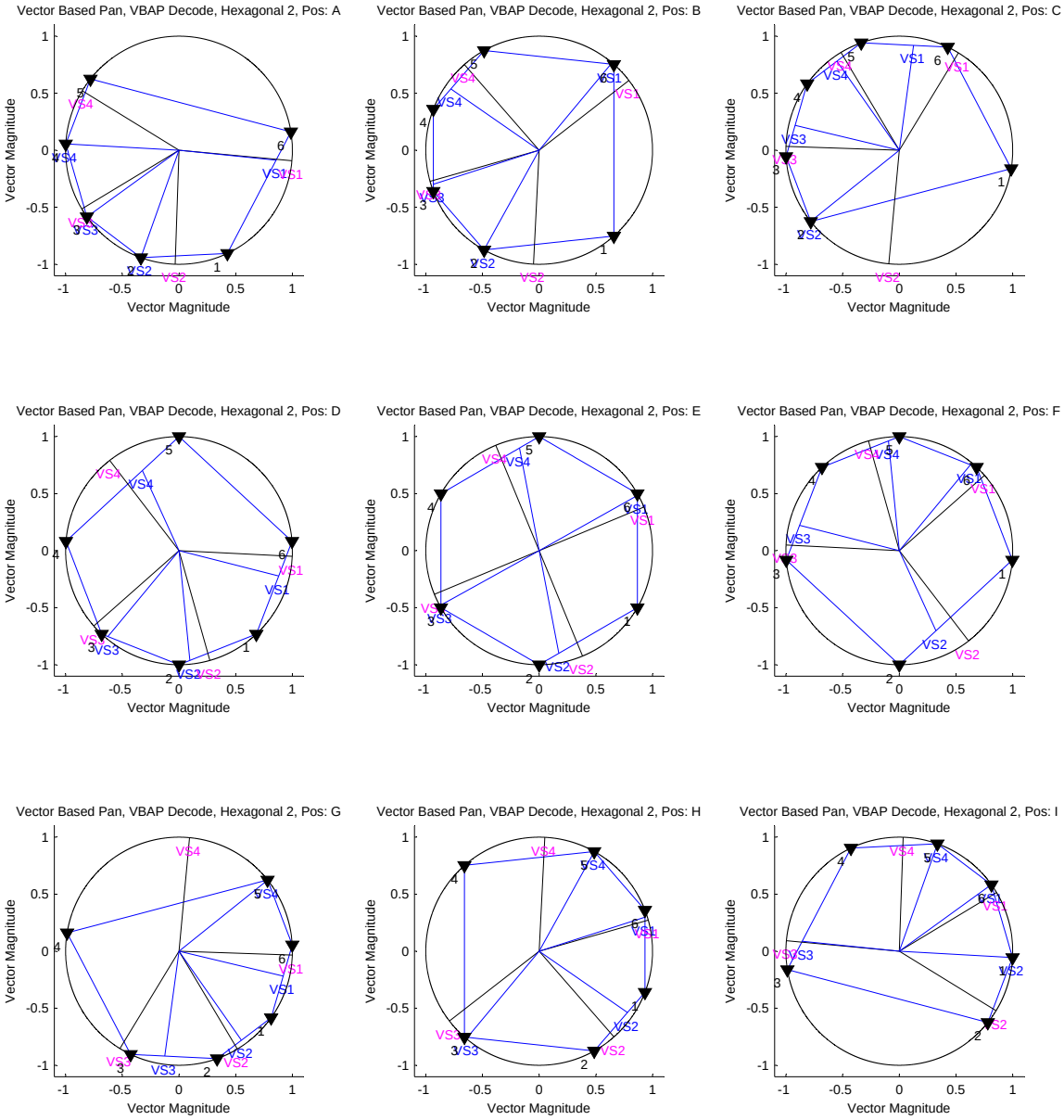


Figure C.4: *VBAP energy vectors: Hexagonal setup 2*

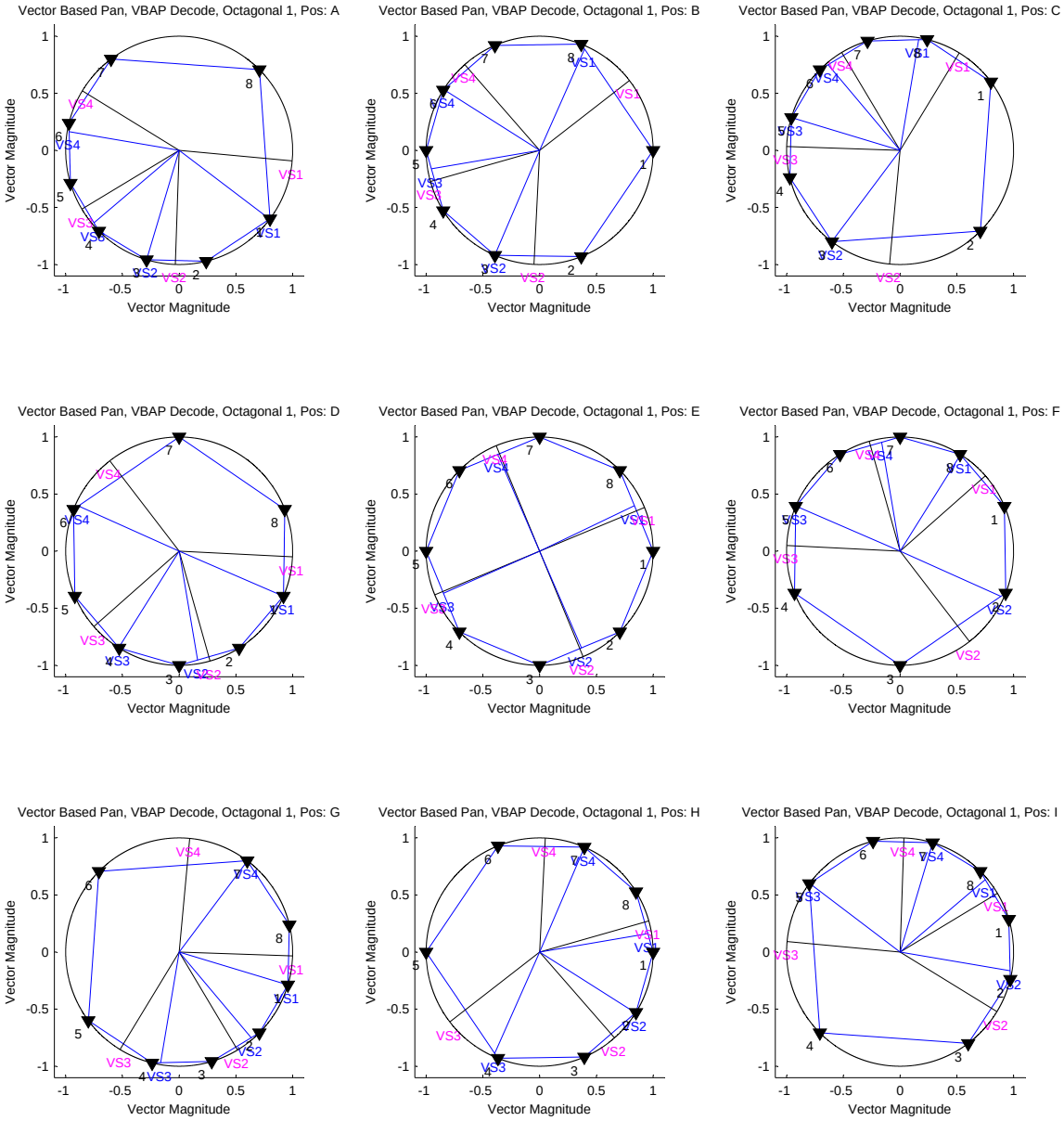


Figure C.5: *VBAP energy vectors: Octagonal setup 1*

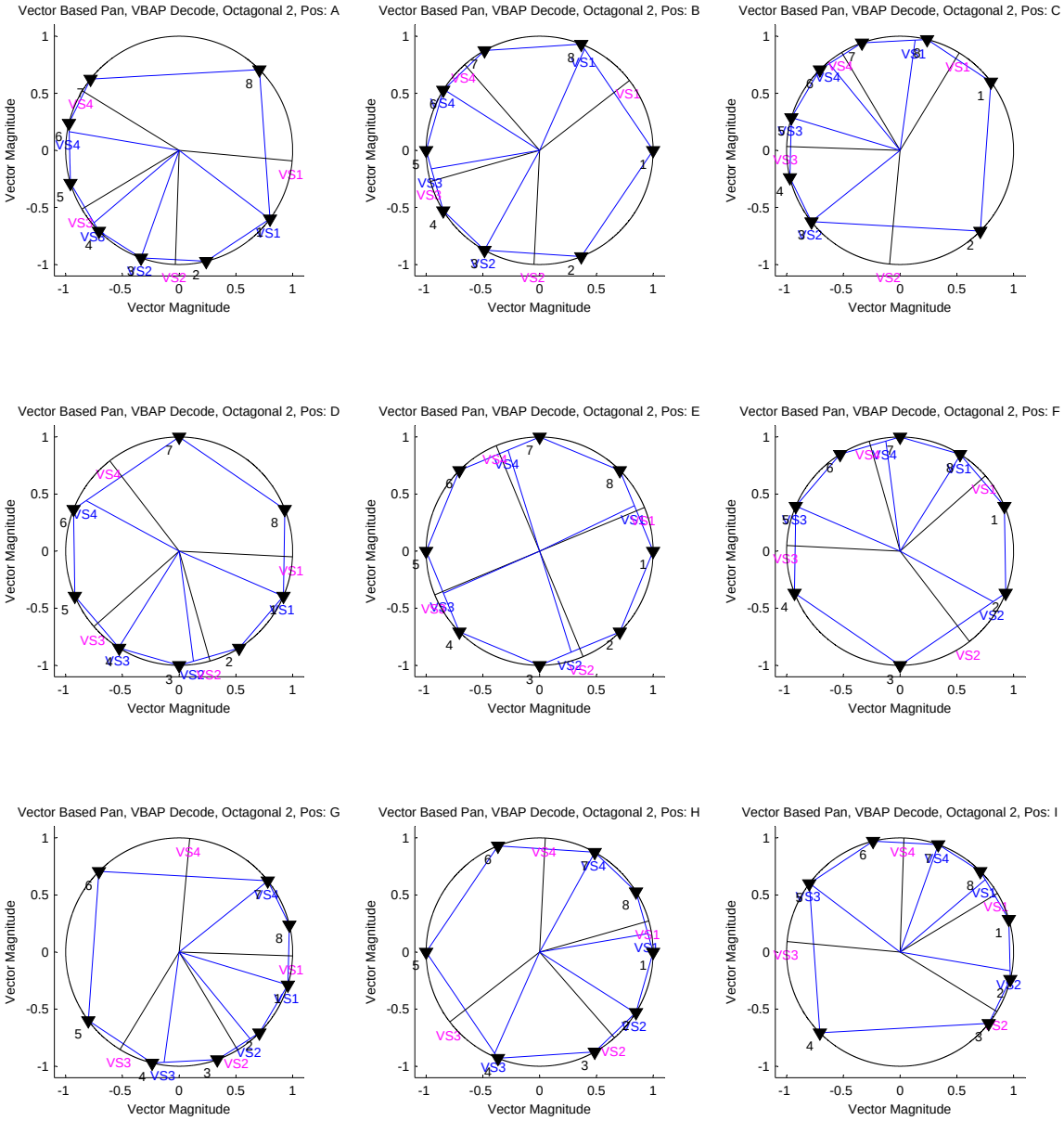


Figure C.6: *VBAP energy vectors: Octagonal setup 2*

D

1st Order Ambisonic Energy Vectors

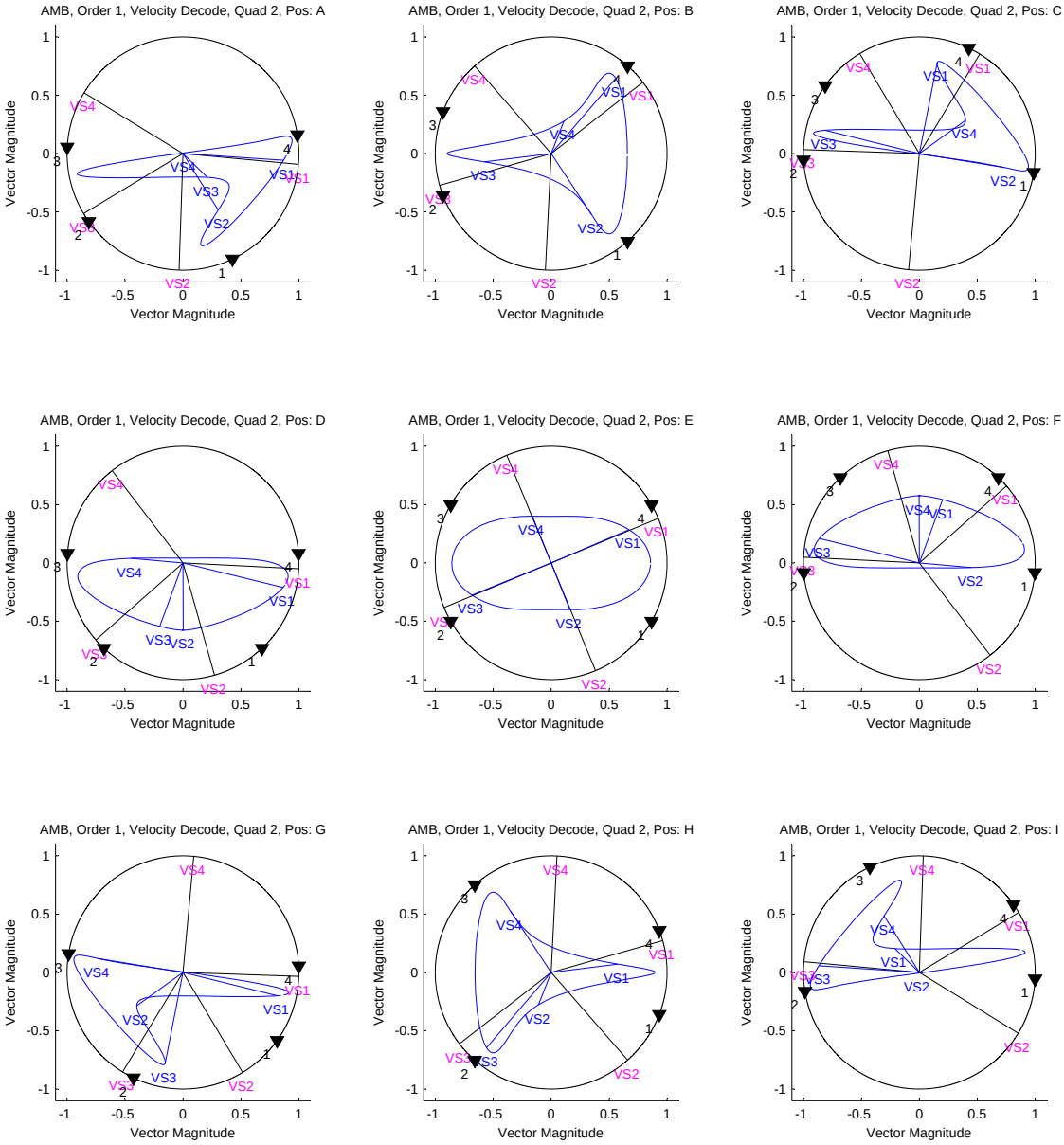


Figure D.1: 1^{st} *Order Ambisonic energy vectors: Quad setup 2, Velocity decode.*

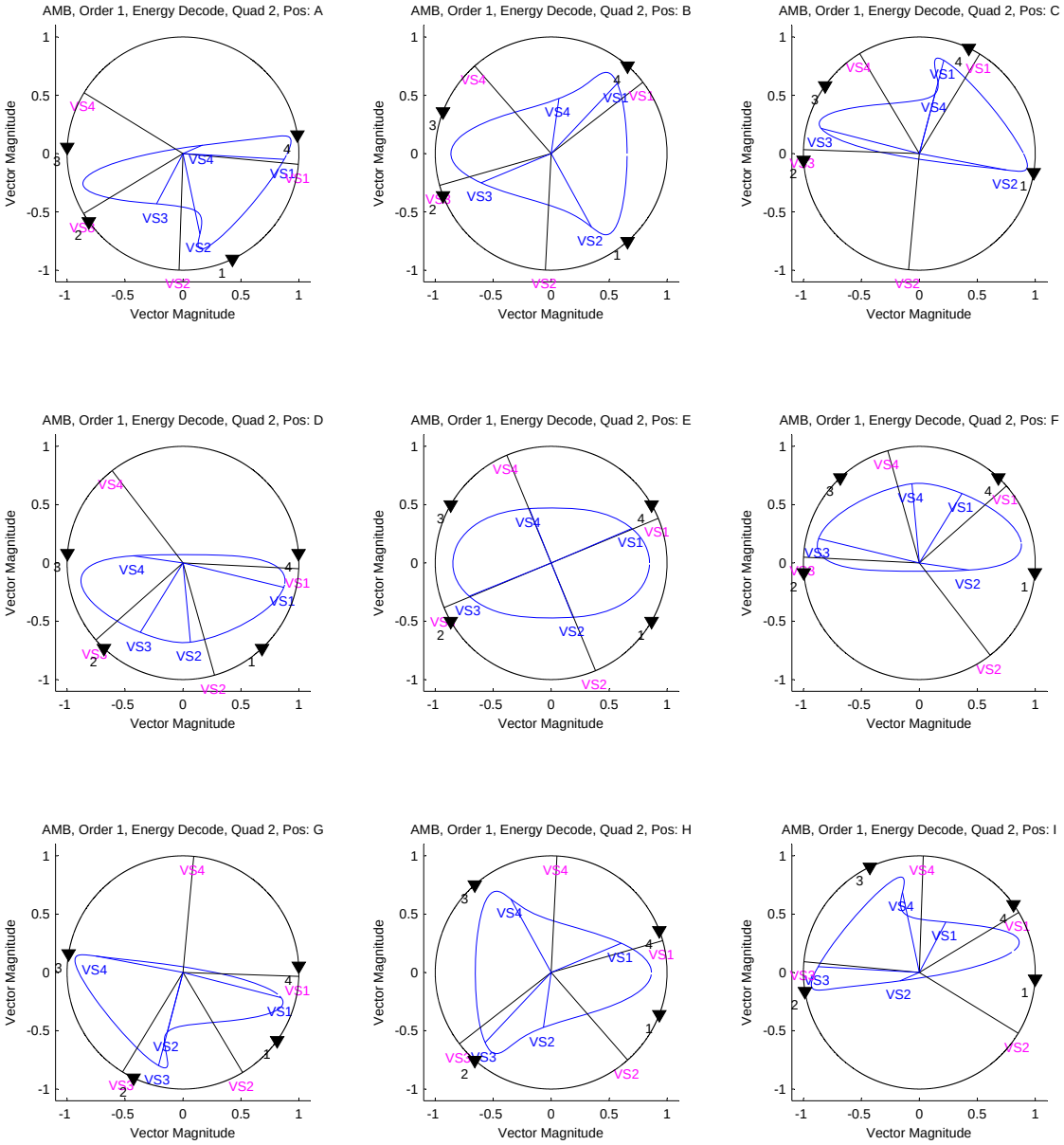


Figure D.2: 1^{st} *Order Ambisonic energy vectors: Quad setup 2, Energy decode.*

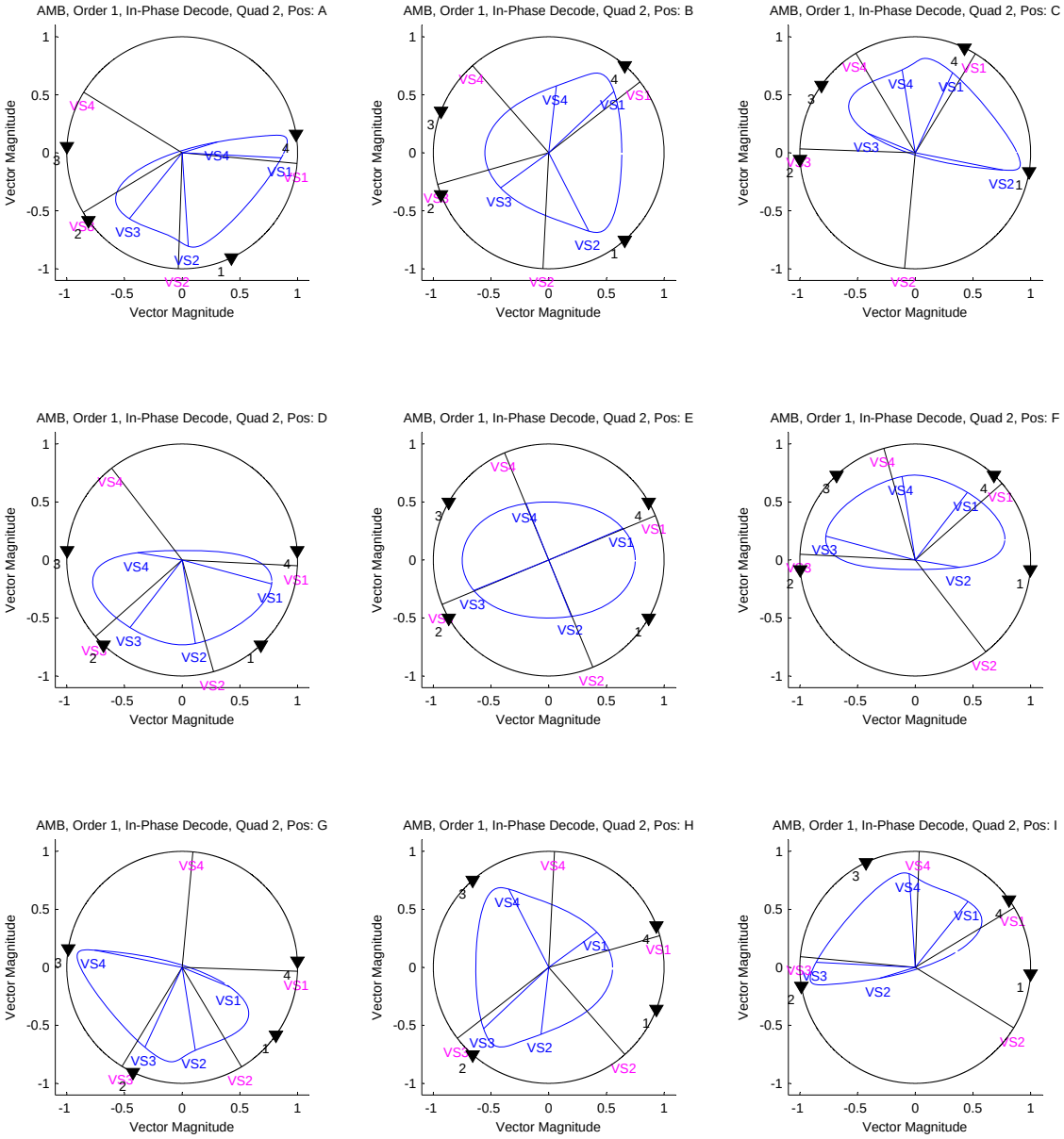


Figure D.3: 1^{st} *Order Ambisonic energy vectors: Quad setup 2, In-phase decode.*

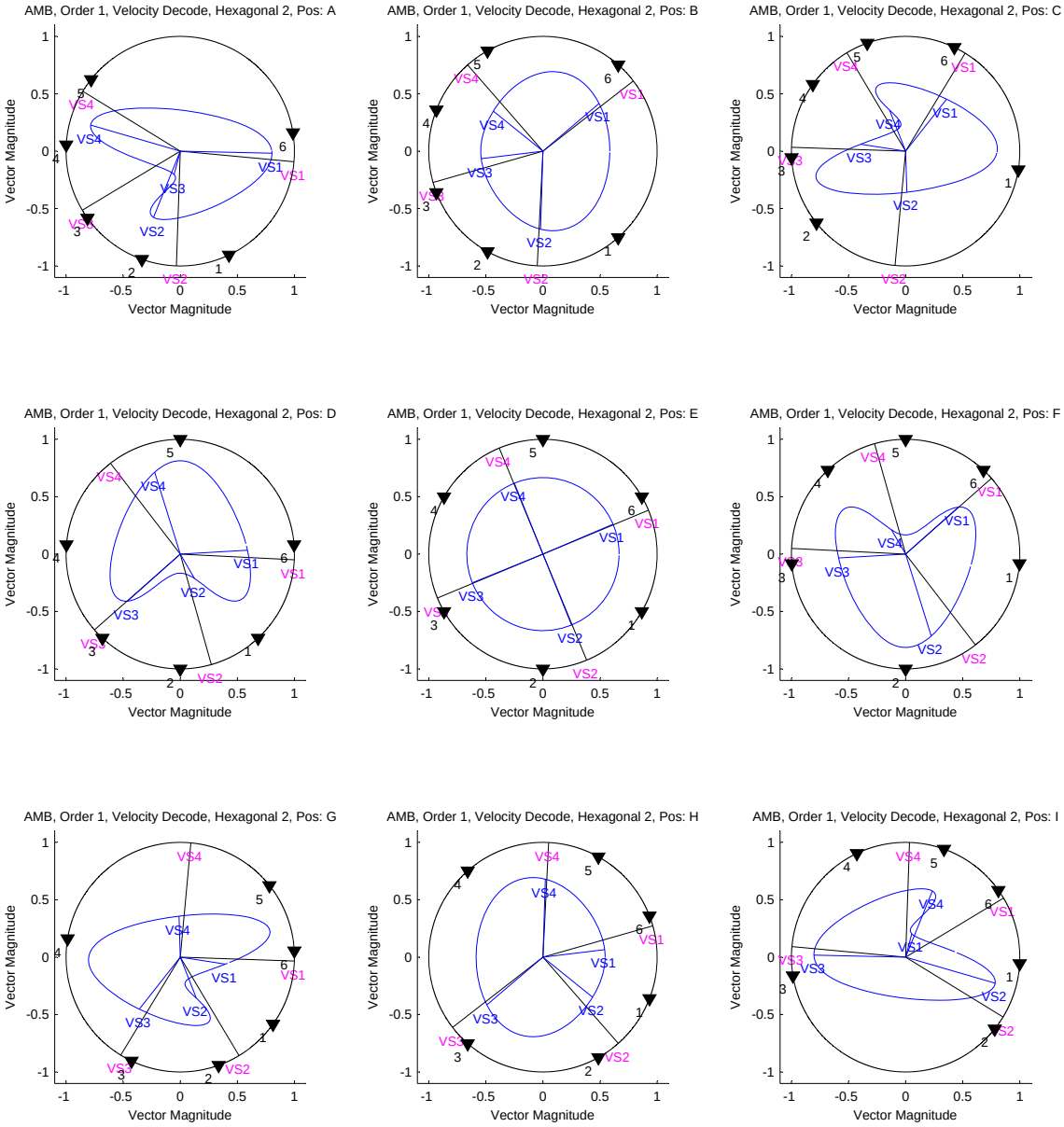


Figure D.4: 1^{st} *Order Ambisonic energy vectors: Hexagonal setup 2, Velocity decode.*

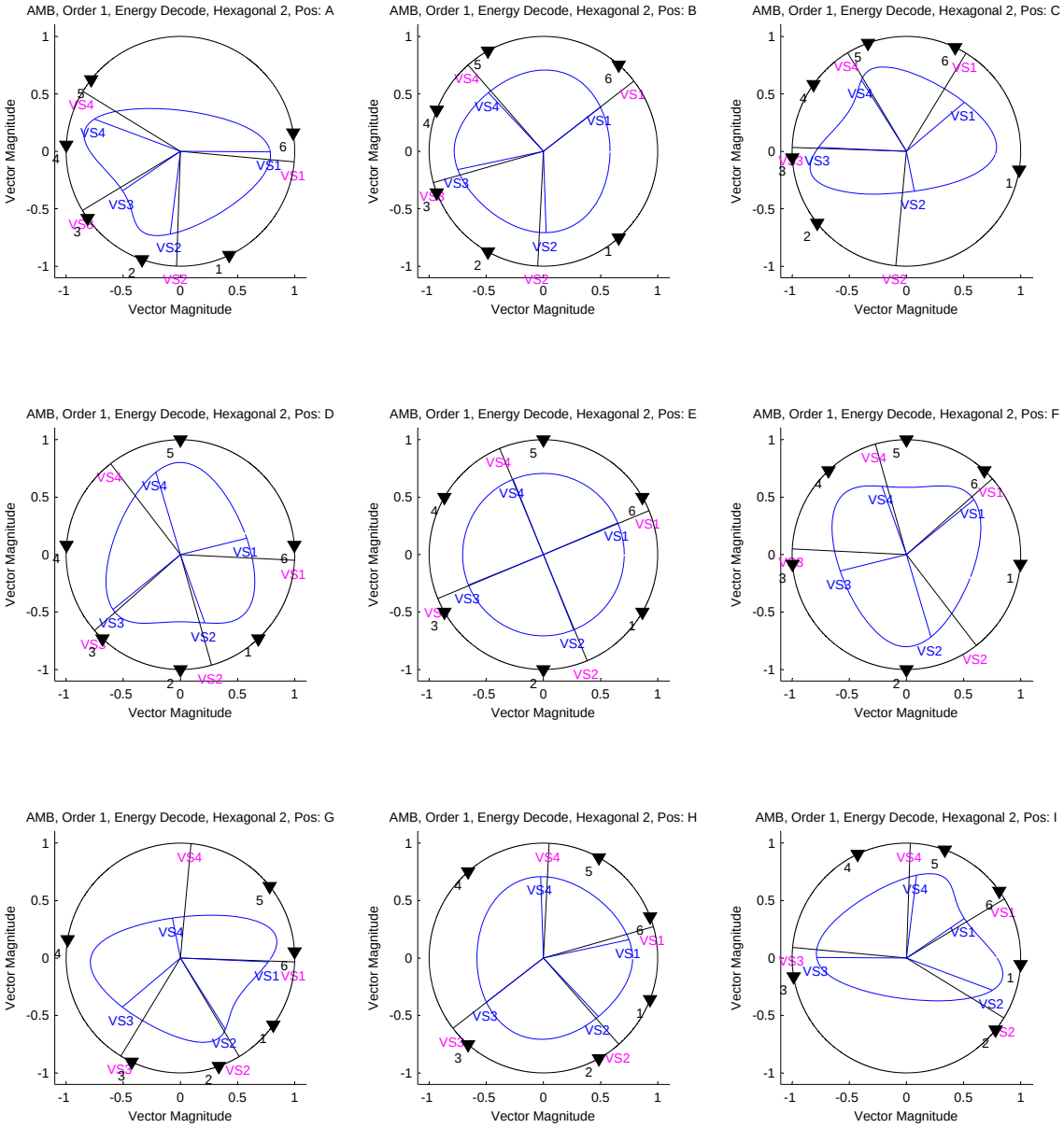


Figure D.5: 1^{st} *Order Ambisonic energy vectors: Hexagonal setup 2, Energy decode.*

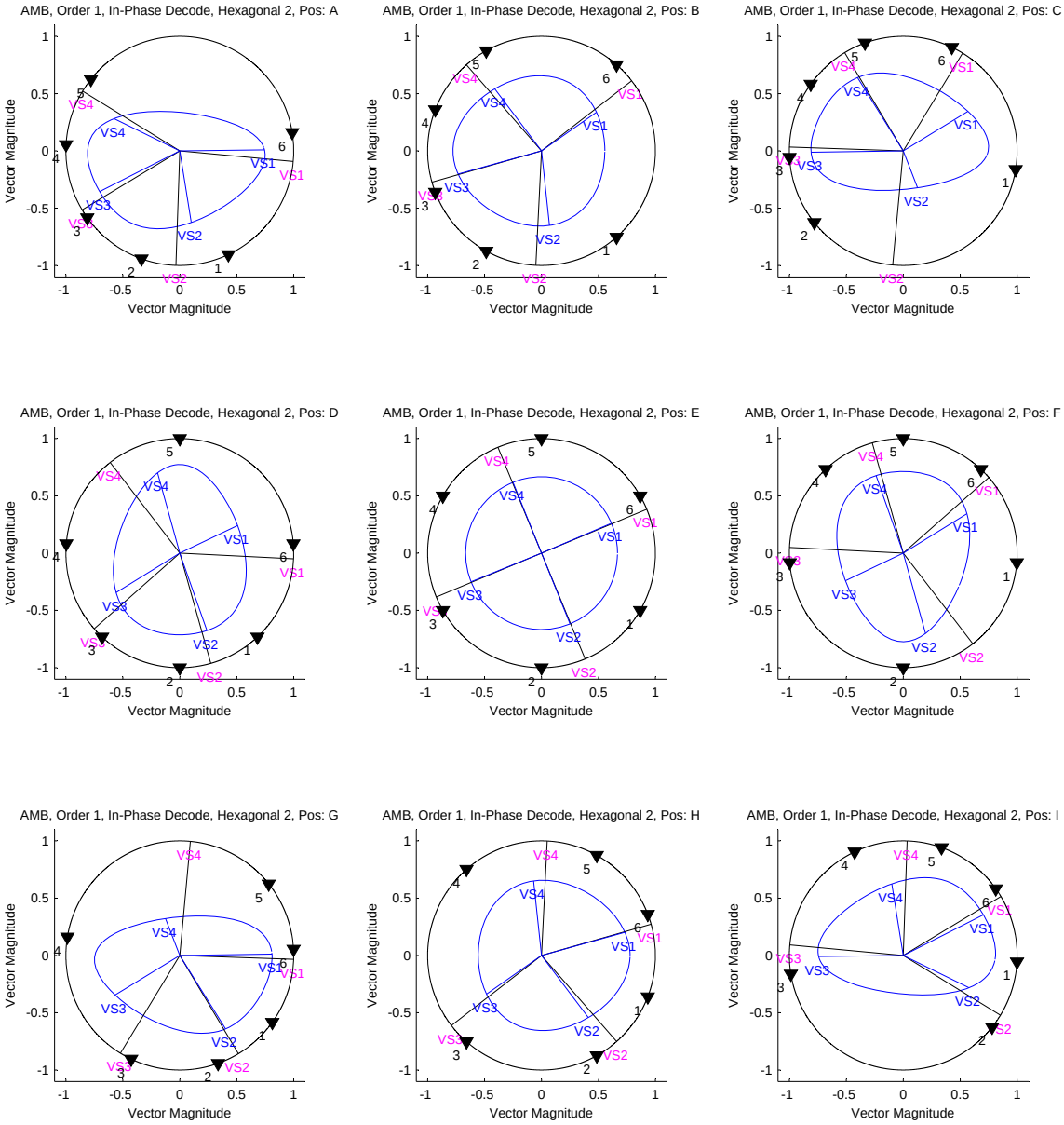


Figure D.6: 1^{st} *Order Ambisonic energy vectors: Hexagonal setup 2, In-phase decode.*

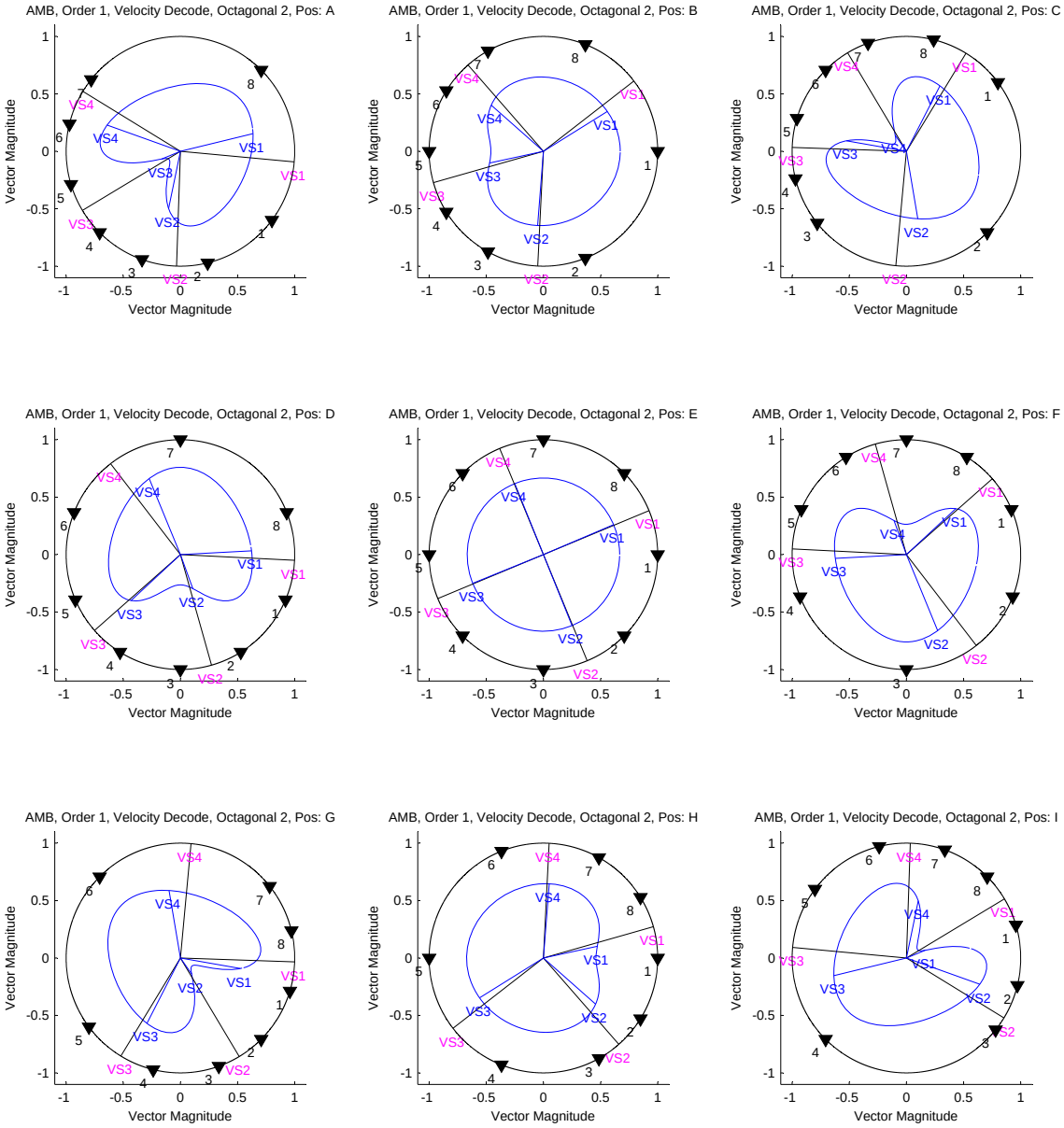


Figure D.7: 1^{st} *Order Ambisonic energy vectors: Octagonal setup 2, Velocity decode.*

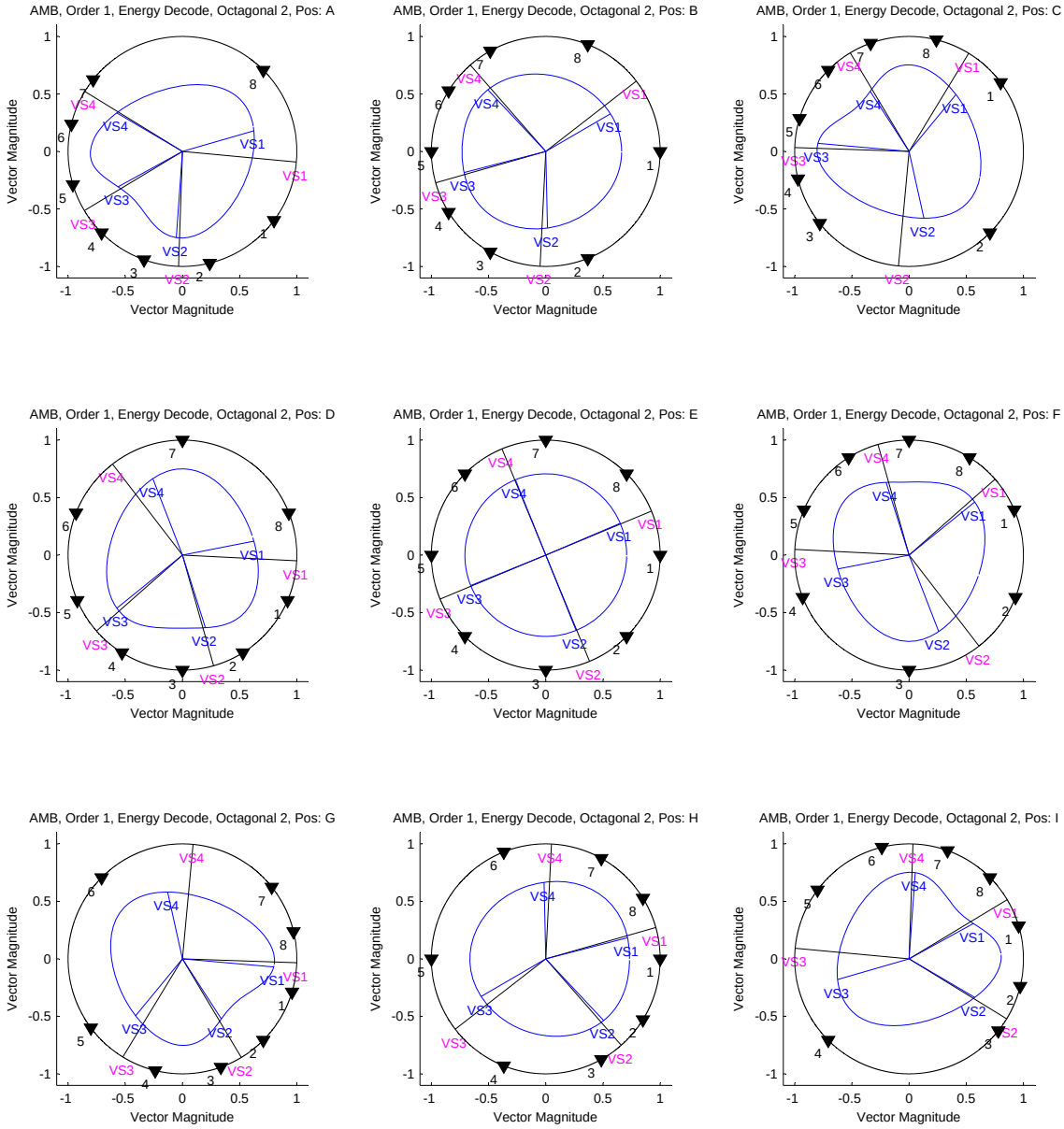


Figure D.8: 1^{st} *Order Ambisonic energy vectors: Octagonal setup 2, Energy decode.*

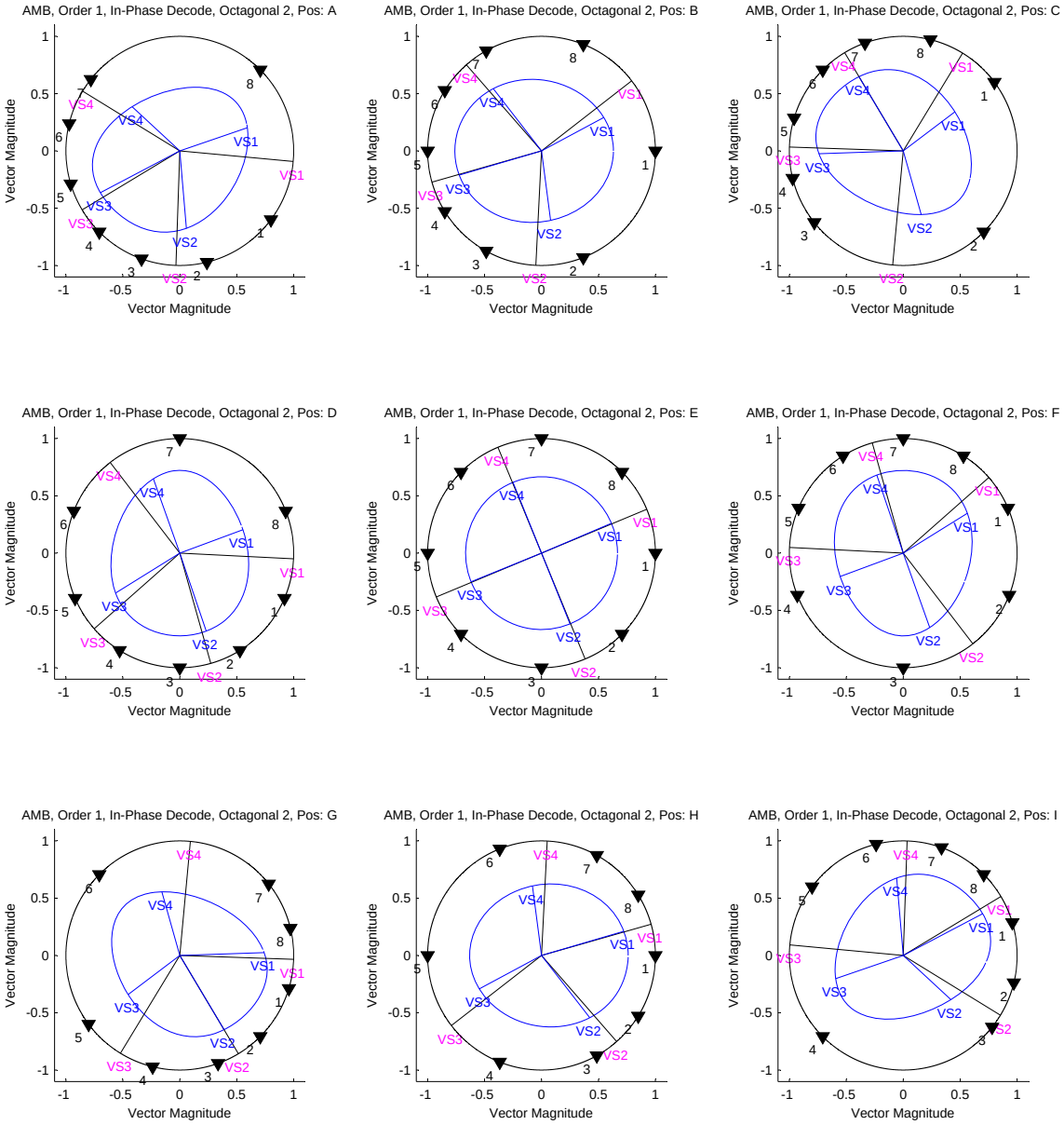


Figure D.9: 1^{st} *Order Ambisonic energy vectors: Octagonal setup 2, In-phase decode.*

E

Higher Order Ambisonic Energy Vectors

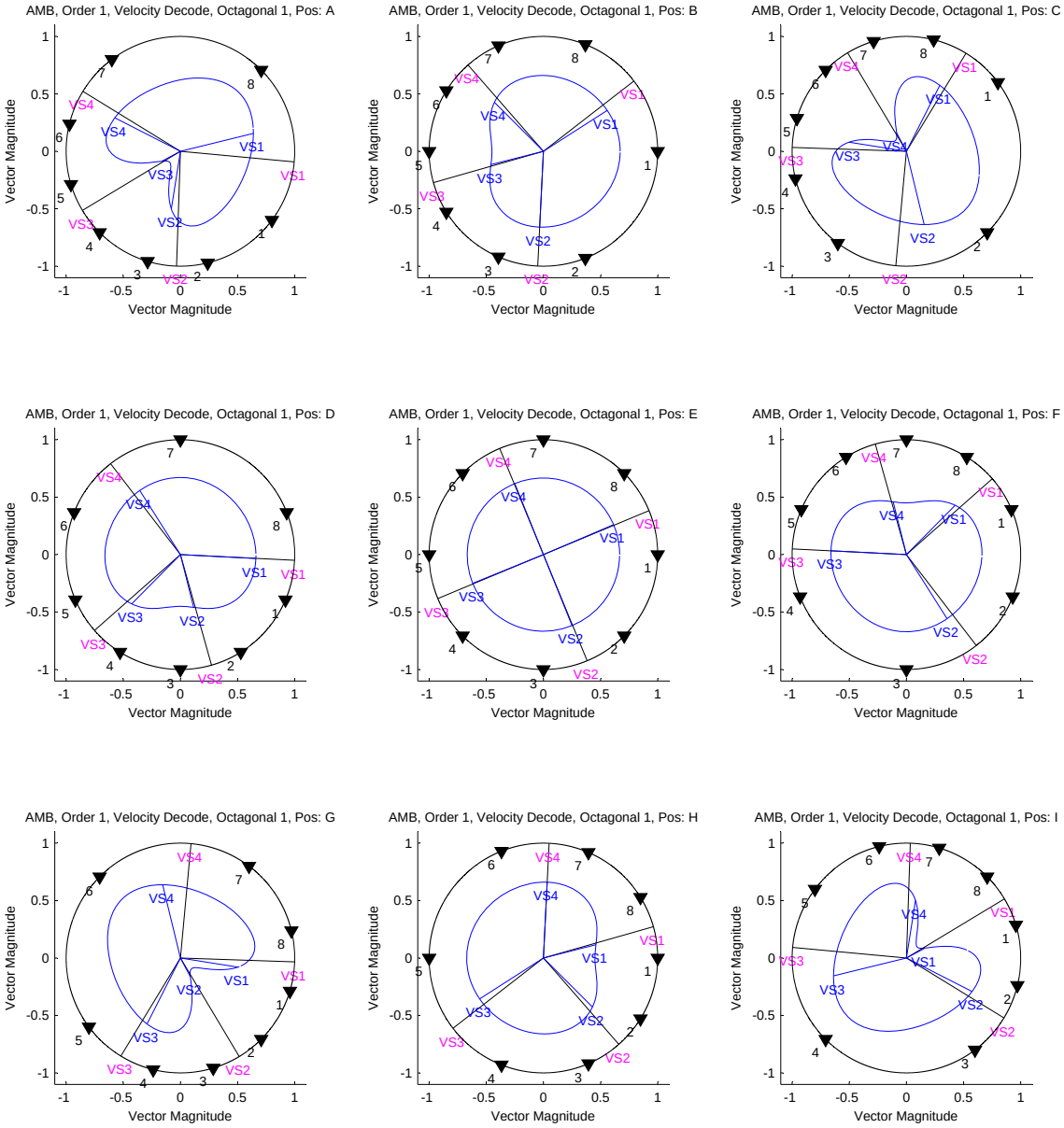


Figure E.1: 1^{st} *Order Ambisonic energy vectors: Octagonal setup 1, Velocity decode.*

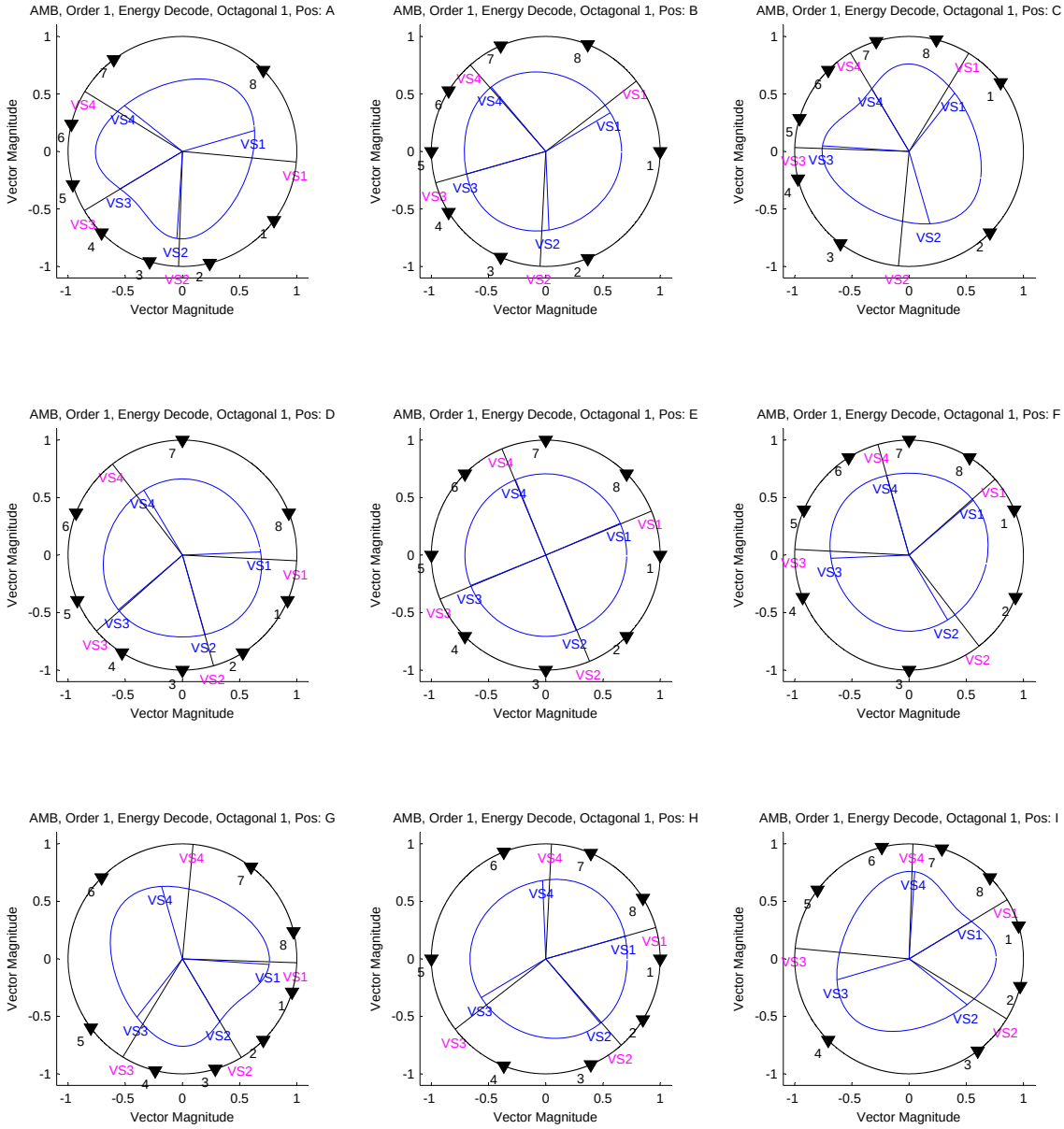


Figure E.2: 1^{st} *Order Ambisonic energy vectors: Octagonal setup 1, Energy decode.*

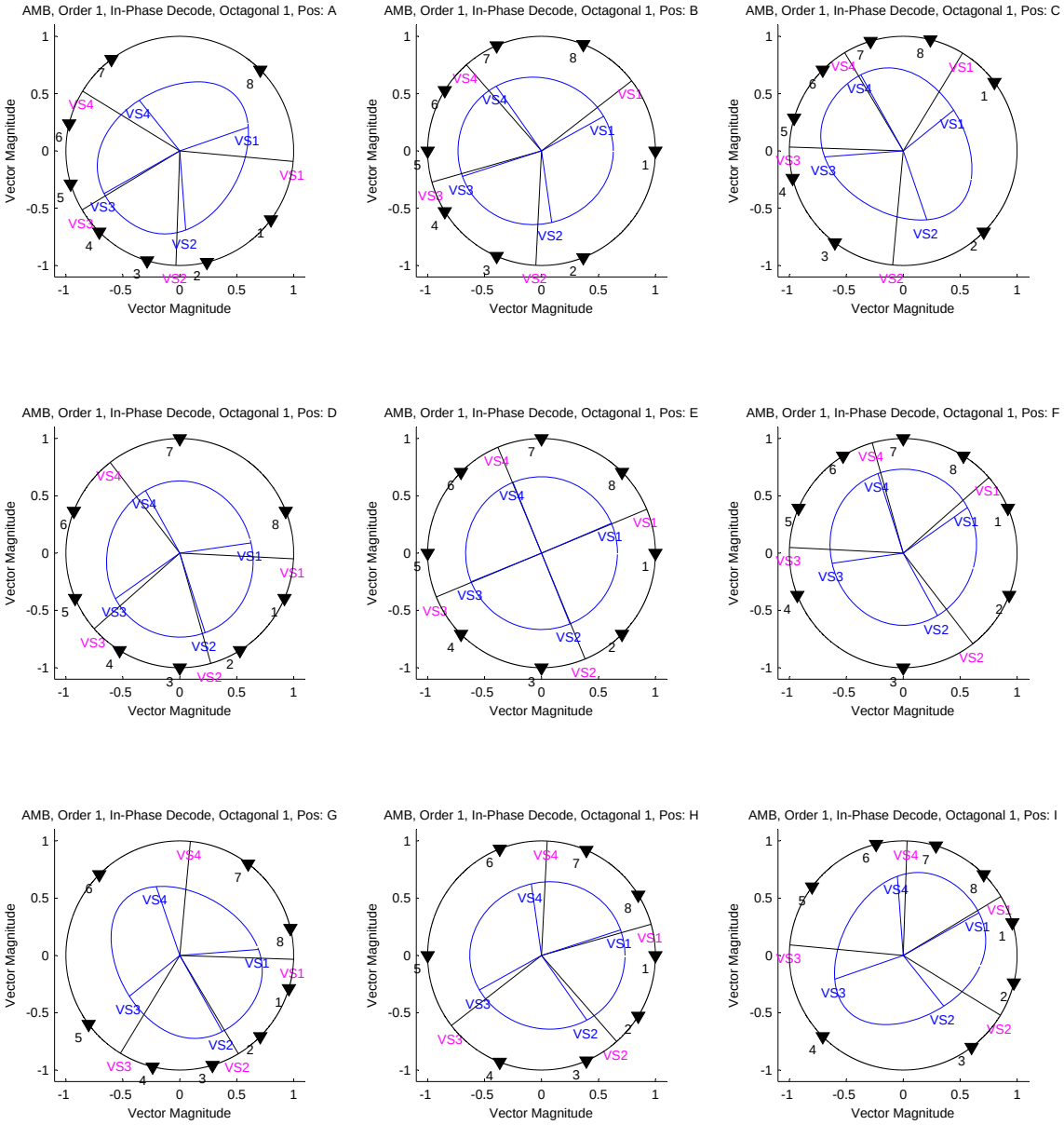


Figure E.3: 1^{st} *Order Ambisonic energy vectors: Octagonal setup 1, In-phase decode.*

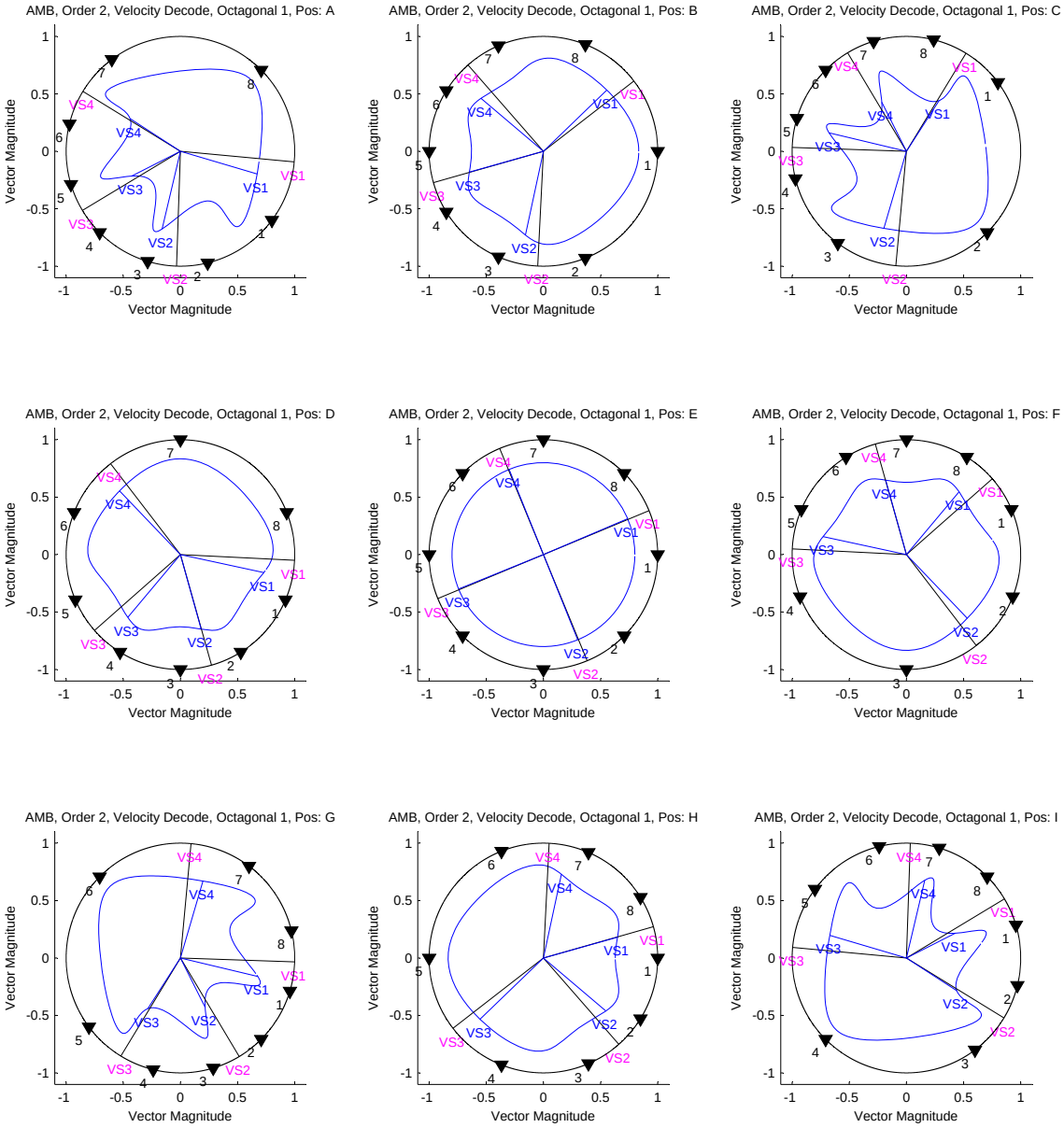


Figure E.4: 2^{nd} *Order Ambisonic energy vectors: Octagonal setup 1, Velocity decode.*

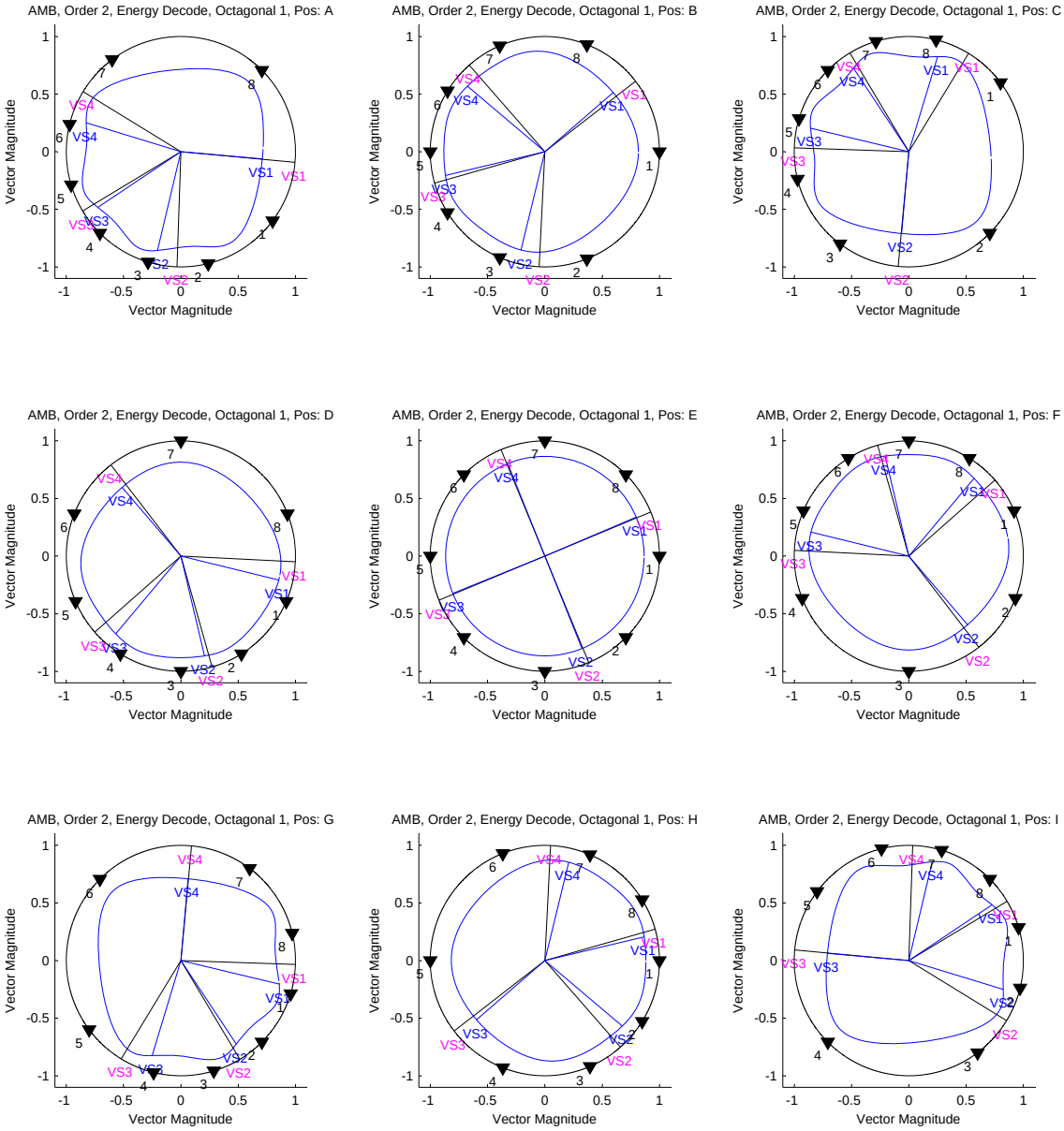


Figure E.5: 2^{nd} *Order Ambisonic energy vectors: Octagonal setup 1, Energy decode.*

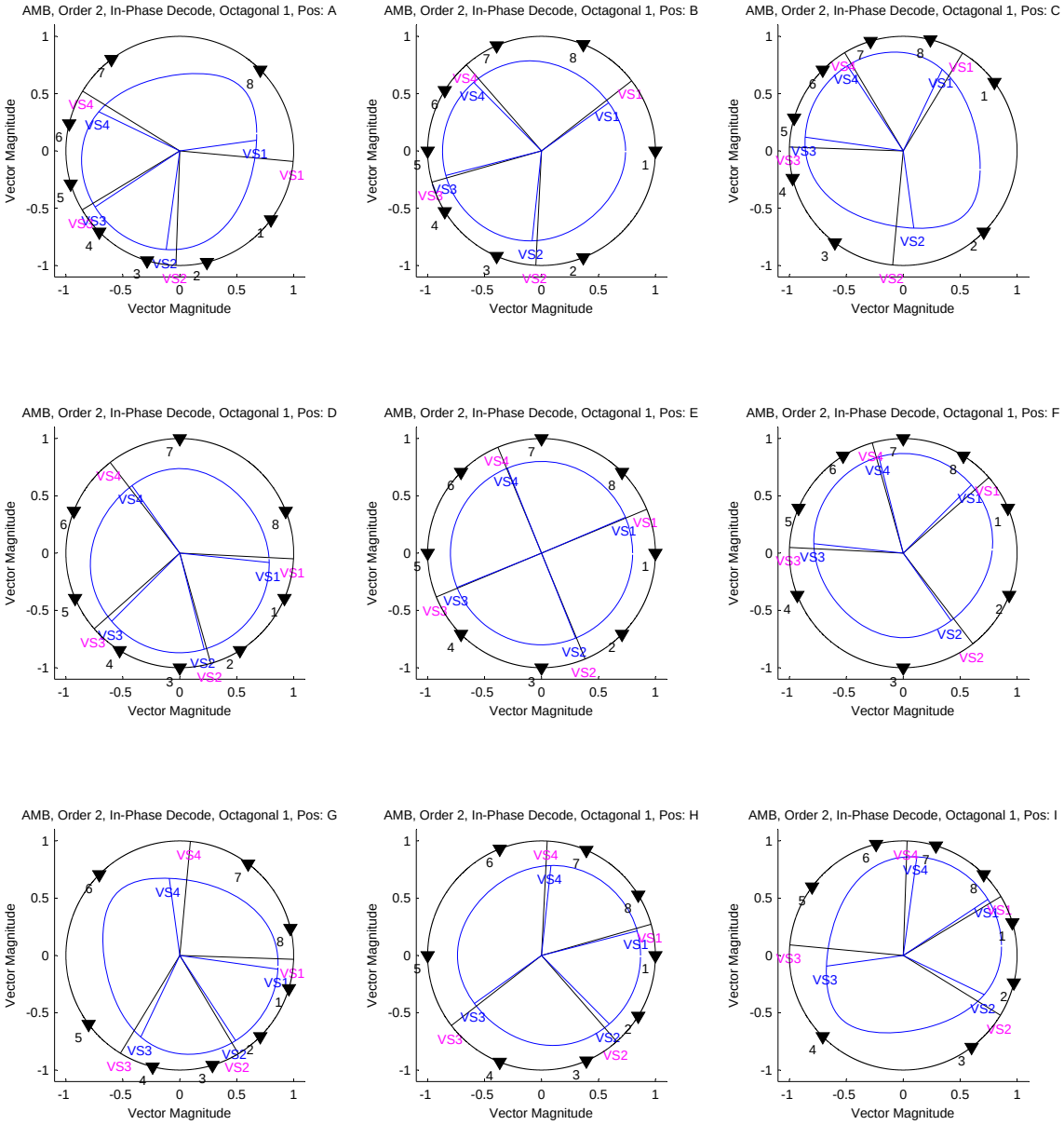


Figure E.6: 2^{nd} *Order Ambisonic energy vectors: Octagonal setup 1, In-phase decode.*

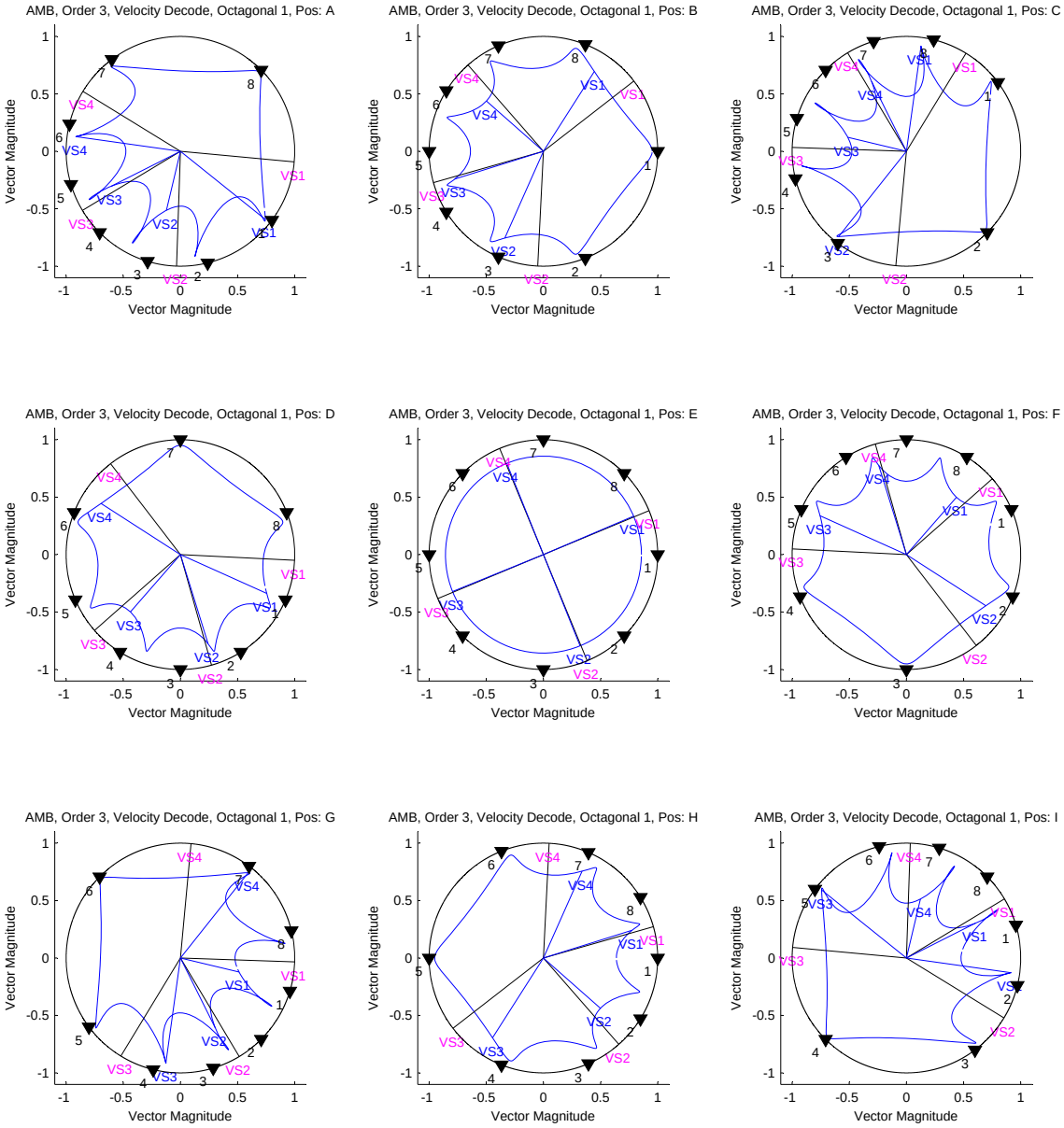


Figure E.7: 3^{rd} *Order Ambisonic energy vectors: Octagonal setup 1, Velocity decode.*

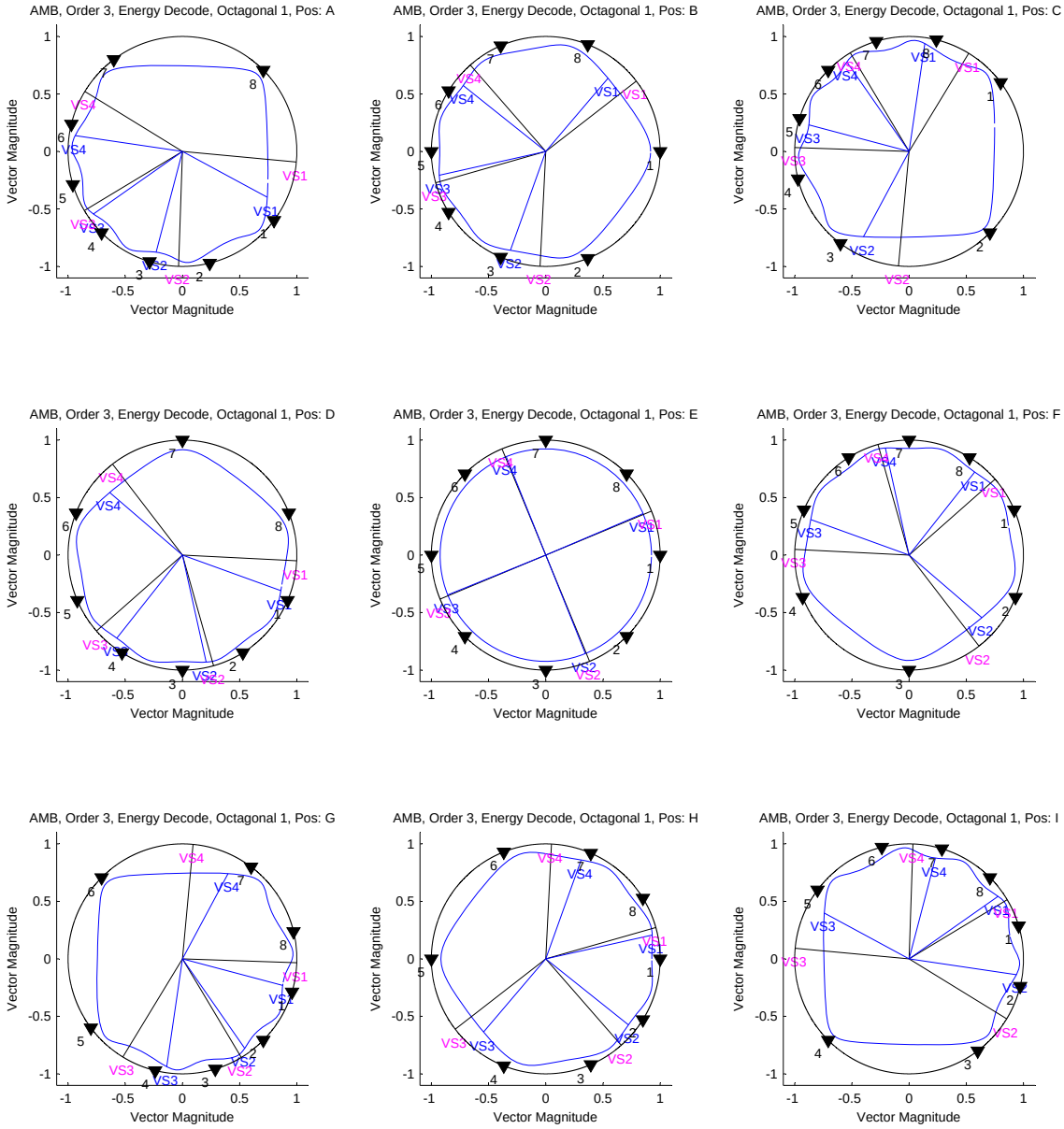


Figure E.8: 3^{nd} *Order Ambisonic energy vectors: Octagonal setup 1, Energy decode.*

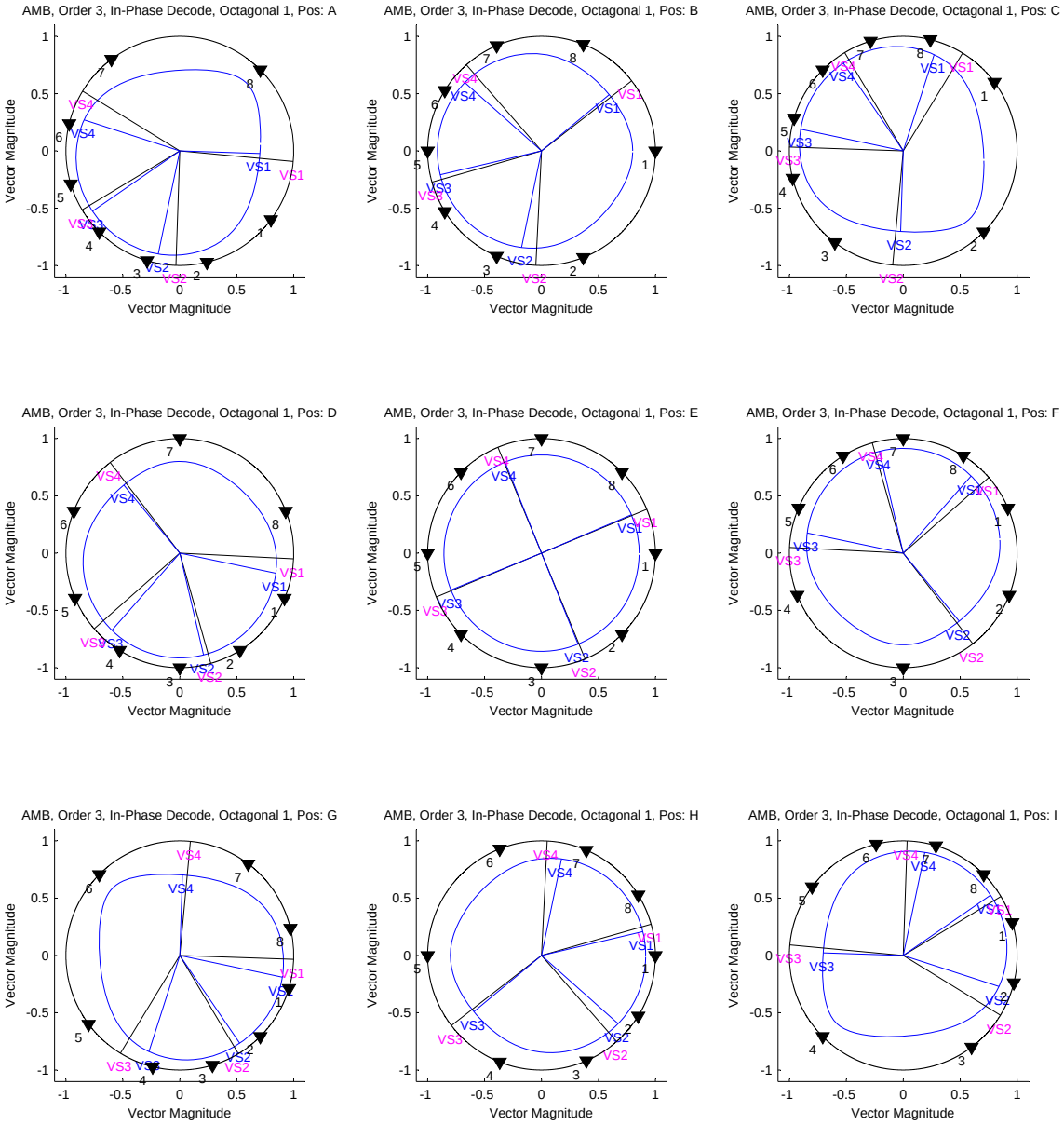


Figure E.9: 3^{nd} *Order Ambisonic energy vectors: Octagonal setup 1, In-phase decode.*

F

Speech Data

F.1 TIMIT Speech Corpus Phrases

The following speech phrases were taken from the TIMIT speech corpus database [Fisher et al., 1986] and recorded with female speech at 96kHz, 24-bit resolution.

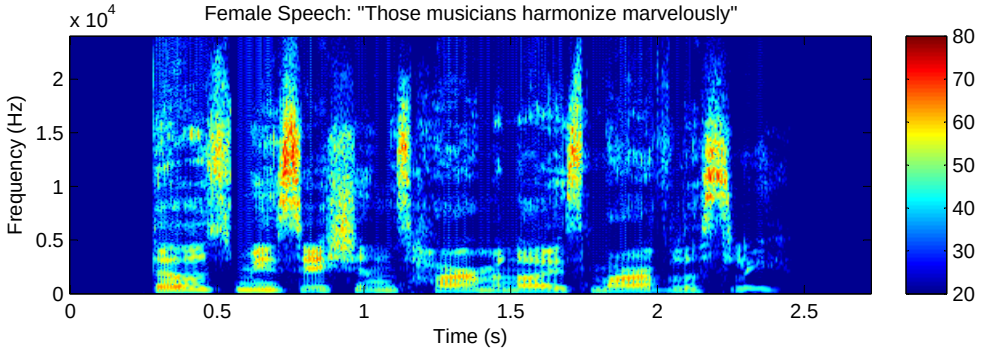


Figure F.1: *Sample 68: "Those musicians harmonize marvelously."*

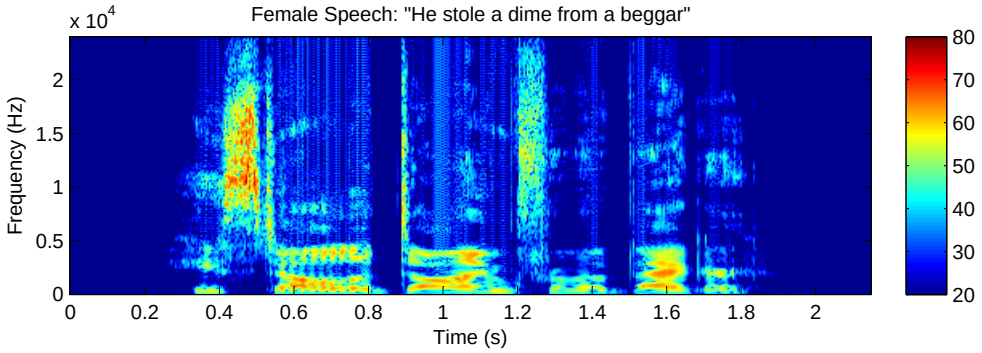


Figure F.2: *Sample 158: "He stole a dime from a beggar"*

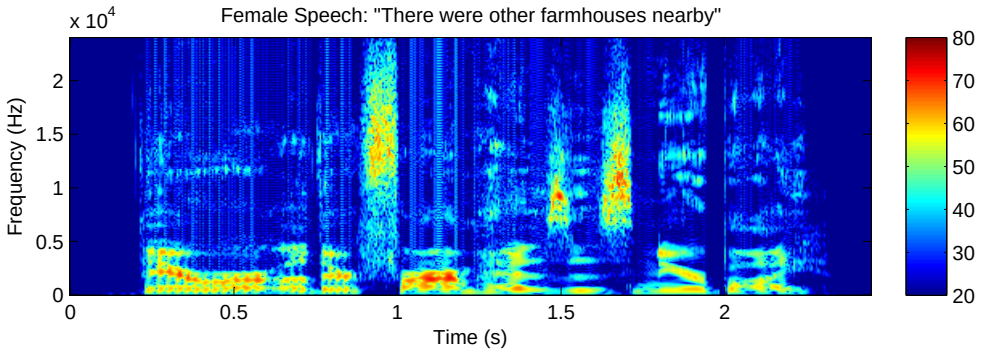


Figure F.3: *Sample 248: "There were other farmhouses nearby"*

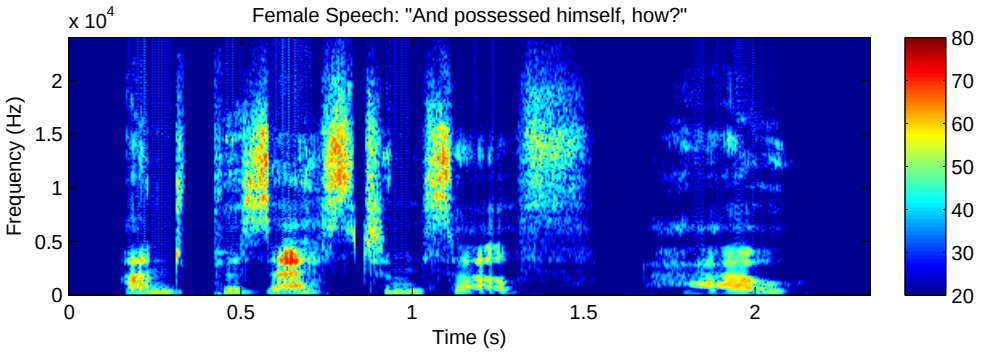


Figure F.4: *Sample 338: "And possessed himself, how?"*

G

ISM and Real Measurement Comparison

G.0.1 Test Environment 1: SRIR Measurements at Violin Positions

ISO 3382 OCTAVE BAND ACOUSTICALPARAMETERS

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB	12.852	44.94	54.657	55.032	57.702	61.121	62.509	67.927	68.493	71.437	73.322	75.561
Noise [dB	0.571	10.415	11.623	6.549	1.809	1.338	0.709	4.299	1.735	5.132	7.749	17.04
strenGth [c	56.148	24.06	14.343	13.968	11.298	7.879	6.491	1.073	0.507	2.437	3.678	1.439
C50 [dB]	1.78	12.127	13.185	12.869	52.519	67.451	1.#IO	1.#IO	1.#IO	1.#IO	1.#IO	71.61
C80 [dB]	3.146	1.513	2.341	4.604	7.618	8.443	11.469	13.467	16.248	24.14	13.719	14.691
D50 [%]	60.106	5.774	4.583	4.911	0.001	0	0	0	0	0	0	0
Ts [ms]	461.409	111.298	101.003	86.335	71.217	66.758	59.973	57.273	54.995	53.265	56.754	56.027
EDT [s]	25.903	0.432	0.714	0.744	0.545	0.516	0.017	0.008	0.005	0.003	0.006	0.006
T20 [s]		0.896	0.974	1.073	0.946	0.848	0.853	0.751	0.518		0.765	0.776
r T20		0.981	0.978	0.99	0.995	0.998	0.995	0.992	0.988		0.994	0.986
T30 [s]	0	0.994	0.992	1.005	0.979	0.898	0.913	0.767	0.536	0.334	0.812	0.852
r T30	101	0.972	0.992	0.994	0.997	0.998	0.996	0.998	0.997	0.969	0.997	0.995
Tuser [s]		0.867	0.974	0.965	0.983	0.889	0.868	0.76	0.52	0.334	0.785	0.831
r Tuser		0.982	0.984	0.984	0.995	0.996	0.999	0.999	0.998	0.963	0.999	0.996
LE												
LF												
LFC												

RTU = RT User (10. dB, 30. dB)

Table G.1: *ISO-3382 acoustical parameters: Test environment 1, 1m from source, height = , Genelec 1029a, W channel.*

ISO 3382 OCTAVE BAND ACOUSTICALPARAMETERS

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB	8.95	42.633	51.28	53.185	54.26	56.849	57.798	63.108	62.7	65.442	68.148	70.144
Noise [dB	2.192	4.803	15.801	13.657	9.973	16.019	11.6	7.242	7.135	1.036	18.849	21.59
strenGth [c	60.05	26.367	17.72	15.815	14.74	12.151	11.202	5.892	6.3	3.558	8.852	6.856
C50 [dB]	2.91	2.388	1.906	2.827	1.739	2.882	6.737	8.877	10.947	18.447	8.388	9.046
C80 [dB]	6.366	3.137	0.717	1.814	4.818	6.452	9.832	11.788	14.654	24.147	11.462	11.746
D50 [%]	66.153	36.588	39.202	34.279	59.879	66.009	82.509	88.534	92.558	98.59	87.341	88.923
Ts [ms]	71.198	102.957	102.415	86.316	64.733	56.008	40.779	34.863	30.088	24.864	35.186	33.925
EDT [s]	0.702	0.704	0.863	0.839	0.904	0.802	0.637	0.452	0.276	0.005	0.563	0.548
T20 [s]	0.661	0.923	1.021	0.898	0.902	0.967	0.926	0.748	0.542	0.298	0.782	0.771
r T20	0.946	0.981	0.981	0.987	0.998	0.993	0.996	0.997	0.996	0.97	0.997	0.998
T30 [s]	0.656	0.965	1.075	0.995	0.972	1	0.942	0.794	0.563	0.342	0.864	0.867
r T30	0.944	0.98	0.993	0.989	0.997	0.997	0.999	0.998	0.998	0.99	0.996	0.995
Tuser [s]	0.771	0.867	1.077	0.989	0.974	1.036	0.964	0.786	0.57	0.344	0.861	0.847
r Tuser	0.907	0.97	0.992	0.991	0.995	0.996	0.998	0.999	0.999	0.987	0.997	0.996
LE												
LF												
LFC												

RTU = RT User (10. dB, 30. dB)

Table G.2: *ISO-3382 acoustical parameters: Test environment 1, 2m from source, height = , Genelec 1029a, W channel.*

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB	7.034	40.343	47.131	50.789	51.638	54.501	54.217	59.507	58.816	59.447	64.367	65.812
Noise [dB	0.981	15.033	26.598	24.406	23.009	27.146	25.1	24.477	7.99	2.508	31.593	33.432
strenGth [c	61.966	28.657	21.869	18.211	17.362	14.499	14.783	9.493	10.184	9.553	12.633	11.188
C50 [dB]	1.707	3.381	1.158	2.952	0.367	1.568	3.232	5.643	7.301	14.44	5.182	5.683
C80 [dB]	6.93	0.691	0.834	2.102	3.679	5.022	6.088	8.394	11.749	20.163	8.277	8.718
D50 [%]	40.301	31.463	43.372	33.634	52.11	58.928	67.792	78.572	84.306	96.528	76.732	78.727
Ts [ms]	96.256	116.72	98.61	87.506	73.854	61.678	49.607	38.594	30.389	17.989	39.653	37.579
EDT [s]	0.728	0.962	1.11	0.927	0.94	0.77	0.858	0.655	0.398	0.166	0.653	0.632
T20 [s]		0.862	0.777	0.974	0.897	0.891	0.912	0.749	0.522	0.319	0.799	0.825
r T20		0.989	0.979	0.989	0.997	0.997	0.998	0.997	0.997	0.988	0.998	0.998
T30 [s]	0	0.76	0.784	0.951	0.789	0.891	0.878	0.757	0.558	0.338	0.816	0.832
r T30	101	0.971	0.989	0.99	0.99	0.998	0.998	0.999	0.997	0.995	0.999	0.999
Tuser [s]		0.737	0.756	0.944	0.803	0.885	0.912	0.778	0.57	0.342	0.84	0.854
r Tuser		0.963	0.98	0.987	0.986	0.997	0.998	0.998	0.998	0.997	0.999	0.999
LE												
LF												
LFC												

RTU = RT User (10. dB, 30. dB)

Table G.3: *ISO-3382 acoustical parameters: Test environment 1, 4m from source, height = , Genelec 1029a, W channel.*

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB	4.585	35.117	46.399	51.572	50.791	53.371	52.567	56.158	55.339	54.45	61.517	62.846
Noise [dB	1.69	6.142	8.522	0.932	2.6	0.717	2.488	6.725	6.794	0.142	4.782	12.985
strenGth [c	64.415	33.883	22.601	17.428	18.209	15.629	16.433	12.842	13.661	14.55	15.483	14.154
C50 [dB]	6.712	9.202	4.326	1.465	1.68	2.729	2.666	3.196	6.042	12.276	3.686	3.94
C80 [dB]	9.739	0.777	3.65	6.022	3.504	4.56	4.759	6.266	9.507	17.593	6.266	6.747
D50 [%]	82.425	10.729	26.972	58.354	59.549	65.211	64.883	67.612	80.079	94.41	70.029	71.242
Ts [ms]	21.67	127.808	96.556	66.83	69.568	60.327	58.792	46.856	32.796	16.541	47.388	45.708
EDT [s]	0.436	0.908	0.941	0.715	0.923	0.855	0.833	0.764	0.516	0.244	0.753	0.728
T20 [s]		0.92	1.11	0.976	0.952	0.944	0.911	0.799	0.565	0.332	0.847	0.868
r T20		0.966	0.99	0.962	0.995	0.998	0.998	0.998	0.999	0.997	0.999	0.998
T30 [s]	0	0.881	1.011	1.068	0.946	0.984	0.948	0.814	0.573	0.35	0.895	0.921
r T30	101	0.979	0.993	0.982	0.998	0.998	0.998	0.999	0.999	0.997	0.998	0.998
Tuser [s]		0.885	0.965	1.179	0.948	0.972	0.954	0.822	0.564	0.35	0.898	0.932
r Tuser		0.949	0.985	0.983	0.995	0.997	0.999	0.999	0.999	0.996	0.999	0.999
LE												
LF												
LFC												

RTU = RT User (10. dB, 30. dB)

Table G.4: *ISO-3382 acoustical parameters: Test environment 1, 8m from source, height = , Genelec 1029a, W channel.*

G.0.2 Test Environment 1: SRIR Measurements at Female Vocal Positions

ISO 3382 OCTAVE BAND ACOUSTICALPARAMETERS

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB	11.916	45.035	54.323	54.577	57.272	60.67	61.995	67.667	67.713	71.423	72.92	75.272
Noise [dB	2.084	5.575	6.43	2.181	2.262	0.024	2.44	6.671	3.521	2.622	10.888	14.175
strenGth [c	57.084	23.965	14.677	14.423	11.728	8.33	7.005	1.333	1.287	2.423	4.08	1.728
C50 [dB]	7.753	5.27	5.211	6.109	8.804	11.242	13.68	15.263	17.963	27.02	15.697	16.637
C80 [dB]	8.459	9.421	7.868	8.983	10.268	12.787	15.434	18.122	21.415	31.955	18.074	19.14
D50 [%]	85.632	77.092	76.852	80.322	88.361	93.012	95.89	97.11	98.427	99.802	97.377	97.877
Ts [ms]	78.286	65.056	52.44	38.154	25.982	18.893	13.623	11.188	9.485	7.587	10.754	10.042
EDT [s]	0.95	0.44	0.666	0.692	0.583	0.411	0.013	0.007	0.005	0.002	0.006	0.006
T20 [s]	4.121	0.864	0.981	1.015	0.88	0.841	0.835	0.713	0.506		0.723	0.721
r T20	0.871	0.981	0.984	0.993	0.992	0.994	0.992	0.991	0.981		0.992	0.984
T30 [s]	3.184	0.979	1.004	0.969	0.908	0.908	0.864	0.719	0.523	0.342	0.772	0.808
r T30	0.858	0.983	0.994	0.996	0.995	0.996	0.998	0.998	0.996	0.973	0.997	0.995
Tuser [s]		0.912	0.969	1.007	0.896	0.9	0.861	0.718	0.522	0.327	0.748	0.774
r Tuser		0.991	0.986	0.991	0.992	0.995	0.997	0.998	0.994	0.955	0.998	0.996
LE												
LF												
LFC												

RTU = RT User (10. dB, 30. dB)

Table G.5: *ISO-3382 acoustical parameters: Test environment 1, 1m from source, height = , Genelec 1029a, W channel.*

ISO 3382 OCTAVE BAND ACOUSTICALPARAMETERS

Freq. [Hz	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB	8.147	42.845	51.135	52.084	53.797	56.268	57.263	62.553	61.808	65.338	67.568	69.654
Noise [dB	5.758	12.763	2.406	0.664	2.281	0.339	1.542	3.515	3.953	2.751	6.2	14.826
strenGth [c	60.853	26.155	17.865	16.916	15.203	12.732	11.737	6.447	7.192	3.662	9.432	7.346
C50 [dB]	8.494	1.491	1.408	0.871	3.896	4.763	8.307	10.027	11.237	19.983	9.587	10.22
C80 [dB]	1.#IO	5.159	0.947	4.371	7.738	7.797	10.806	12.714	15.204	25.149	12.541	12.889
D50 [%]	87.608	41.5	41.963	54.996	71.036	74.965	87.132	90.961	93.005	99.006	90.092	91.318
Ts [ms]	149.203	95.542	95.311	68.071	47.111	44.011	29.451	23.53	19.38	13.138	23.705	22.068
EDT [s]	0.533	0.814	0.85	0.847	0.745	0.751	0.536	0.481	0.363	0.005	0.548	0.553
T20 [s]	0.193	0.927	1.044	0.753	0.896	0.899	0.908	0.749	0.533	0.329	0.772	0.756
r T20	0.869	0.988	0.986	0.983	0.996	0.999	0.995	0.996	0.998	0.967	0.996	0.998
T30 [s]	0	0.829	1.072	0.913	0.961	0.917	0.942	0.762	0.562	0.349	0.826	0.832
r T30	101	0.991	0.995	0.982	0.997	0.999	0.998	0.998	0.998	0.993	0.997	0.996
Tuser [s]	0.082	0.828	1.098	0.954	0.974	0.909	0.963	0.758	0.556	0.346	0.82	0.812
r Tuser	0.871	0.98	0.994	0.974	0.995	0.999	0.999	0.998	0.999	0.995	0.999	0.998
LE												
LF												
LFC												

RTU = RT User (10. dB, 30. dB)

Table G.6: *ISO-3382 acoustical parameters: Test environment 1, 2m from source, height = , Genelec 1029a, W channel.*

ISO 3382 OCTAVE BAND ACOUSTICALPARAMETERS

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB	49.823	40.162	46.466	50.055	50.427	53.802	53.789	58.838	58.077	59.46	63.766	65.292
Noise [dB	13.178	9.255	8.371	1.09	2.424	0.803	2.094	4.879	3.703	2.966	5.762	16.676
strenGth [c	118.823	28.838	22.534	18.945	18.573	15.198	15.211	10.162	10.923	9.54	13.234	11.708
C50 [dB]	1.#IO	2.012	1.451	1.718	2.109	3.14	4.99	6.898	8.568	16.392	6.674	7.353
C80 [dB]	1.#IO	1.375	2.77	5.268	3.766	6.571	7.099	8.878	12.23	20.955	9.077	9.786
D50 [%]	30.99	38.619	58.274	59.761	61.909	67.329	75.935	83.04	87.792	97.756	82.3	84.464
Ts [ms]	1212.038	108.049	83.286	71.159	64.309	47.961	37.527	27.88	19.668	6.49	27.968	25.126
EDT [s]		0.979	1.074	0.827	1.003	0.732	0.851	0.708	0.448	0.176	0.679	0.625
T20 [s]		0.956	1.14	0.985	0.923	0.894	0.855	0.726	0.557	0.337	0.767	0.795
r T20		0.985	0.982	0.989	0.996	0.995	0.997	0.998	0.999	0.991	0.999	0.998
T30 [s]		1.082	1.209	0.99	0.92	0.908	0.904	0.773	0.58	0.351	0.824	0.86
r T30		0.989	0.991	0.995	0.998	0.998	0.997	0.998	0.999	0.997	0.998	0.997
Tuser [s]		1.044	1.274	0.97	0.91	0.916	0.906	0.758	0.577	0.355	0.826	0.85
r Tuser		0.98	0.987	0.986	0.995	0.996	0.998	0.997	0.999	0.999	0.997	0.998
LE												
LF												
LFC												

RTU = RT User (10. dB, 30. dB)

Table G.7: *ISO-3382 acoustical parameters: Test environment 1, 4m from source, height = , Genelec 1029a, W channel.*

ISO 3382 OCTAVE BAND ACOUSTICALPARAMETERS

Freq. [Hz	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB	49.823	34.792	45.919	50.71	49.688	52.576	51.688	55.987	54.904	54.256	61.049	62.352
Noise [dB	8.032	5.755	6.202	0.198	1.886	0.696	1.601	5.284	5.839	0.591	5.576	13.943
strenGth [c	118.823	34.208	23.081	18.29	19.312	16.424	17.312	13.013	14.096	14.744	15.951	14.648
C50 [dB]	1.#IO	13.323	5.122	2.283	0.01	0.62	0.912	1.599	3.559	8.795	1.36	1.714
C80 [dB]	1.#IO	5.876	0.136	5.862	2.903	3.943	4.29	5.074	7.928	14.405	5.432	5.846
D50 [%]	191.116	4.446	23.515	37.155	49.944	46.44	44.769	59.101	69.414	88.342	57.767	59.74
Ts [ms]	2054.998	144.374	111.902	78.977	81.505	72.124	70.372	60.608	47.541	31.733	60.071	58.357
EDT [s]		0.927	1.006	0.653	0.907	0.774	0.786	0.773	0.51	0.271	0.724	0.691
T20 [s]		0.916	1.115	0.902	0.853	0.942	0.92	0.77	0.579	0.361	0.829	0.85
r T20		0.98	0.973	0.99	0.989	0.998	0.998	0.998	0.998	0.995	0.998	0.998
T30 [s]		1.019	1.037	1.035	0.92	0.918	0.915	0.79	0.591	0.364	0.848	0.879
r T30		0.988	0.989	0.987	0.994	0.999	0.999	0.999	0.999	0.998	0.999	0.999
Tuser [s]		1.081	0.914	1.083	0.9	0.924	0.914	0.805	0.593	0.366	0.865	0.892
r Tuser		0.99	0.989	0.989	0.981	0.997	0.998	0.999	0.998	0.997	0.999	0.999
LE												
LF												
LFC												

RTU = RT User (10. dB, 30. dB)

Table G.8: *ISO-3382 acoustical parameters: Test environment 1, 8m from source, height = , Genelec 1029a, W channel.*

G.0.3 Test Environment 1: ISM Simulation at Performance Distances

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB	51.03	58.595	56.437	56.875	57.171	60.014	63.829	64.122	62.301	65.782	70.351	71.131
Noise [dB	13.6	12.127	8.34	10.723	9.432	11.286	10.555	15.41	32.238	29.23	4.841	15.379
strenGth [c	17.97	10.405	12.563	12.125	11.829	8.986	5.171	4.878	6.699	3.218	6.649	5.869
C50 [dB]												
C80 [dB]												
D50 [%]	0.001	0	0	0	0	0	0	0	0	0	0	0
Ts [ms]	283.414	218.59	201.185	199.633	203.069	201.953	202.4	199.469	191.733	184.957	197.768	196.242
EDT [s]	1.6	0.454	0.438	0.334	0.648	0.759	0.627	0.629	0.54	0.473	0.61	0.584
T20 [s]	1.549	1.168	0.82	0.846	0.744	0.697	0.73	0.673	0.588	0.401	0.679	0.788
r T20	0.997	0.936	0.985	0.981	0.996	0.994	0.997	0.999	0.998	0.997	0.999	0.996
T30 [s]	1.546	1.335	0.896	0.875	0.838	0.72	0.75	0.701	0.575	0.407	0.715	0.885
r T30	0.999	0.978	0.993	0.994	0.992	0.997	0.999	0.999	0.999	0.998	0.999	0.991
Tuser [s]	1.56	1.292	0.918	0.918	0.831	0.676	0.766	0.682	0.588	0.404	0.708	0.833
r Tuser	0.998	0.955	0.986	0.996	0.996	0.996	0.999	0.998	0.998	0.997	0.999	0.999
LE												
LF												
LFC												

RTU = RT User (10. dB, 30. dB)

Table G.9: *ISO-3382 acoustical parameters: Test environment 1, 1m from source, height = 1.2, Directivity = Omnidirectional, W channel.*

ISO 3382 OCTAVE BAND ACOUSTICAL PARAMETERS

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB	34.95	48.466	54.187	57.668	60.825	62.749	66.818	69.22	68.028	69.337	74.551	74.607
Noise [dB	19.731	13.964	10.029	8.668	12.627	8.072	5.369	7.47	27.617	29.163	0.562	0.252
strenGth [c	34.05	20.534	14.813	11.332	8.175	6.251	2.182	0.22	0.972	0.337	2.449	2.393
C50 [dB]												
C80 [dB]												
D50 [%]	0	0	0	0	0	0	0	0	0	0	0	0
Ts [ms]	281.7	236.772	234.971	219.566	221.518	221.383	220.131	218.831	209.141	197.92	215.575	213.688
EDT [s]	1.3	0.823	0.766	0.559	0.744	0.727	0.611	0.673	0.609	0.449	0.638	0.627
T20 [s]	1.008	0.905	0.768	0.777	0.652	0.668	0.753	0.653	0.545	0.399	0.662	0.652
r T20	0.982	0.983	0.957	0.993	0.994	0.995	0.997	0.998	0.998	0.998	0.998	0.998
T30 [s]	1.054	0.956	0.866	0.82	0.717	0.702	0.784	0.714	0.567	0.396	0.715	0.707
r T30	0.992	0.992	0.982	0.994	0.992	0.997	0.998	0.996	0.998	0.999	0.997	0.996
Tuser [s]	1.062	0.934	0.922	0.798	0.699	0.697	0.784	0.697	0.559	0.4	0.703	0.698
r Tuser	0.981	0.981	0.967	0.994	0.994	0.994	0.999	0.998	0.998	0.999	0.999	0.999
LE												
LF												
LFC												

RTU = RT User (10. dB, 30. dB)

Table G.10: *ISO-3382 acoustical parameters: Test environment 1, 2m from source, height = 1.2, Directivity = Omnidirectional, W channel.*

ISO 3382 OCTAVE BAND ACOUSTICALPARAMETERS

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB	44.76	52.133	58.541	59.425	63.111	66.03	70.288	71.943	71.769	74.365	77.924	78.204
Noise [dB	22.468	21.939	2.638	0.012	8.333	6.951	4.394	5.157	25.452	27.273	2.125	24.715
strenGth [c	24.24	16.867	10.459	9.575	5.889	2.97	1.288	2.943	2.769	5.365	0.924	1.204
C50 [dB]												
C80 [dB]												
D50 [%]	0.039	0.005	0	0	0	0	0	0	0	0	0	0
Ts [ms]	363.283	289.846	239.658	240.027	229.108	229.737	232.184	229.01	221.364	211.842	225.866	223.523
EDT [s]	2.12	0.989	1.05	0.778	0.608	0.73	0.644	0.63	0.525	0.428	0.594	0.57
T20 [s]	2.168	1.667	0.811	0.921	0.662	0.719	0.773	0.675	0.54	0.394	0.686	0.679
r T20	0.998	0.988	0.982	0.99	0.99	0.995	0.996	0.999	0.999	0.998	0.999	0.999
T30 [s]	1.978	1.808	0.867	0.935	0.76	0.715	0.749	0.72	0.567	0.396	0.708	0.733
r T30	0.99	0.994	0.992	0.996	0.989	0.997	0.998	0.995	0.997	0.999	0.999	0.996
Tuser [s]	1.997	1.867	0.943	0.888	0.732	0.692	0.753	0.674	0.542	0.402	0.695	0.71
r Tuser	0.995	0.995	0.986	0.992	0.983	0.996	0.996	0.999	0.999	0.999	0.999	0.999
LE												
LF												
LFC												

RTU = RT User (10. dB, 30. dB)

Table G.11: *ISO-3382 acoustical parameters: Test environment 1, 4m from source, height = 1.2, Directivity = Omnidirectional, W channel.*

ISO 3382 OCTAVE BAND ACOUSTICALPARAMETERS

Freq. [Hz]	31.5	63	125	250	500	1000	2000	4000	8000	16000	A	Lin
Signal [dB	40.894	55.228	65.567	67.869	65.654	70.087	74.278	73.73	71.67	74.665	80.192	80.497
Noise [dB	19.036	0.777	8.518	4.895	1.773	3.101	4.744	1.928	19.536	29.294	9.122	12.605
strenGth [c	28.106	13.772	3.433	1.131	3.346	1.087	5.278	4.73	2.67	5.665	3.192	3.497
C50 [dB]												
C80 [dB]												
D50 [%]	0	0	0	0	0	0	0	0	0	0	0	0
Ts [ms]	275.587	278.266	266.713	235.829	249.983	248.294	245.585	241.407	235.373	226.552	241.407	239.704
EDT [s]	0.932	0.775	0.924	0.722	0.705	0.711	0.601	0.676	0.526	0.381	0.614	0.604
T20 [s]	0.829	1.066	0.841	0.805	0.798	0.676	0.768	0.694	0.6	0.409	0.709	0.711
r T20	0.99	0.953	0.979	0.994	0.993	0.991	0.997	0.998	0.997	0.999	0.998	0.998
T30 [s]	0.955	1.045	0.955	0.868	0.777	0.76	0.802	0.742	0.6	0.433	0.759	0.767
r T30	0.987	0.985	0.985	0.994	0.997	0.989	0.998	0.997	0.999	0.998	0.997	0.997
Tuser [s]	0.955	0.937	0.998	0.822	0.734	0.734	0.809	0.735	0.587	0.436	0.749	0.757
r Tuser	0.984	0.977	0.985	0.992	0.997	0.992	0.998	0.998	0.997	0.999	0.998	0.998
LE												
LF												
LFC												

RTU = RT User (10. dB, 30. dB)

Table G.12: *ISO-3382 acoustical parameters: Test environment 1, 8m from source, height = 1.2, Directivity = Omnidirectional, W channel.*

H

Measured Wavefields

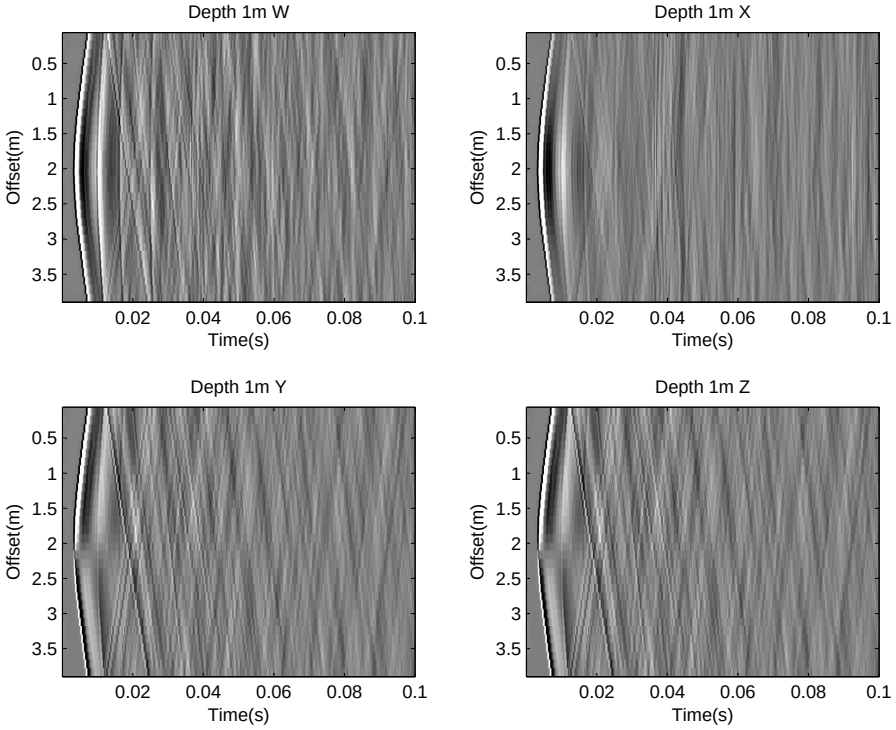


Figure H.1: *B-Format components of measured wavefield with source at 1m.*

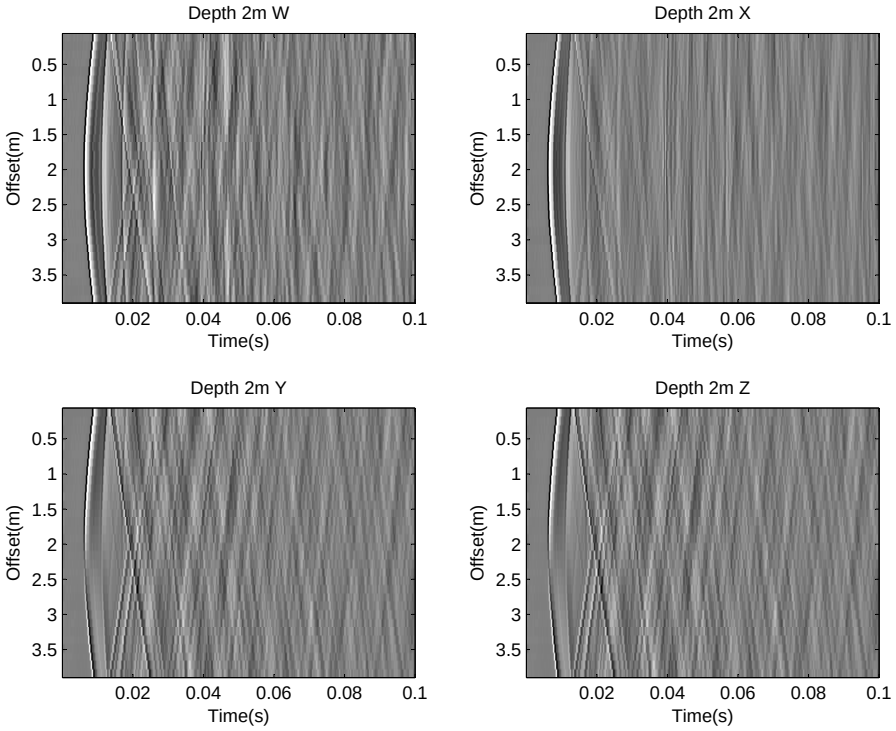


Figure H.2: *B-Format components of measured wavefield with source at 2m.*

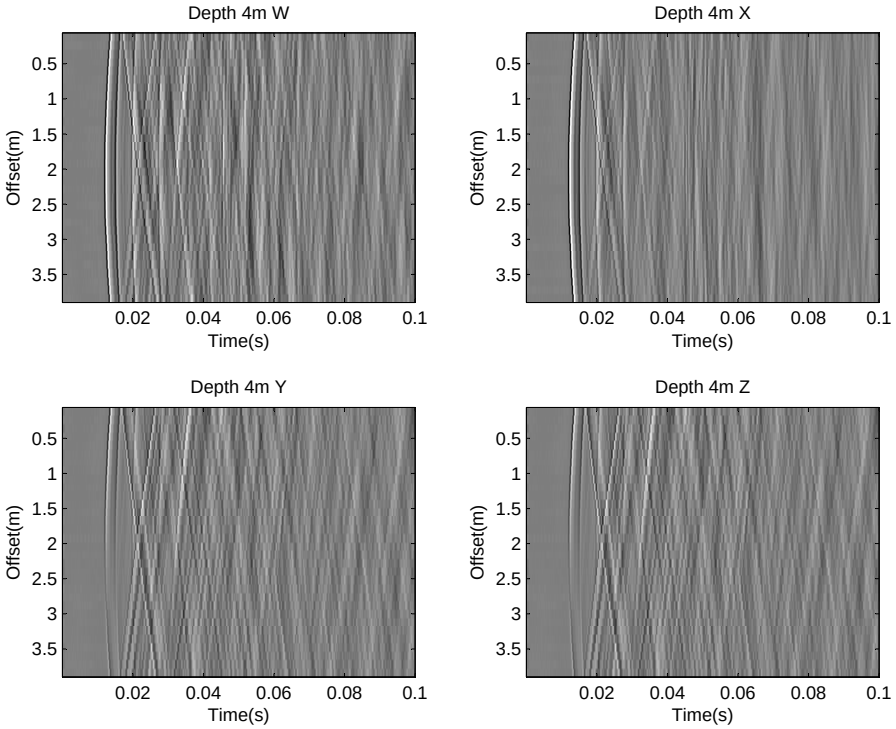


Figure H.3: *B-Format components of measured wavefield with source at 4m.*

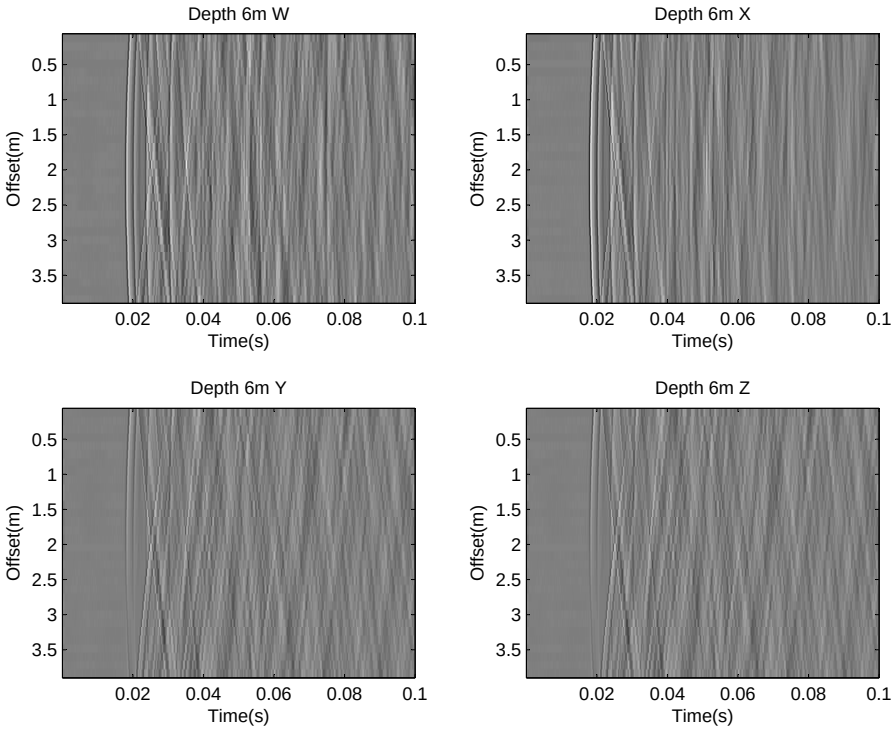


Figure H.4: *B-Format components of measured wavefield with source at 6m.*

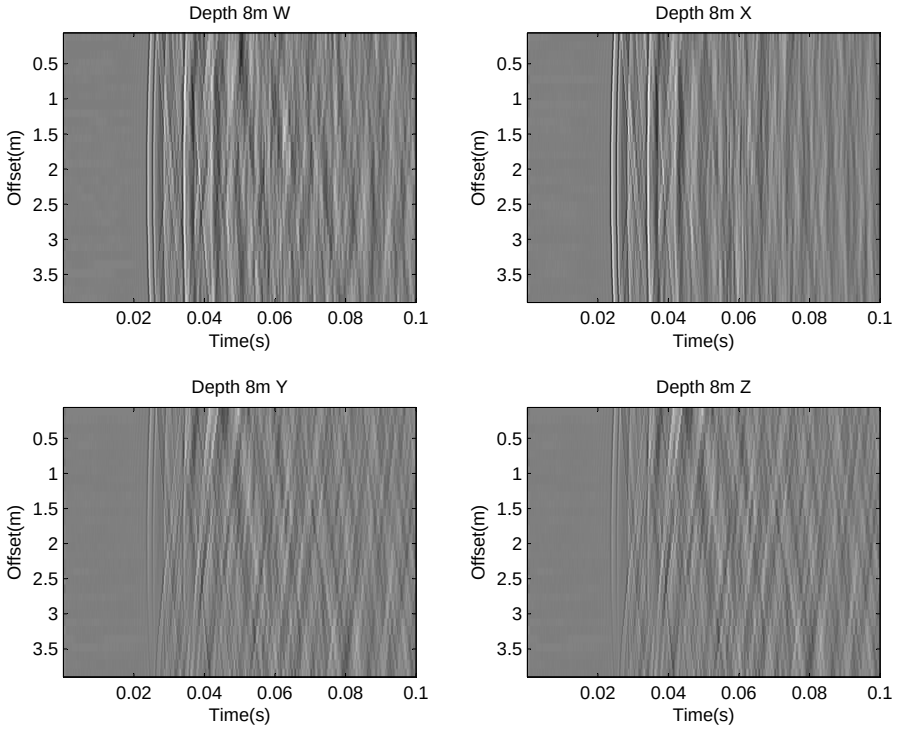


Figure H.5: *B-Format components of measured wavefield with source at 8m.*

I

DVD Contents

The DVD accompanying this thesis contains media that supports the thesis work. A html file, 'readme.html' contained in the root folder, allows for easy access to the material. All media is located in the 'data' folder, and the folders are organized according to the chapter names. The contents of the data disc are as follows:

Chapter 2: Spatial Hearing

Section 2.4.3 Localisation Experiment I: Evaluation of IACC-PHAT Through Perceived Localization in a Reverberant Environment

Male Speech Binaural recordings, 2 Channel, 44.1kHz, 16Bit

- Position 1, Loudspeaker 2
- Position 2, Loudspeaker 2
- Position 3, Loudspeaker 2
- Position 4, Loudspeaker 2
- Position 5, Loudspeaker 2
- Position 6, Loudspeaker 2
- Position 7, Loudspeaker 2
- Position 8, Loudspeaker 2
- Position 9, Loudspeaker 2
- Position 1, Loudspeaker 6
- Position 2, Loudspeaker 6
- Position 3, Loudspeaker 6
- Position 4, Loudspeaker 6
- Position 5, Loudspeaker 6
- Position 6, Loudspeaker 6
- Position 7, Loudspeaker 6
- Position 8, Loudspeaker 6
- Position 9, Loudspeaker 6
- Position 1, Loudspeaker 10
- Position 2, Loudspeaker 10
- Position 3, Loudspeaker 10
- Position 4, Loudspeaker 10
- Position 5, Loudspeaker 10
- Position 6, Loudspeaker 10
- Position 7, Loudspeaker 10
- Position 8, Loudspeaker 10
- Position 9, Loudspeaker 10

- Position 1, Loudspeaker 14
- Position 2, Loudspeaker 14
- Position 3, Loudspeaker 14
- Position 4, Loudspeaker 14
- Position 5, Loudspeaker 14
- Position 6, Loudspeaker 14
- Position 7, Loudspeaker 14
- Position 8, Loudspeaker 14
- Position 9, Loudspeaker 14

Chapter 4: A Comparative Study of Spatialization Techniques Under Real Listening Conditions

Section 4.2.4: Subjective Listening Experiment II.A: Localisation of VBAP and 1st Order Ambisonics Under Real Listening Conditions

Ambisonic Test Files, 8 Channel, 44.1kHz, 16Bit

- Female Speech, VS1
- Female Speech, VS2
- Female Speech, VS3
- Female Speech, VS4
- Male Speech, VS1
- Male Speech, VS1
- Male Speech, VS1
- Male Speech, VS1
- White Noise, VS1
- White Noise, VS1
- White Noise, VS1
- White Noise, VS1
- Music, VS1
- Music, VS1
- Music, VS1
- Music, VS1

VBAP Test Files, 8 Channel, 44.1kHz, 16Bit

- Female Speech, VS1
- Female Speech, VS2

- Female Speech, VS3
- Female Speech, VS4
- Male Speech, VS1
- Male Speech, VS1
- Male Speech, VS1
- Male Speech, VS1
- White Noise, VS1
- White Noise, VS1
- White Noise, VS1
- White Noise, VS1
- Music, VS1
- Music, VS1
- Music, VS1
- Music, VS1

Section 4.3.2: Subjective Listening Experiment III: Analysis of Higher Order Ambisonic Localisation

Higher Order Ambisonic Test Files, 1st Order, 8 Channel, 44.1kHz, 16Bit

- Basic Decode, Female Speech, VS1
- Basic Decode, Female Speech, VS2
- Basic Decode, Female Speech, VS3
- Basic Decode, Female Speech, VS4
- Basic Decode, Female Speech, VS1
- In Phase Decode, Female Speech, VS2
- In Phase Decode, Female Speech, VS3
- In Phase Decode, Female Speech, VS4
- Energy Decode, Female Speech, VS1
- Energy Decode, Female Speech, VS2
- Energy Decode, Female Speech, VS3
- Energy Decode, Female Speech, VS4
- Shelf Decode, Female Speech, VS1
- Shelf Decode, Female Speech, VS2
- Shelf Decode, Female Speech, VS3
- Shelf Decode, Female Speech, VS4

Higher Order Ambisonic Test Files, 2nd Order, 8 Channel, 44.1kHz, 16Bit

- Basic Decode, Female Speech, VS1
- Basic Decode, Female Speech, VS2
- Basic Decode, Female Speech, VS3
- Basic Decode, Female Speech, VS4
- Basic Decode, Female Speech, VS1
- In Phase Decode, Female Speech, VS2
- In Phase Decode, Female Speech, VS3
- In Phase Decode, Female Speech, VS4
- Energy Decode, Female Speech, VS1
- Energy Decode, Female Speech, VS2
- Energy Decode, Female Speech, VS3
- Energy Decode, Female Speech, VS4
- Shelf Decode, Female Speech, VS1
- Shelf Decode, Female Speech, VS2
- Shelf Decode, Female Speech, VS3
- Shelf Decode, Female Speech, VS4

Higher Order Ambisonic Test Files, 3rd Order, 8 Channel, 44.1kHz, 16Bit

- Basic Decode, Female Speech, VS1
- Basic Decode, Female Speech, VS2
- Basic Decode, Female Speech, VS3
- Basic Decode, Female Speech, VS4
- Basic Decode, Female Speech, VS1
- In Phase Decode, Female Speech, VS2
- In Phase Decode, Female Speech, VS3
- In Phase Decode, Female Speech, VS4
- Energy Decode, Female Speech, VS1
- Energy Decode, Female Speech, VS2
- Energy Decode, Female Speech, VS3
- Energy Decode, Female Speech, VS4
- Shelf Decode, Female Speech, VS1
- Shelf Decode, Female Speech, VS2
- Shelf Decode, Female Speech, VS3
- Shelf Decode, Female Speech, VS4

Chapter 5: Virtual Acoustic Recording

Section 5.3.3.2 Perceptual Effects of Coloration in the Convolution Chain

Trinity College Chapel Choir, Sicut Cervus, 44.1kHz, 24Bit

- Actual Choir performance, Ambisonics, 7.1 (Front High)
- Virtual Choir performance, Uncompensated, Ambisonics, 7.1 (Front High)
- Virtual Choir performance, Compensated, Ambisonics, 7.1 (Front High)
- Actual Choir performance, Hamasaki Square, 4.0
- Virtual Choir performance, Uncompensated, Hamasaki Square, 4.0
- Virtual Choir performance, Compensated, Hamasaki Square, 4.0
- Actual Choir performance, OCT 7 Array, 7.1 (Front High)
- Virtual Choir performance, Uncompensated, OCT 7 Array, 7.1 (Front High)
- Virtual Choir performance, Compensated,OCT 7 Array, 7.1 (Front High)

Section 5.3.3 Equalization of the Convolution Chain

Direct Field Recordings, 96kHz, 24Bit

- Direct Field Female Vox, 1m, Uncompensated
- Direct Field Female Vox, 1m, Compensated
- Direct Field Female Vox, 2m, Uncompensated
- Direct Field Female Vox, 2m, Compensated
- Direct Field Female Vox, 4m, Uncompensated
- Direct Field Female Vox, 4m, Compensated
- Direct Field Female Vox, 8m, Uncompensated
- Direct Field Female Vox, 8m, Compensated
- Direct Field Violin, 1m, Uncompensated
- Direct Field Violin, 1m, Compensated
- Direct Field Violin, 2m, Uncompensated
- Direct Field Violin, 2m, Compensated
- Direct Field Violin, 4m, Uncompensated
- Direct Field Violin, 4m, Compensated
- Direct Field Violin, 8m, Uncompensated
- Direct Field Violin, 8m, Compensated

Section 5.4.2 Binaural Comparison of Reproduced Recordings

Binaural Recordings, 96kHz, 24Bit

- Binaural, Female Vox, 1m, Actual
- Binaural, Female Vox, 1m, Virtual
- Binaural, Female Vox, 2m, Actual
- Binaural, Female Vox, 2m, Virtual
- Binaural, Female Vox, 4m, Actual
- Binaural, Female Vox, 4m, Virtual
- Binaural, Female Vox, 8m, Actual
- Binaural, Female Vox, 8m, Virtual
- Binaural, Violin, 1m, Actual
- Binaural, Violin, 1m, Virtual
- Binaural, Violin, 2m, Actual
- Binaural, Violin, 2m, Virtual
- Binaural, Violin, 4m, Actual
- Binaural, Violin, 4m, Virtual
- Binaural, Violin, 8m, Actual
- Binaural, Violin, 8m, Virtual

Section 5.4.3 Subjective Experiment IV.A: Subjective Attributes of Actual Vs. Virtual Convolution Based Recordings

Custom test software for derivation of subjective attribute descriptions.

5.1 Comparison of SRIR Convolution

5.5 Directivity Enhancement for Virtual Acoustic Recordings

Source simulation radiator auralization comparison, 5.1, 44.1kHz, 16 bit.

- Violin, 1m, Non-SSR render
- Violin, 1m, SSR render
- Violin, 2m, Non-SSR render
- Violin, 2m, SSR render
- Violin, 4m, Non-SSR render
- Violin, 4m, SSR render
- Violin, 8m, Non-SSR render

- Violin, 8m, SSR render

Chapter 7: Virtual Acoustic Reproduction

Section 7.2.2 Factorization Analysis using the CIPIC Dataset

Binaural, Stereo Virtual Loudspeakers, 44.1kHz, 16 bit.

- 180 long angle independent component, 20 long angle dependent component: Full convolution
- 180 long angle independent component only.
- 20 long angle dependent components only.
- 190 long angle independent component, 20 long angle dependent component: Full convolution
- 190 long angle independent component only.
- 20 long angle dependent components only.

Supplementary Material

Further material supporting the thesis work, including SRIR measurements, listening test software, test environment 1 models and equipment measurements can be found in the relevant chapter folders under 'Data'.