Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

Optimised Real-time Rendering of Auditory Events in Immersive Virtual Environments

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

Marcin Gorzel
Trinity College Dublin, May 2014

Signal Processing and Media Applications
Department of Electronic and Electrical Engineering
Trinity College Dublin

To my family.

Abstract

This project looks at the problem of the capture or synthesis of acoustic events in reverberant spaces and their subsequent plausible reproduction in a virtual version of the original space, otherwise known as a Virtual Auditory Environment (VAE). Of particular concern is identification and perceptually accurate reconstruction of the important acoustic cues that enable us to localise sound sources in the whole 3-D space, with a special emphasis on the perception of auditory distance.

Such an auditory presentation can be realised with the use of both multichannel loudspeaker arrays and headphones. The latter are able to provide a personalised sound field to a single user, minimising the influence of the listening environment and providing a better sense of immersion. However, one of the problems that needs to be addressed is the user interaction and how listener movements affect the experience. Such walk-through auralisations present several challenges for production engineers, most significant of which are the generation of correct room acoustic responses for a given source-listener position as well as identification of the most optimal sound reproduction schemes that can minimise the computational burden.

A framework is proposed that considers the parametrisation of real-world sound fields and their subsequent real-time auralisation using a hybrid image source model/measurement-based approach. Two different test models are constructed based on existing spaces with significantly different acoustic properties: a middle sized lecture hall and a large cathedral interior. Various optimisation techniques, including order reduction of Head Related Transfer Function using approximate factorisation and Room Impulse Response decomposition using directional analysis and diffuseness estimation are incorporated. The subjective evaluation shows that the technique proposed could be successfully used in order to adapt the convolution based auralisation to real-time and interactive scenarios, including virtual reality applications or video games.

Since the method proposed enables re-rendering of the directional components of sound fields using higher spatial resolution, tests have been conducted whether increasing the resolution may contribute to a better perception of sound source distance. Results show that it is not the case and correct perception of distance can be assured even at low spatial reconstruction orders. However, since spatial localisation of sounding objects is affected not only by the auditory cues but also by other modalities such as vision, further user studies are performed which show the effect of incongruent audio-visual cues on photo-realistic VAEs presentations.

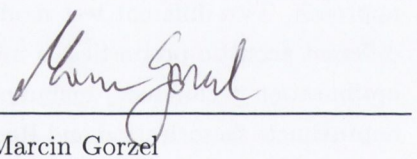
Lastly, to account for existing loudspeaker based auralisations (e.g. in a '5.1. surround sound' format), a novel approach to binaural mix down using sound field stabilisation with head-tracking is presented. Objective analysis shows that better localisation accuracy can be expected than in the case of other similar methods.

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,



Marcin Gorzel

May 28, 2014.

Acknowledgments

There are many people that made my research possible, who gave me their great support, advice and constructive comments. I am forever grateful to:

Prof. Frank Boland, my supervisor - thanks to you I was able to pursue the topics in the field of spatial audio that interest me most. I really appreciate your great guidance, advice, motivation and expertise you have been sharing with me over the last three years.

Dr. Fionnuala Conway - I will never forget that you helped me get started and guided me through the beginnings of this road. I really appreciate all your advice and financial support that you gave me.

Dr. Gavin Kearney, currently University of York - thanks for the great amount of time we spent together working on numerous projects, but in particular, for your constant inspiration, motivation and incredible knowledge of spatial audio you shared with me - I was so lucky to be able to work with you!

Dr. Dermot Furlong - I am the most grateful for every kind of support you offered me at so many occasions. Thank you for allowing me to be a part of your great achievement which the Music and Media Technologies Course certainly is.

Prof. Anil Kokaram, currently Google Inc., for the great inspiration you gave me as well as for many opportunities to present and share with others what I do. Also, I am very grateful to Google Inc. for the financial assistance they kindly offered to help me proceed with this project.

Prof. Henry Rice, from the Department of Mechanical Engineering, for giving me great support and opportunity to collaborate on the Science Foundation's Metropolis Project and for making it possible for me to present my research at numerous occasions.

My current and former colleagues, collaborators and friends from the Sigmedia group, especially: Dr. Claire Masterson, Dr. Darren Kavanagh, Dr. David Corrigan, Dr. François Pitié, Dr. Gary Baugh, Ian Joseph Kelly and Brian O'Toole - you guys are best of the best!

Present and former staff and students in the Music and Media Technologies and Interactive Digital Media courses and in particular, those students I had this great pleasure to supervise: Tom Martin, Marc Balbirnie, Tracy Nagle, Antonio Pedro Santos and Ronan Corrigan.

Natasa Paterson - you are not only the most talented composer, singer and cellist, but also the best friend, advisor and motivator. Thank you for sharing so many ideas and so many constructive discussions.

Liam O'Sullivan - I learn something from you every day and... I will definitely remember the "pop quiz" and your "dzień dobry" every morning, probably until the end of my days!

People in the Audio Lab, University of York, in particular Prof. David Howard, Dr. Damian Murphy, Jude Brereton, Aglaia Foteinou and Andrew Chadwick as well as the most helpful staff in the department of Theatre Film & Television - thank you for all your help that made the experiments in this thesis possible, as well as for your constructive comments.

John Squires - for all your incredible skills in so many different fields and always being happy to help.

Staff in the Department of Electronic and Electrical Engineering: Conor Nolan, Robbie Dempsey, Shane Hunt, Jenny Kirkwood, Nora Moore and Bernie - without you everyday life in college would not be possible.

My futsal team! I know, I was busy over the past few months... but now I am coming back!

Those I did not name, but still remember about - especially, those who devoted their time and effort to participate in my listening tests.

I am forever indebted to my family: Grażyna, Grzegorz, Paweł, Daniela, Eugeniusz, Agnieszka, Arkadiusz and Oliwia - every day you give me power, motivation and love that I need. And although we are all in different parts of the world, I know that you are always close.

Finally, and the most importantly, the most grateful feelings I direct to my beloved wife, Karolina. You are the one who suffered the most from my (mental) absence and at the same time who gave the support, strength and understanding I so much needed. I do not even attempt to put in words what do I owe to you, because no such words exist. You are inseparable part of my life.

Contents

List of Acronyms

ANOVA ANalysis Of VAriance

AEP Ambisonic Equivalent Panning

ASW Apparent Source Width

BGE Blender Game Engine

BRIR Binaural Room Impulse Response

CG Computer Graphics

FOA First Order Ambisonics

GPU Graphics Processing Unit

GUI Graphical User Interface

HOA Higher Order Ambisonics

HRIR Head Related Impulse Rresponse

HRTF Head Related Transfer Function

HSD Honestly Significant Difference

IACC Inter Aural Cross Correlation Coefficient

IACF Inter Aural Cross Correlation Function

IMU Inertial Measuring Unit

ORTF Office de Radiodiffusion Télévision Française

PCPP Pairwise Constant Power Panning

RIR Room Impulse Response

RT_{60} Reverberation Time

SOA Second Order Ambisonics

SRIR Spatial Room Impulse Response

TOA Third Order Ambisonics

VAE Virtual Auditory Environment

VBAP Vector Base Amplitude Panning

WFS Wave Field Synthesis

1 Introduction

Methods for delivering realistic spatial sound presentations have been a subject of research over the last few decades. With the emergence of modern computing technology as well as advancements in digital signal processing and the theory of spatial hearing, it is now becoming possible to re-create real acoustic events in virtual settings with high levels of faithfulness and immersion. Therefore, spatial audio technologies have their presence in a multitude of applications, most notably in film, music and video games. However, the most promising, but at the same time the most challenging uses, are in the real-time and interactive virtual reality e.g. video games, "serious games" (for example, simulations created for training, scientific research or guided tour purposes etc.) [12] and immersive audio-visual installations. Creation of such Virtual Auditory Environments (VAEs) often invoke the term *auralisation* which stems from its visual counterpart - *visualisation*.

For example, a real-time walk-through auralisation can refer to the process of recording and subsequent faithful reconstruction of some acoustic event in the virtual settings, so that the user can explore and interact with it. The element of interactivity and randomness, as in video games, is one of the biggest challenges here, since due to non-linearity[1] of auditory presentations, subsequent acoustic conditions cannot be easily predicted. This fact necessitates the use of efficient rendering techniques that allow for rapid re-calculation of the reproduced signals so that the current listener-source(s)-environment states are reflected.

The complete acoustic path from a sound source to a listener can be modelled as a trans-

[1]Here "non-linearity" refers to the manner of the sequential presentation of the audio-video content in video game playing.

mission through numerous transfer functions pertaining to the particular stages of the signal propagation. The transfer functions are responsible for: 1) shaping of the source radiation pattern (i.e. the way in which acoustic energy is emitted in different spatial directions); 2) sound interaction with boundaries of the enclosed space which is commonly referred to as the Room Impulse Response (RIR); 3) depending on the reproduction method, sound interaction with the physical features of humans body, which is commonly referred to as Head Related Transfer Function. In the interactive auralisations, the main challenge is that, although each of these stages can be treated as linear, at the same time they are time-varying (LTV). Thus, the interactive auditory scenes require that the resultant transfer function is constantly updated in order to accommodate for all these changes.

Naturally sounding results can be achieved, for example, with the convolution reverberation technique. In this technique the acoustic "fingerprint" of a particular space, which is called its acoustic Room Impulse Response (RIR), is measured and then used as a filter in the process of auralisation. The source audio material is usually recorded under anechoic conditions or at least in the direct field so that the amount of ambient sound is minimised. Also, directivity filters are used in order to recreate the frequency dependent radiation pattern belonging to a particular source. Then, after filtering with the RIR, sound is spatialised using arrays of loudspeakers or headphones. This situation is illustrated in Figure 1.1.

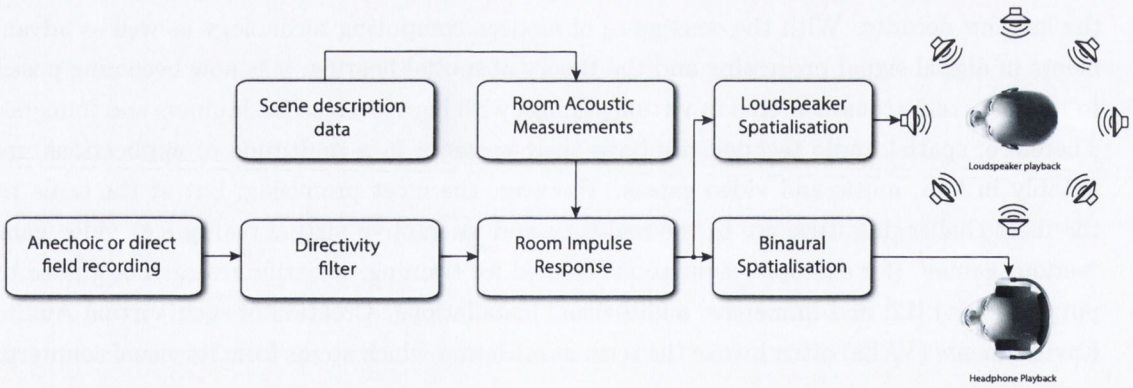


Figure 1.1: Block diagram showing consecutive stages in the database auralisation process

However, one of the problems with this method is that acoustic behaviour varies quite dramatically whenever the spatial configuration of the radiating source changes with respect to the listener. In order to accommodate for multiple listening perspectives, a large dataset of these "fingerprints" needs to be collected. This methodology is often referred to as *database auralisation* [105].

However, the measurement process is often laborious, tedious, and prone to error. For these reasons, techniques of approximation and interpolation of acoustic measurements have been studied [135]. Nevertheless, even with a densely defined grid of different acoustic responses,

there are still problems in terms of the information storage and the processing power required to apply these changes in real-time.

On the other hand, state-of-the-art numerical simulation techniques not only allow for computation of acoustic responses belonging to existing places but also facilitate prediction of the acoustic behaviour in conceived spaces. In general, methods which make use of the synthesised RIRs are often referred to as *computational auralisation*. The results of such simulations can be very accurate and, in some cases, may mitigate the needs of physical measurements altogether. For example, highly accurate results can be achieved using so-called wave-based methods, such as the Finite Element Method (FEM), Boundary Element Method (BEM) [176] or Finite-Difference Time-Domain (FDTD) method [200]. Sadly, the complexity and the amount of mathematical operations involved is usually so vast that only off-line processing is feasible. Naturally, it is still possible to render a database of pre-defined acoustic responses and proceed as in the case of the database auralisation. However, in such a situation many of the aforementioned problems still hold and the solution cannot be considered as optimal.

So far, real-time auralisations using numerical acoustic simulations have been shown to be feasible only for small room volumes and limited frequency ranges [201]. To make the computations faster, often highly parallel processor architectures are employed, such as in the Graphics Processing Units (GPU), instead of Central Processing Units (CPU). Considering the exponential growth in terms of computing power capabilities, this fact definitely gives some hope that in the future numerically based real-time acoustic simulations will be amongst the most desirable methods for real-time walk-through auralisations. In the meantime however, we shall still seek more efficient solutions to the problem.

Acoustic responses can be also computed using less complex geometrical approaches which treat acoustic waves as light rays. Image Source Method (ISM) [8] or ray-tracing are just two examples. These methods are well suited to the mid- to high-frequency region simulations, although they do not consider phenomena such as diffraction or scattering. As the result, they offer inferior accuracy than two previous approaches. A block diagram showing the principles of computational auralisation is shown in Figure 1.2.

Lastly, there is also a third option, where a hybrid system makes use of both measured and computed acoustic responses. However, hybrid approaches require thorough psychoacoustic analysis of the RIR structure in order to determine best areas of operation for both paradigms. So far, they have been largely focused on the synthesis of the diffuse decay as opposed to early parts of the RIRs [33,214] and other ways still seem to be relatively unexplored. A block diagram for such a system is illustrated in Figure 1.3.

1.1 Focus of this Thesis: Perceptually Equivalent Auralisation

This thesis investigates possible solutions to the real-time auralisation problem, and in particular, methods that are perceptually driven and which would allow for highly convincing results

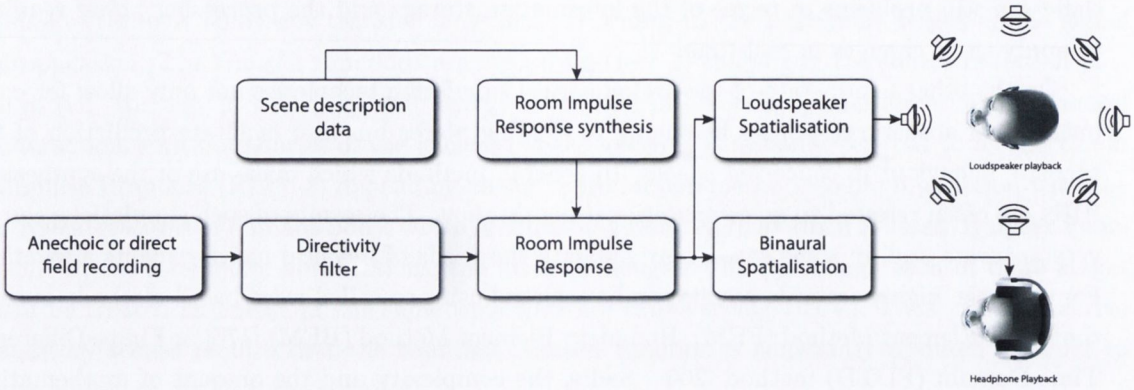


Figure 1.2: Block diagram showing consecutive stages in the computational auralisation process

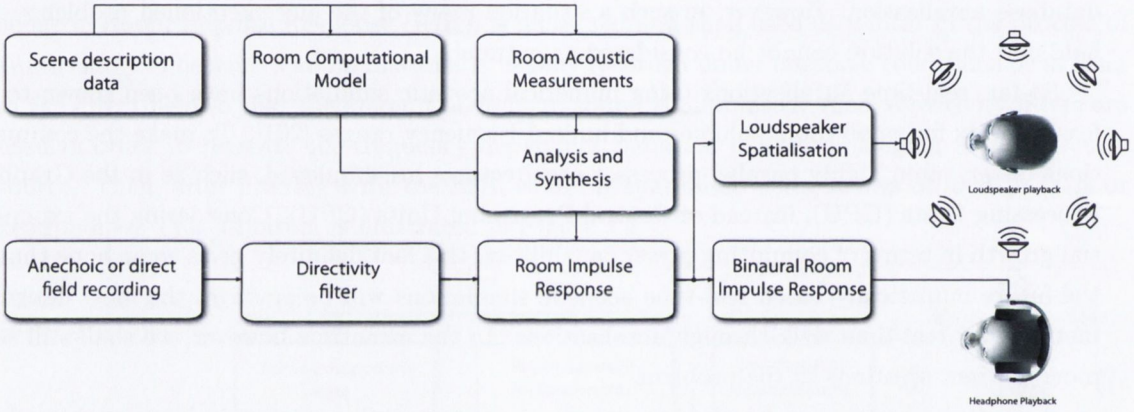


Figure 1.3: Block diagram showing consecutive stages of the auralisation process

whilst maintaining a low overall implementation cost. As a possible solution, it is proposed that measured RIRs are first analysed in terms of their directional content. Then, the diffuse component can be pre-filtered with the source material off-line. On the other hand, the directional components can be synthesised at run-time using some highly efficient computational model. The psychoacoustic justification to the above is that the room acoustic properties can be considered in terms of their directional and diffuse properties. The diffuse field is statistically homogeneous across the whole space and does not carry important directional information. That is why it does not need to be constantly updated nor rendered with high directional accuracy. In contrast, the directional components are crucial to the process of directional discrimination of sound sources and they require an optimal method of spatialisation.

In order for the above to be properly addressed, several stages to this investigation are suggested. First of all, it is necessary that human perceptual abilities and limitations with regards to spatial hearing are carefully studied and used as a psychoacoustical scaffold for all the technical aspects in this work. On the reproduction side then, it is required that the system

is capable of reproducing vital acoustic cues responsible for providing unambiguous information about the source's spatial location.

So, secondly, a review of the existing audio rendering techniques should be performed. Many techniques have been proposed in literature that offer different, or even variable, levels of acoustic fidelity for a single user or distributed audiences. Some of them differ significantly from the point of view of their underlying principles and assumptions so that not all of them are equally suited for the given task. In this quest, of a particular interest is the identification of systems that allow for controlled sound field degradation in terms of its directional resolution so that different components of the RIRs can be rendered with different accuracy.

The outcomes of these investigations should be used in the development of a novel approach to a hybrid database/computational auralisation with the use of some practical examples. These examples shall explain the implementation aspects of Virtual Auditory Environments as well as evaluate its success (or lack thereof). Two demonstrations will be presented and will be centred around the idea of the virtual acoustic recording, where a real-world acoustic event is recorded and subsequently *virtualised* so that the listener can explore and interact with the acoustic space.

Of note, a convincing spatial audio presentation requires not only that the sufficient directional information is provided but also that the strong sensation of auditory distance is retained. It is however unclear how the perception of distance is affected when the directional resolution of a sound field is degraded. This area seems to be still relatively poorly understood as compared to other aspects of spatial hearing, like the directional discrimination. It is therefore required to perform a thorough investigation into auditory distance perception, and in particular, the effect of directional sound field accuracy on the evoked sensations of distance. In this regard, identification of perceptual equivalence between the real and virtual acoustic environments would constitute a significant contribution to knowledge in the field. These investigations should be performed for both loudspeaker and headphone listening. The latter mode is particularly relevant in the context of this work since it can be used to effectively deliver personalised audio to the listener. However, some problems need to be addressed when it comes to the real-time headphone auralisation. These are mainly concerned with the required Head Related Impulse Response dataset sizes, amount of necessary filtering and finally, lengths of the individual filters.

Lastly, not all already existing auralisations are designed for headphone listening. This is true for a myriad of so-called surround sound material (using popular 5.1 or 7.1 formats) including interactive sound tracks of video games. In some of the cases, like in auralisations for video games, we do not have access to individual audio assets but only to a multichannel loudspeaker signal. However, in-game headphone auralisations still can be improved. For this reason, a binaural mix down is often used. In this case, it is also crucial to account for user's head movements so that the sound sources in the sound field retain their original spatial locations. This can be assured by employing a head-tracking device that constantly monitors the current orientation of the head. Then, one of the various sound field transformation strategies can be implemented in order to counteract these movements. In this work, a unified solution is proposed

that allows for various strategies to sound field rotation designated for arbitrary loudspeaker configurations. These strategies are first reviewed and analysed from the point of view of the objective localisation criteria as applied to the rotated sound field. Then, a novel approach to applying head-tracking to one of the most popular 5.1 surround sound layouts is offered.

Therefore, with all the above in mind, the main hypothesis can be formulated as follows:

> **Convolution-based auralisation can be adapted and efficiently implemented in the context of real-time multi-source and dynamic auralisation, so that the perceptual effect is enhanced whilst maintaining a low overall computational cost.**

Due to a vast amount of possible approaches and solutions to this problem, it is finally useful to define certain assumptions and restrictions that can set the application target to this work. First of all, we are going to assume a personalised presentation, as opposed to multi-user or distributed audience scenarios. Such situations are very common in video game playing experience or training and simulation in the virtual reality context. Only these two targets already constitute an enormous field of modern multimedia applications and fully justify the importance of investigations into high fidelity real-time audio. Due to the fact that for such applications, both loudspeaker playback systems as well as personalised binaural rendering are applicable, we shall consider both modes in the appropriate level of detail.

1.2 Thesis outline

The previous part of this chapter set the context of this work and defined its primary areas of focus. The remainder of this thesis is organised as follows.

Chapter 2: Spatial Perception of Sound

This chapter describes the fundamentals of spatial hearing. Mechanisms for the perception of auditory source direction, distance and motion are discussed. A review of research on the perception of spatial sound is presented. This chapter forms a perceptual scaffold for the work presented in subsequent chapters.

Chapter 3: Spatial Reproduction of Sound Fields

Here, spatial audio reproduction methodologies are reviewed. These methodologies are discussed from the point of view of their underlying principles and how they address the problem of constructing auditory localisation cues. Particular focus is on the suitability of different spatialisation schemes to real-time walk-through auralisations as far as a single user is concerned. Therefore, a significant amount of space is devoted to headphone reproduction as a means providing personalised listening experience.

Chapter 4: Real-Time Walk-Through Auralisations

Here, the proposed solution to the real-time interactive auralisation problem is proposed. Two applications of perceptually optimised rendering schemes for VAEs are presented, both dealing with the problem of acoustic recording and subsequent auralisation with the use of virtual acoustics. The first example is a traditional Irish trio captured in a medium size reverberant hall. The second example is the performance of a choir in the Christ Church Cathedral in Dublin, which is one of the most valuable (architecturally and acoustically) spaces in Ireland.

The work presented in this chapter utilises a number of tools facilitating the real-time rendering of auditory scenes discussed in previous chapters. These include the use of techniques for decomposing RIRs and HRIRs into the convolution of their direction dependent and direction independent components; a method for designing hybrid reverberators that utilises both measurement-based and computation-based techniques; decomposition of sound source spatial radiation patterns using spherical harmonic basis functions; optimised binaural reproduction utilising a limited number of filters for all-around reproduction.

Chapter 5: Perception of Auditory Distance under Real and Virtual Conditions

This experimental part of work looks further at the evaluation of the VAEs and more precisely, at the problem of the perception of auditory distance in VAEs as compared to real environments. Virtual acoustic conditions are simulated using original sound field decomposition with different level of accuracy. In this way, the impact of directional resolution in auditory scenes on the perceived distance is tested and compared against the real-world scenarios. The experiments are conducted in the context of both loudspeaker and headphone listening. Lastly, the effect of visual cues on distance judgements is also evaluated.

Chapter 6: Sound Field Stbilisation

In this chapter, sound field rotation is discussed as a method of stabilisation of virtual loudspeaker sound fields, so that user head movements are accounted for. Unification of the Ambisonic approach (where a sound field is regarded as a full representation of all the sources in the scene) and various panning laws is offered so that the rotation can be applied to arbitrary multichannel audio formats and subsequently rendered over headphones. Formal objective analysis of existing panning algorithms is made where objective localisation criteria are used as predictors of perceived localisation accuracy in different areas of the rotated sound field. Finally, a novel approach utilising the direct non-linear optimisation is presented alongside the analysis of the results obtained.

Chapter 7: Conclusions

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

1.3 Contributions of this thesis

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

- A novel, perceptually optimised framework for creating Virtual Auditory Environments using a hybrid measurement-computation based engine for real-time generation of Spatial and Binaural Room Impulse Responses
- Virtual audio-visual walk-through demonstration based on the traditional Irish trio recording in the medium size reverberant hall (Printing House Hall, Trinity College Dublin, Ireland)
- Virtual audio-visual walk-through demonstration based on the choir performance in the Christ Church Cathedral in Dublin, Ireland
- An investigation into perceptual effects of HRIR simplification and its application to real-time audio
- An investigation into perception of auditory distance in First and Higher Order Ambisonic sound fields presented over head-tracked headphones
- An investigation into perception of auditory distance in First and Higher Order Ambisonic sound fields presented over the periphrastic loudspeaker array
- An investigation into the impact of visual cues on the tolerance of audio-visual incongruence in the context of the perception of distance
- A novel algorithm for performing rotations of arbitrary horizontal sound fields with an example based on the popular *ITU 5.1* loudspeaker layout.

1.4 Publications

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

Journal publications:

- J1:"Distance Perception in Interactive Virtual Acoustic Environments using First and Higher Order Ambisonic Sound Fields" by Gavin Kearney, Marcin Gorzel, Henry Rice and Frank Boland, in *Acta Acustica united with Acustica*, Volume 98, Number 1, pp. 61-71(11), January/February 2012

- J2: "HRIR Order Reduction using Approximate Factorisation" by Claire Masterson, Gavin Kearney, Marcin Gorzel and Frank Boland in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1808 - 1817, 2012

Conference publications:

- C1: "On Loudspeaker Rendering of Auditory Distance in Higher Order Ambisonics" by Gavin Kearney, Marcin Gorzel and Frank Boland, *in Proceedings of the Acoustics 2012 Conference*, Nantes, France, April 2012

- C2: "Distance Perception in Virtual Audio-Visual Environments" by Marcin Gorzel, David Corrigan, Gavin Kearney, John Squires and Frank Boland, in *Proceedings of the 25th Audio Engineering Society UK Conference*, York , UK, March 2012

- C3: "Real-Time Walkthrough Auralisation of the Acoustics of Christ Church Cathedral, Dublin" by Gavin Kearney, Marcin Gorzel, Fiona Smyth, Frank Boland, Henry Rice and Donal Lennon , in *Proceedings of the Institute of Acoustics: Auditorium Acoustics Conference*, Dublin, Ireland, May, 2011

- C4: "On the Perception of Dynamic Sound Sources in Ambisonic Binaural Renderings" by Marcin Gorzel, Gavin Kearney, Henry Rice and Frank Boland, *in Proceedings of the 41st Audio Engineering Society Conference on Game Audio*, London, UK, February, 2011

- C5: "Application of HRIR Factorisation to Game Audio" by Gavin Kearney, Marcin Gorzel, Henry Rice and Frank Boland, in *Proceedings of the 41st Audio Engineering Society Conference on Game Audio*, London, UK, February, 2011

- C6: "A Video Database for the Development of Stereo-3D Post-Production Algorithms" by D. Corrigan, F. Pitié, V. Morris, A. Rankin, M. Linnane, G. Kearney, M. Gorzel, M. O'Dea, C. Lee and A. Kokaram, in Proceedings of the European Conference on Visual Media Production (CVMP '10), London, UK, November, 2010

- C7: "Virtual Acoustic Recording: An Interactive Approach" by Marcin Gorzel, Gavin Kearney, Frank Boland and Henry Rice, in *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx10)*, Graz, Austria, September, 2010

- C8: "Optimised Virtual Loudspeaker Reproduction" by Claire Masterson, Gavin Kearney, Marcin Gorzel, Henry Rice and Frank Boland, in *Proceedings of the Irish Signals and Systems (ISSC) Conference*, Cork, Ireland, June 2010

- C9: "Depth Perception in Interactive Virtual Acoustic Environments using Higher Order Ambisonic Soundfields" by Gavin Kearney, Marcin Gorzel, Frank Boland, Henry Rice, in *Proceedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, Paris, France, May, 2010

Other contributions and poster presentations:

- "Real and Virtual Acoustic Recording for Interactive Walkthrough Auralization", Gavin Kearney, Marcin Gorzel and Frank Boland, Virtualisation and Heritage Synposium, University of York, Uk, February, 2012.
- "Web-based Tutorial for Ambisonics Theory", Marcin Gorzel, Gavin Kearney, Frank Boland, Irish Sound Science and Technology Association Convocation, Limerick, Ireland, August, 2011
- "Optimization of Ambisonic Renderings for Virtual Reality Applications", Marcin Gorzel, Second Irish Workshop on Music and Audio Signal Processing, Trinity College Dublin, Ireland, 13th January, 2010

In the journal publication 1 (J1) the author participated in the experimental design, construction of the loudspeaker rig for the HRIR data acquisition, SRIR data acquisition as well as the conduct of all the stages of user trials (HRIR data acquisition and listening tests). He also developed the test software (including head-tracking integration) and performed the statistical analysis of the results.

In J2 the author participated in the evaluation of the HRIR factorisation algorithm, and in particular, the experimental design, conduct of the listening tests and analysis of the results.

In C1 the author participated in the experimental design as well as the conduct of the user trials (HRIR data acquisition and listening tests). He also developed the test software (including head-tracking integration) and performed the statistical analysis of the results.

In C2 the author designed the experiment, participated in the data acquisition process (both audio and video data), developed the test software, conducted the user trials and analysed the results.

In C3 the author developed and implemented the real-time hybrid convolution engine (including head-tracking integration), participated in the data acquisition process (including Spatial and Binaural Room Impulse Response data collection, multi-microphone recording of the Christ Church Choir performance as well as geometrical and visual space characterisation). He also created a fully interactive, detailed 3-D visual model of the Cathedral based on AutoCAD architectural drawings and LiDAR laser survey and integrated the model with the audio engine.

In C4 the author designed and conducted the experiment, developed the test software (including head-tracking integration) and analysed the results.

In C5 the author participated in the evaluation of the HRIR factorisation algorithm, and in particular, the experimental design, conduct of the listening tests and analysis of the results.

In C6 the author recorded a multichannel soundtrack to accompany the video in the database. The recording format was B-Format Ambisonics (Using the *Soundfield MKV* system) and 4-Channel *Zoom H2* recordings subsequently converted to B-Format.

In C7 the author developed and implemented the real-time hybrid convolution engine (including head-tracking integration), participated in the data acquisition process (including Room Impulse Response data collection, multi-microphone recording of the traditional Irish trio performance as well as geometrical and visual space characterisation). He also created a fully interactive, detailed 3-D visual model of the hall based on the geometrical measurements and integrated the model with the audio engine.

In C8 the author participated in the Ambisonics implementation of the optimised virtual loudspeaker approach.

Finally, in C9 the author participated in the experimental design, SRIR data acquisition process and the conduct of the user trials. He also developed the test software (including head-tracking integration) and performed the statistical analysis of the results.

2

Spatial Perception of Sound

Rendering of auditory scenes with high fidelity and faithfulness requires that the sound reproduction system reconstructs with sufficient accuracy important auditory localisation cues. These cues are however often subtle. That is why, in the optimal design it is crucial to deeply analyse human abilities and limitations with regard to perception of auditory information from different locations in space. It is already well understood that many different mechanisms are responsible for providing cues for localisation in a free field and in enclosed spaces, in the horizontal plane, in the median plane and in the distance. Finally, humans make use of the available information in quite a different way when the sound source is stationary and when it is in motion. All these aspects are discussed in this chapter with hope to identify areas of higher and lower importance to the recreation of spatial sound scenes in virtual auditory environments.

2.1 Sound Localisation in a Free Field

Human abilities to localise sound vary depending on the actual sound object's placement. Localisation error can be of about 3° - 4° for auditory events occurring centrally in front of the listener or even 10° and more for sounds localised laterally [26]. However, these values are representative for an average person. In practice, localisation accuracy varies across individuals and listening conditions and can be either lower (even 1° - 2° for sources localised in the front) or higher.

The term "localisation blur" is often used when talking about localisation accuracy. It describes the smallest shift of the sound source's position which creates an audible change of its perceived localisation. In experimental setups, the localisation blur can be often measured as

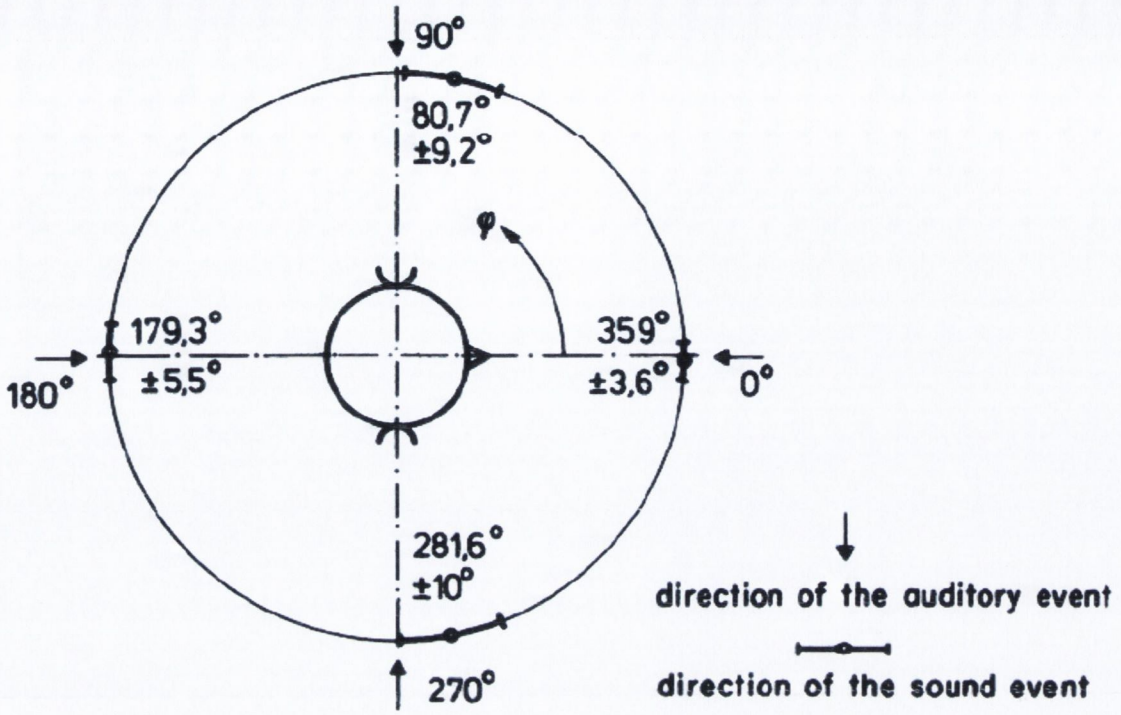


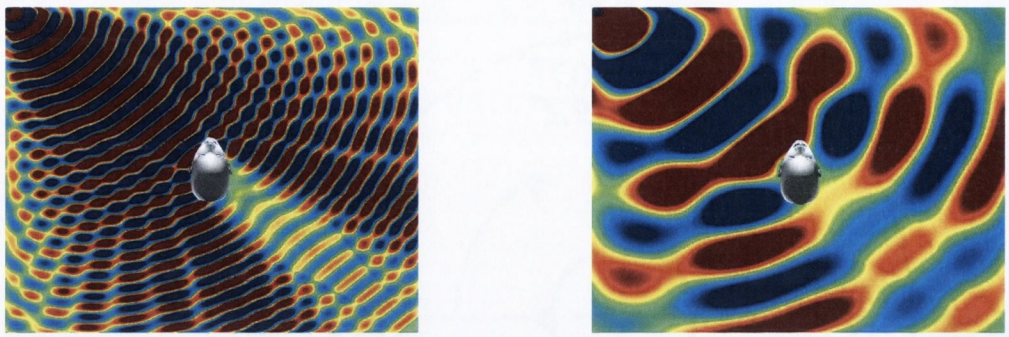
Figure 2.1: Localisation blur in the horizontal plane for white noise pulses of 100ms duration, from [26]

a Minimum Audible Angle (MAA). As we shall see in the subsequent sections, the localisation blur varies in different regions of the sphere encircling the head with clearly inferior localisation experienced at elevated angles. To understand this phenomenon it is necessary to get an insight into the theory commonly used to explain human sound localisation, known as "duplex theory", which is introduced in the next section.

2.1.1 Localisation in the Horizontal Plane

Localisation in the horizontal plane has been investigated and discussed quite intensively by many authors. For example, Figure 2.1 illustrates results presented by Blauert [26] (after Preibisch-Effenberger [179] and Haustein and Schirmer [93]) for white noise pulses of 100ms duration. In two large scale experiments (600 and 900 participants respectively) untrained subjects were asked to either align a movable loudspeaker with a static one or to position a movable loudspeaker in one of four directions: in the front, to the left, at the back or to the right.

There are two main physical phenomena that help explain the mechanisms of spatial hearing in terms of horizontal plane: Interaural Level Difference (ILD) and Interaural Time Difference (ITD). Both were first observed and described by Lord Rayleigh already in 1907 [191] as a part of his theory, known today as the "duplex theory".



(a) (b)

Figure 2.2: Illustration of the shadowing effect of the head using numerical acoustic simulation: (a) High ILD resulting from of high frequency wave; (b) Low ILD resulting from low frequency wave

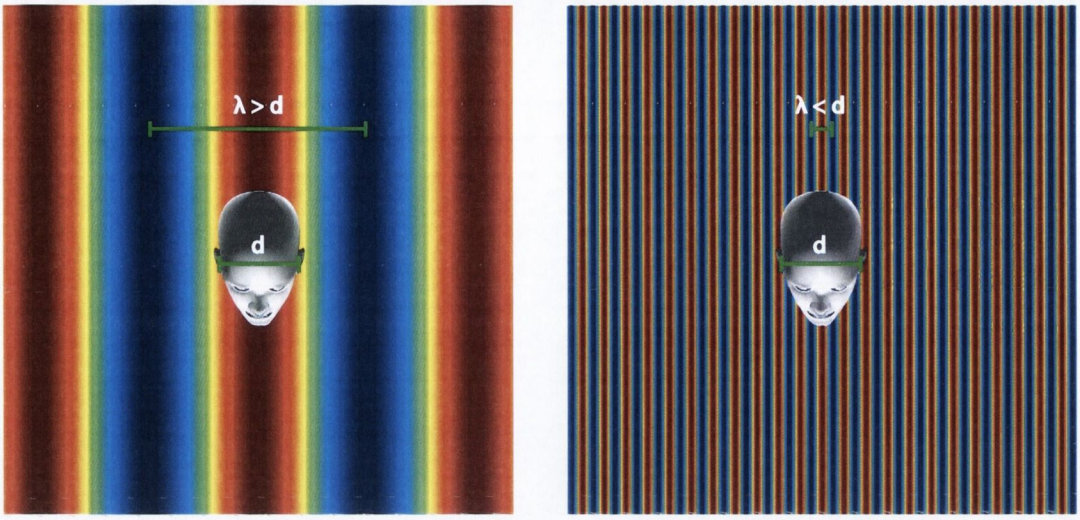
Rayleigh realised that for a sound source localised to one side of the head, the sound pressure level of the acoustic wave is lower at the ear on the opposite side due to the shadowing effect of the head. Moreover, phases of the two ear signals will also differ because of the distance the wave has to traverse from the source to the left and right ear. He referred to the intensity difference as (ILD) and the time difference as (ITD).

However, these two phenomena are not equally effective across the audible spectrum. For low frequencies (below $700Hz$) the head is no longer a significant obstacle for a sound wave and, due to diffraction, the shadowing effect is minimised. After further studies, different authors suggested that the effectiveness of the ILD cue is only above $3kHz$, noting a significant drop below $1.5Hz$ [26]. The effect of the head shadowing for low and high frequency sound is illustrated in Figure 2.2.

On the other hand, if the length of the acoustic wave is shorter than the doubled distance between ears, a phase difference at the ears may suggest a source localisation different than the actual localisation of the sound source. This phenomenon is known as "phase ambiguity", which reduces the effectiveness of the ITD cue above $700Hz$ and exhibits a significant drop of performance above $1.5kHz$ [91]. The effect of phase ambiguity is illustrated in Figure 2.3.

It is a common agreement that ITD provides a sufficient cue for the sound localisation for frequencies below $1.5kHz$ and ILD for frequencies above $1.5kHz$. It is however hardly surprising that significant drop of localisation accuracy is usually observed for pure tones with frequencies between $1 - 2kHz$ - the range in which the performance of both ILD and ITD is rather poor. So, mid-frequency pure tone signals are generally considered as difficult to localise.

When explaining the ILD and ITD mathematically, spherical or elliptical head models are often used. However, these models may lead to a paradox that is commonly referred to in the



(a) (b)

Figure 2.3: Illustration of the phase ambiguity effect: (a) wavelength λ larger than head diameter d causing unambiguous interpretation of phase; (b) wavelength λ smaller than head diameter d causing multiple, ambiguous interpretations of phase

literature as the "cone of confusion". For example, there exist multiple lateral locations where ILDs of the same value can be found. This is in fact also true for ITDs. The cone can be visualised if one draws circles around the straight line connecting both ears (the ears axis) as presented in Figure 2.4. It is then easy to observe that sound sources placed anywhere on the area of the cone will yield the same values of ILD and ITD. So, from the above it can be inferred that without additional cues, the lateral localisation of sound objects, and especially front-back localisation of elevated pure tone sources, could not be performed at all. It is therefore undeniable that there must exist other cues that help to mitigate the lateral localisation problems as well as allow for the localisation of sound sources in the median plane (where the ILDs and ITDs are also very similar). These cues are going to be presented next. Nonetheless, the weakness of the ILD and ITD cues concerning lateral placement of sound sources can at least partially explain the inferior localisation accuracy of such sources [26].

2.1.2 Localisation in the Vertical Plane

Since the vertical or median plane can be regarded as a point of symmetry for all the models of the head, sounds localised in this plane will result in ILD and ITD values close to zero. As mentioned before, this fact can already explain the deterioration of the ability to localise elevated sound sources in the median plane. Figure 2.5 shows again the results form a perceptual study presented by Blauert [26] (after Damaske and Wagner [46]) for a continuous speech stimulus, but this time across the vertical angles. The main focus of this experiment was to compare how

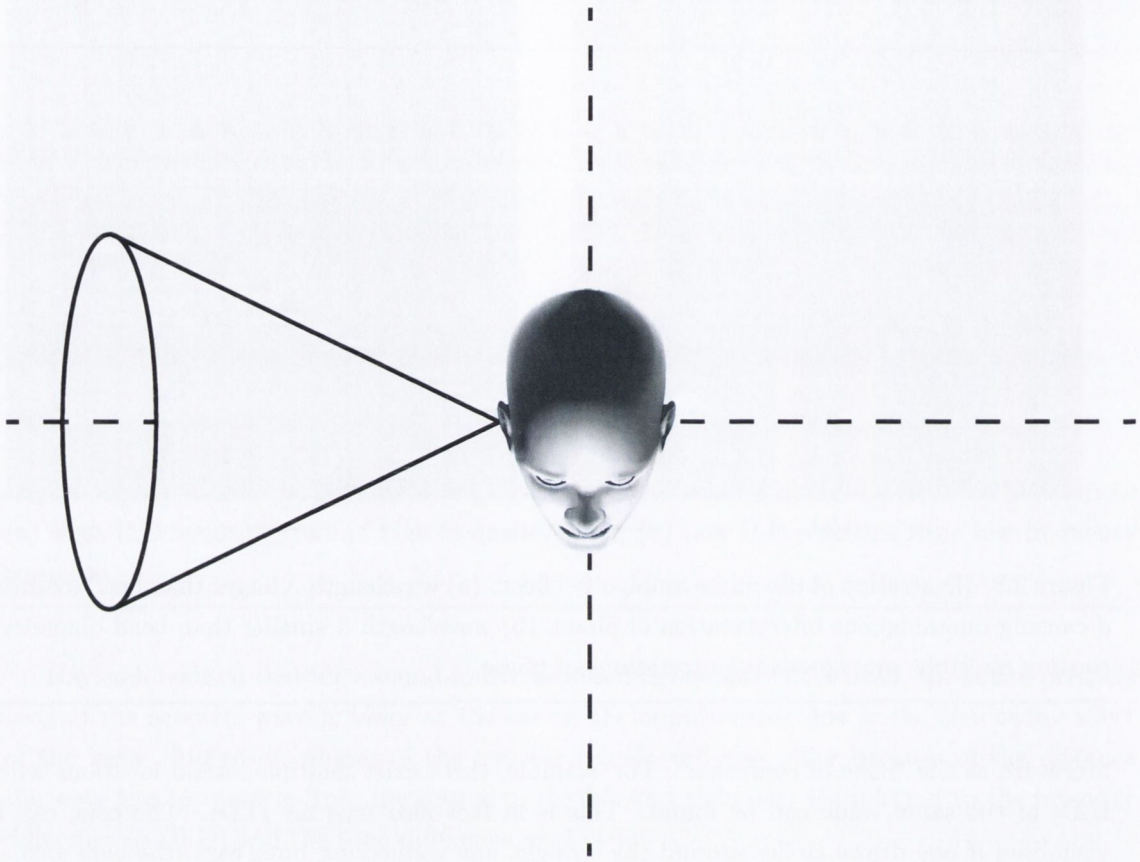


Figure 2.4: Illustration of the "cone of confusion"

subjects perceive real elevated sound sources versus binaural recordings thereof taken using a dummy head microphone. First thing to notice is a significant deterioration in terms of the angular source discrimination with much higher standard deviation values as compared to the localisation in the horizontal plane.

Interestingly, Barbour [17] showed that in the median plane, localisation of phantom image sources (which are virtual sources resulting from two or more physical sources emitting the same signal but with different amplitude and/or phase) is not significantly different from the localisation of real sources. However, in his study he considered only sources localised frontally. His findings are presented in Figure 2.6.

What can be inferred from the above discussion is that indeed there must exist other cues that permit detection of sound elevation in the absence or severe reduction of usability of ILD and ITD cues. These cues are contained in the spectral changes that some parts of human body (mainly torso, head and ear pinnae) impose on the impinging sound wave. The changes can be encapsulated in a transfer function that modifies the incoming left and right ear signals depending on the incidence angle. These transfer functions are already well established in the

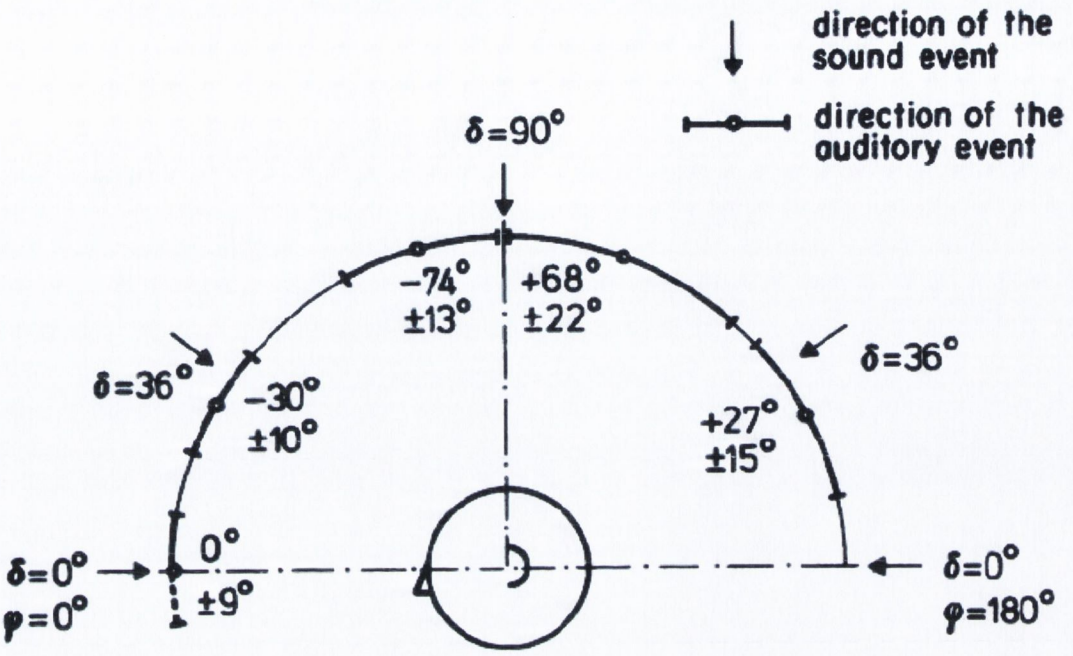


Figure 2.5: Localisation blur in the median plane for the continuous speech signal, from [26]

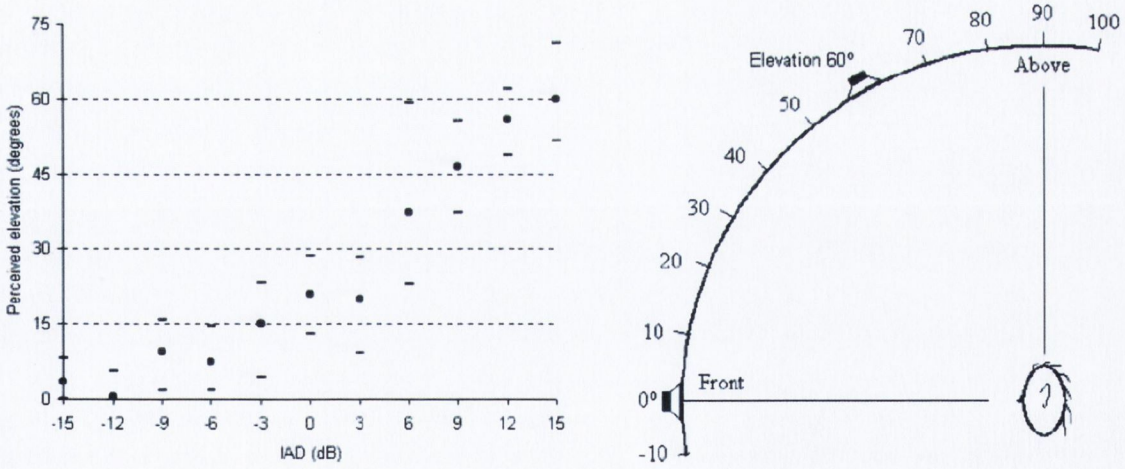


Figure 2.6: Localisation blur of the amplitude-based phantom source images in the vertical hemisphere, as obtained by Barbour [17]. The results are presented as the perceived elevation against the Interchannel Amplitude Difference (IAD) of two loudspeakers at 0° and 60° elevation respectively.

theory of sound localisation and have broad applications e.g. in virtual reality applications. They are commonly referred to as Head Related Transfer Functions (HRTFs).

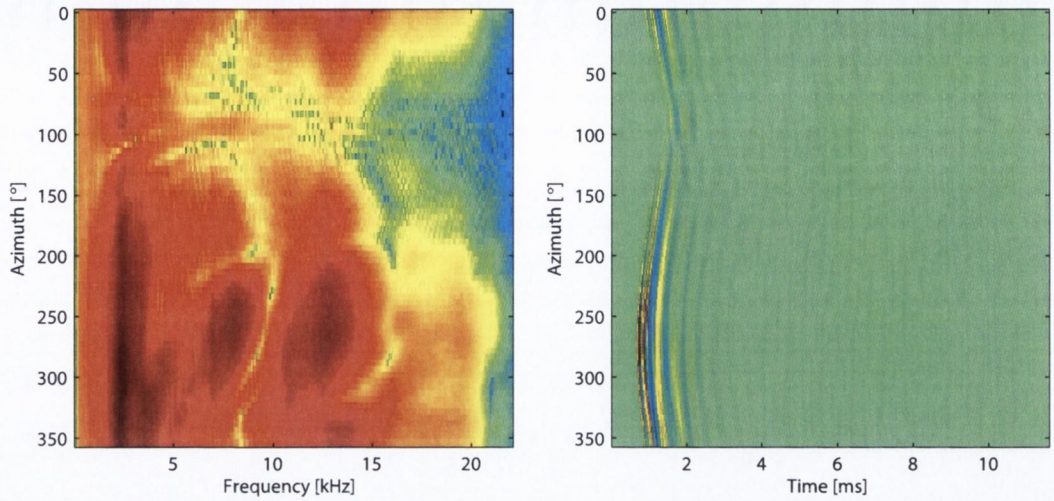
2.1.3 Head Related Transfer Function

It was discovered that the spectrum of the sound source measured in the ear canal varies depending on the sound source placement in the space surrounding the head. It means that head, outer ear, shoulders and upper torso must have impact on the sound wave's spectrum that eventually reaches the eardrum. In a sense, the human body can be described as an acoustic system that continuously performs angle-dependant filtering of the incoming left and right ear signals.

Therefore, by comparing the received signal spectrum to the spectrum of the signal that would have been received in the absence of the listener, the transfer function of this "head-related filter" can be derived. If the filter is characterised in terms of its response to the acoustic impulse (Dirac delta signal), the time domain sequence of the head-related filter is commonly referred to as Head Related Impulse Response or HRIR. Thus, its Fourier transformed counterpart is named Head Related Transfer Function or HRTF.

HRIRs are usually measured under anechoic conditions using real subjects or binaural mannequins and densely defined source locations. Commonly, tiny microphone capsules are placed at the entrance of a blocked ear canal (blocked *meatus*) on each side of the head. Then, to obtain a pair of HRIRs for a given source location, an impulsive sound should be emitted from the source and recorded by the left and right ear microphones. However, in practice better results can be obtained using techniques based on deconvolution (e.g. using exponentially swept sine tones [59]). This approach allows for simultaneous recording of the impulse response and nonlinear distortion in the audio signal path that can be subsequently removed from the final filters. However, it may significantly prolong the process of measurement, which for large databases can reach even up to several hours. It poses many difficulties, especially if the real subjects are concerned. We have to remember that even small head deviations from the initial position can result in HRIRs that are no longer correct. That is why, a significant body of research is currently devoted to the methods of simplification and paralelisation (e.g. by reciprocity) of HRIR measurement [130,186,244] or even simulation or modelling of the HRIR filters [7,14,88,129,151]. Also, full datasets of HRTFs have already been published (e.g. [6, 71, 97, 233]). An exemplary set of HRTFs for one ear and different source angles is presented in Figure 2.7(a). The time domain representation (HRIR) is also visualised in Figure 2.7(b). In the time domain it is quite easy to identify how the overall energy changes with the source angle and how it is reflected in the ILD. Similarly, samples which contain the maximum energy occur at different time samples what reflects the changes in ITD.

Convolving anechoic audio signals with a set of HRIRs results in a binaural recording. This technique can provide a very convincing and faithful illusion of a sound coming from a particular direction, provided that left and right ear signals are delivered using headphones (i.e. there is no



(a) (b)

Figure 2.7: Example of HRIR - HRTF pair. Left ear HRTFs (a) and HRIRs (b) obtained from Gardner and Martin's *KEMAR* dataset [71] for $0° - 360°$ azimuth and $0°$ elevation angles

cross-talk between the ear signals or the cross-talk is significantly minimised). This is because HRTFs contain all the necessary cues allowing for localisation of sound sources around the sphere surrounding the human head - ILD, ITD and spectral cues. It was shown that they are also sufficient to avoid the front-back confusion and determine the source elevation [91]. However, problems occur when one tries to listen to a binaural recording in which source audio had been convolved with somebody else's HRIRs. Such an experience can be compared to "listening through somebody else's ears". Despite the horizontal localisation could be in general acceptable, the perception of elevation is usually severely impaired [227]. Also, front-back discrimination is very often impossible. This is again down to the spectral cues that are no longer matching the everyday listening experience.

Although there exist no "generic" HRTFs that would provide a virtual listening experience equivalent to the real-world experience for everybody, it has been shown that it is possible to "learn" or adapt to spectral changes in HRTFs, at least to some extent. In [95] Hofman showed that a tested group of people wearing special ear molds for a period of up to 6 weeks that altered their HRTF's spectrum were able to significantly increase their localisation performance toward the end of the experiment. Also, it was shown that the experience with non-individualised HRTFs can be improved if other cues are available to the listener, e.g. due to head motion [55, 126, 229]. These cues can be incorporated into the audio rendering system with the use of head-tracking or face-tracking devices and by switching between the HRTFs to reflect the current change in source location. Although there is no unequivocal data to prove the increase of the localisation performance, it was shown that the front-back confusion rates can indeed be significantly reduced [19].

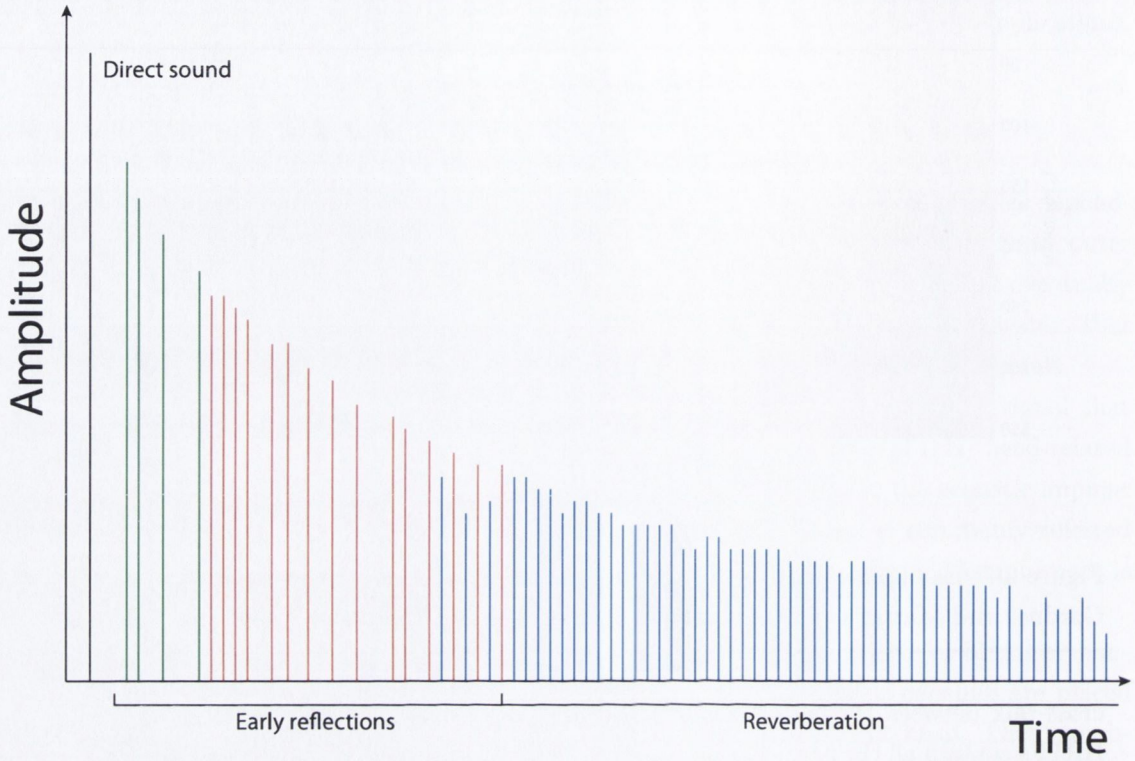


Figure 2.8: Typical echogram of a reverberant space. Black colour indicates the arrival of the direct sound. Marked in green and red are 1^{st} and 2^{nd} order reflections respectively. Dense blue lines signify reflections of order 3 and beyond.

2.2 Sound Localisation in Reverberant Environments

2.2.1 Room Impulse Response

In reverberant spaces sound waves directly emitted from a sound source undergo multiple reflections from the boundaries of that space. Then, at the listener's location, the resultant sound field is a superposition of direct and reflected waves. In the case when the emitted sound is a single impulse (Dirac delta signal), the resultant sound field at any point in the room can be illustrated as an echogram i.e. a plot that shows time of arrivals τ and magnitudes of the initial impulse and subsequent reflections. A typical echogram is illustrated in Figure 2.8. Echograms can be created rather easily by geometrical analysis of reverberant spaces. Usually, the emitted sound wave is treated as a ray of light that, upon each reflection from a boundary, obeys the Snell's law governing the arrival and reflection angles. Depending on how many times the acoustic "ray" bounces of subsequent boundaries on its path, we distinguish 1^{st}, 2^{nd}, ..., n^{th} order reflections.

Each reflection can be obtained as a mirrored source and the mirroring is done symmetrically

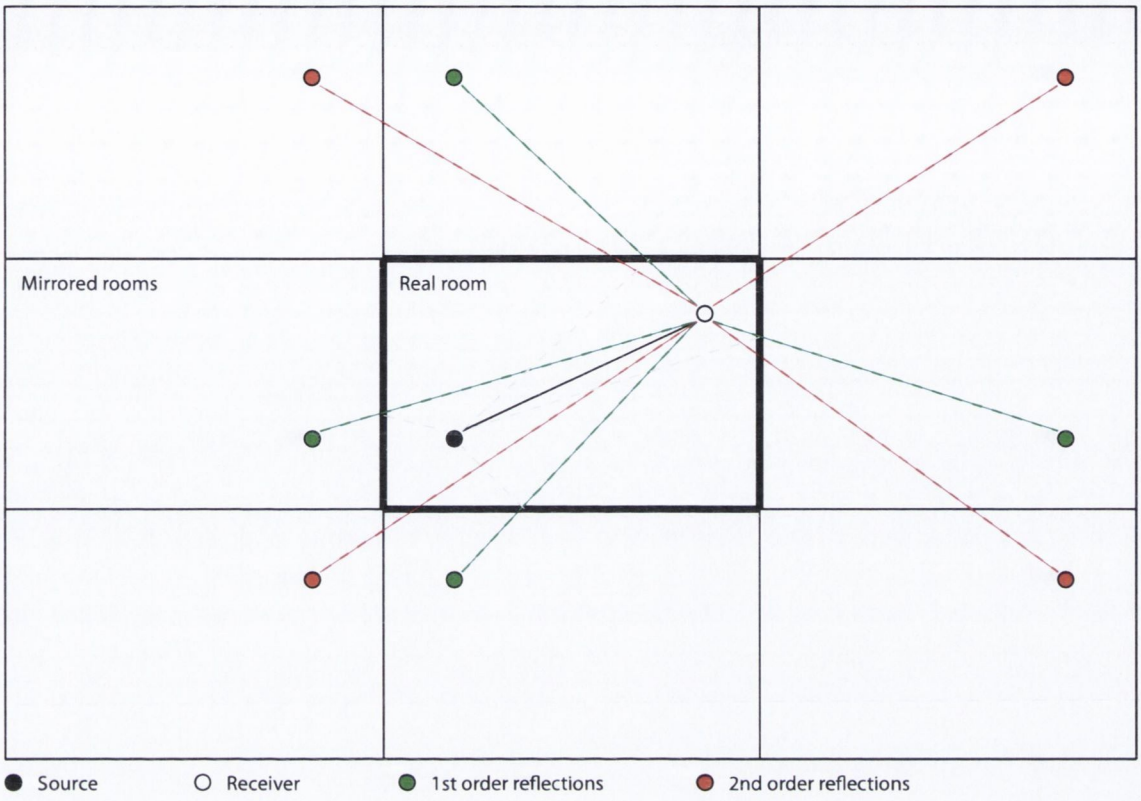


Figure 2.9: Lattice with mirror-image sources used for calculating early reflections in a simple 2-D rectangular room. Dashed lines signify natural paths of the first and second order reflections whereas solid green and red lines indicate paths between the receiver and first and second order mirror images of the source.

against each boundary of the room. This is known as Image Source Method [8] and its principles are illustrated in Figure 2.9. One can expect that obtaining co-ordinates of image sources for rectangular spaces is rather trivial but for more complex geometries it is getting more involved. An efficient algorithm for calculating image sources for simple and compound geometries has been proposed by McGovern in [138] and [139].

In practice, sound pressure level would be measured at the receiver end which would result in a time sequence known as Room Impulse Response (RIR). The difference is that RIR consists of both positive and negative pressure values which reflect frequency dependant changes of *phase* in the signal at each reflection. Nevertheless, it is a common practice to express the reflection (or absorption) coefficients as single scalar values that describe well the average reflective properties of the boundaries in a given environment. However, we have to keep in mind that in reality, reflective properties vary across different materials and boundaries so that they differ in the way they modify spectrum of the reflected waves.

Although the intensity level of subsequent reflections drops with time due to interaction

with room boundaries, their density grows with time up to the point where dense reverberation tail is created. This state is diffuse in nature and homogeneous across different locations of the space. The point in time when the dense discrete reflections become a diffuse reverberation is called the transition point of the RIR. Since the borderline is usually smeared it is also called a transition *period*. Methods of finding the transition points/periods for different room geometries were investigated e.g. by [94, 119, 215].

One of the most common as well as most useful measures to quantify reverberation properties of different rooms is the reverberation time RT_{60}. By definition, it is the time in seconds which takes the emitted broadband acoustic energy to drop to one millionth of its initial value (which is equal to a drop of $60dB$) after the emission has been ceased. Reverberation time can be measured at different points of the space and expressed, for example, in octave bands. This is called spatially averaged reverberation time. Directives for obtaining acoustic parameters of reverberant spaces, including reverberation times, can be found in $ISO - 3382$ [98].

Alternatively, reverberation time can be calculated based on the geometrical data of the space. The empirically derived equations have been proposed by Eyring [57] and Sabine [197]. The reverberation time RT_{60} according to Eyring can be calculated as:

$$RT_{60} = \frac{0.161V}{S\ln(1-\alpha)} \tag{2.1}$$

where V is the total volume of the room, S is its total surface area and α is the average absorption coefficient ($\alpha = 1$ means full absorption of the acoustic energy by the boundary, $\alpha = 0$ means full reflection from the boundary). In rooms where reflective surfaces dominate (i.e. α is small) Eyring's equation can be simplified to Sabine's equation which is:

$$RT_{60} = \frac{0.161V}{S\alpha} \tag{2.2}$$

If we capture the Room Impulse Response with the use of binaural microphones (same as in the case of HRIR measurements) we come up with a Binarual Room Impulse Response (BRIR). BRIRs, similarly as HRIRs, can be used to simulate source location in binaural recordings. The difference is that the recordings now contain the response of the particular room, so the locatedness of source is also determined by the properties of that room. This technique is a basis of the convincing audio auralisation e.g. for virtual reality applications. In some cases however, its practical implementation is not straightforward or not feasible. This topic will be discussed in more detail in Chapter 3

2.2.2 Phantom sources, Precedence Effect and Echoes

Having defined the basic behaviour of the acoustic waves in enclosed spaces we shall now discuss how it affects the perception of spatial sound. As identified earlier, sound field in rooms can be regarded as a superposition of directly emitted and reflected acoustic energy. It means, that direct and reflected wave fronts arrive at listener's ears at different times. Depending on the time lag between these signals we can define three different scenarios from the point of view of perception.

If the time delay between coherent sound waves coming from two different locations is less than approx. $2ms$, a phantom source is formed i.e. the effective localisation of the sound will not point at any of the sources but somewhere in between them. This theory is also known as "summing localisation" [26]. On the other hand, if the time delay is between approx. $2-40ms$, the law of the first wave front dominates and the wave that reaches the listener earlier favours the direction of its source. This is also known as the precedence effect. Lastly, above approx. $40ms$, depending on the relative level of the two waves, the discrete echoes will be audible.

It is therefore clear that the presence of reflected acoustic energy has the potential of altering spatial sound perception in rooms. However, the nature of the alteration is strongly dependent on the room size and geometry. Luckily, some mathematical tools have been designed in order to predict the perceptual effects of a particular space based on objectively measured parameters.

2.2.3 Interaural Cross Correlation Function (IACF)

There is another way of looking at direct and reflected acoustic energy that reach listener's ears. Room reflections also cause ear signals to decorrelate. It is possible to measure the correlation between left and right ear signals with the use of Interaural Cross Correlation function (IACF), defined as:

$$IACF\tau = \frac{\int_{t_1}^{t_2} h_L(t)h_R(t+\tau)dt}{\sqrt{\int_{t_1}^{t_2} h_L^2(t)dt \int_{t_1}^{t_2} h_R^2(t)dt}} \tag{2.3}$$

where h_L and h_R denote time domain left and right ear signals and the integration is done over the interval t_1 to t_2. The IACF, as defined above, is normalised and returns values between -1 and 1. The IACF maximum value is referred to as Interaural Cross Correlation Coefficient (IACC) and is used as the objective measure of the left and right ear signals similarity. The importance of IACC to sound perception in reverberant spaces is that it was shown to correlate well with Apparent Source Width (ASW) [94]. In general, high values of IACC indicate a narrow source and low values of IACC indicate a wide or diffuse source. Hidaka et al. suggested that the objective measure of the ASW is:

$$ASW = |1 - IACC_{E3}| \tag{2.4}$$

where subscript $E3$ denotes that the IACC is measured in the early part of the BRIR (i.e. $< 50ms$) and averaged across 3 octave bands - 500, 1000 and $2000Hz$. In order to express the ASW in terms of degrees it is necessary to know the sound intensity level as well. These relationships were investigated by e.g. Keet et al. and can be found in [110].

It is also important to note that the Interaural Cross Correlation function can be utilised to determine ITD for the incoming signal. This is the time value τ when the maximum value of the IACF occurs:

$$ITD = arg(max_{\tau}|IACF\tau|) \tag{2.5}$$

However, the ITD estimation may not be accurate in more reverberant environments. An improved method for ITD estimation was proposed by Kearney [105].

2.2.4 Former Psychoacoustical Studies on Sound Localisation in Rooms

Ability to localise sound sources in enclosures varies quite significantly depending on temporal and spectral structure of source signals. For example, steady-state sine tones are generally very difficult to localise in rooms [190]. This is mainly due to the fact that reflections create interference waves that misinform ILD cues i.e. large ILDs may occur even at low frequencies [189]. Well defined onsets usually help in such situations since they allow for the precedence effect to operate.

A distinct example of the pure tone localisation difficulties is the Franssen effect [68], also investigated by Hartmann and Rakerd [92]. The effect can be tested experimentally using a setup with two loudspeakers, one to each side in front of the head. There are a few variations of the experiment but in its simplest form the left loudspeaker outputs a sine tone with a strong onset which, after a while, starts to decay exponentially. At the same time, the right loudspeakers starts outputting the same tone with a mild onset slope so that the resultant envelope of the summed signal is rectangular. For a moment, the right loudspeaker is playing on its own to eventually pass the playback back to the left loudspeaker. In the course of the experiment, it has been shown that listeners completely do not realise the fact that there was any contribution from the right loudspeaker.

Localisation of impulsive sounds was shown not to be dependent on the reverberation time [90] when the reverberation time was reduced from $5s$ to $1s$. However, room geometry and the early reflections may still play a significant role. For example, in another study Hartmann and

Rakerd [90] discovered that lowering of the ceiling and effectively re-ordering early reflections made the localisation easier. However, the effect of magnitude of early reflections on source localisation has been tested only in a limited range ($\pm 7dB$) and did not show any significant results.

Therefore, it is quite clear that transients facilitate the localisation in rooms by making it easier for the preceding wave to be detected. However, even if transients are present in the test signal, the precedence effect still does not eliminate all influences of room reflections [189].

Localisation of sounds without transient attacks is more difficult but gets better if the frequency spectrum of a signal gets wider [90]. However, steady-state broadband noise sounds are still localised with reduced accuracy than impulses. The explanation of this effect can be that reflected noise acts as a masker to the test signal. Thus, localisation of steady-state broadband signals also depends on the reverberation time.

2.3 Localisation in the Distance

Throughout the literature, there exists a clear distinction between "distance" and "depth", both understood as perceptual attributes of sound. According to Rumsey [195], "distance" is related to the physical range between the sound source and the listener, whereas "depth" relates to the recreated auditory scene as a whole and concerns the sense of perspective in that scene. In the light of the previous discussion, we shall focus on the former since clearly it expresses the last, 3^{rd} spatial dimension of the possible sound location which is of great concern in this work. At the same time, significantly less attention has been devoted to perception of distance as compared to localisation in horizontal and vertical planes.

2.3.1 Perception of Auditory Distance in a Free Field

Although the human ability to perceive sources at different distances is not fully understood yet, there are however several key factors that are known to contribute to the perception distance. First of them, would certainly be the sound amplitude which is related to the perceived loudness. In general, sound intensity level drops if we move away from a sound source. Since the wave propagation obeys certain physical rules, the extent of amplitude reduction can be quite accurately predicted.

Let us assume a point-like sound source emitting a perfectly spherical sound wave. Sound intensity I of a monopole source at a given distance r can be expressed mathematically as in Equation 2.6. Such a wave traversing the distance from the source to a listener undergoes the process of acoustic energy spreading over a larger and larger area of a sphere. As a result, the sound intensity will diminish according to the inverse-square law and the sound intensity level will drop by $6dB$ with each doubling of the distance (Equation 2.7).

$$I = \frac{p^2}{\rho_0 c r^2} \tag{2.6}$$

$$10\log_{10}\frac{I_{2r}}{I_r} = 10\log_{10}\frac{p^2}{\rho_0 c(2r)^2}\frac{\rho_0 c(1r)^2}{p^2} = 10\log_{10}\frac{1}{4} \approx -6.02dB \tag{2.7}$$

Where I is the sound intensity, ρ_0 - medium density, c - speed of sound in the medium. The expression $\rho_0 c$ is also called the acoustic impedance of a medium, often expressed by the capital letter Z.

In general, sound intensity level, which is non-linearly related to the perceptual loudness, is often referred to as a *monaural* distance cue. Unless the sound source is very familiar (like the voice of a known person), this cue is also said to be *relative*. It means that for humans in most cases it is very difficult to judge the distance to a source simply by presenting an isolated sample of a certain loudness, particularly if the sound source is unfamiliar. On the other hand, presentation of two samples that differ in perceived loudness might lead to the relative comparison and, as a result, the inference so to their relative distance (i.e. which one is closer and which one is further away).

With increasing distance from the sound source, wave fronts change their shape and for sufficiently remote sources a spherical wave becomes similar to a plane wave. It has been hypothesized that this phenomenon can constitute another important distance cue and thus correct reconstruction of the wave fronts could be beneficial for the perception of distance. However, it has been shown by Wittek in [237] that there is no correlation between the wave front curvature and the perceived distance and this subtle cue becomes easily overridden by stronger ones like aforementioned loudness.

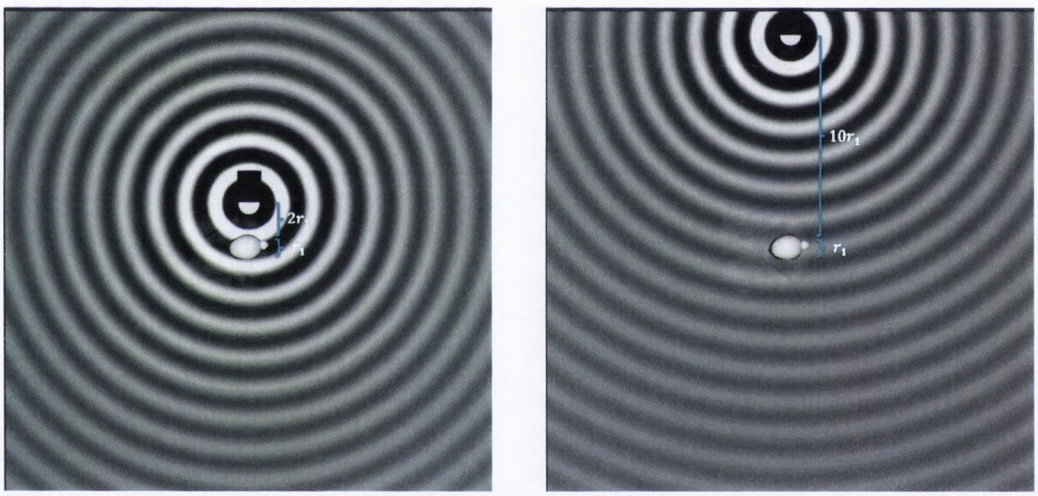
Sound waves travelling over a substantial distance also undergo the process of energy absorption by water molecules in the atmosphere. It is more apparent in high-frequency energy of the wave and leads to spectral changes (i.e. low-pass filtering) of the sound being heard. In general, sound intensity at a certain distance r corrected for the absorption effect of the medium can be expressed by an empirically derived equation of a form:

$$I(r) = I_{1m} e^{\frac{-r}{\alpha}} \tag{2.8}$$

Where I_{1m} is sound intensity for a spherical sound source calculated according to the Equation 2.6 and α is a constant parameter dependant on the frequency of the acoustic wave as well as the temperature, humidity and atmospheric pressure of a medium.

For even larger distances and sound waves characterised with high pressure levels, the propagation speed in a medium ceases to be linear which may lead to additional waveform distortions [44]. It is however unclear if such distortions commit to the perception of distance.

On the other hand, sound sources very close to a listener and localised off-centre can produce substantial level differences at the ears [132]. For sources approaching the head, as in Figure 2.10, the inter-aural distance can constitute a significant fraction of the overall distance to the source which in effect, according to the inverse-square law, can result in the substantial increase of ILD. Calculated ILDs, including the shadowing effect of the head [87] and resulting from sinusoidal waves of different frequencies emitted by a close source, are shown in Figure 2.11. For these reasons, subconscious head movements toward and away from the sound source may be regarded as another important cue [13] since intensity level changes close to the source will be more apparent than far away from it. It can be shown that the rate of change of acoustic pressure with distance ($\frac{dp}{dr}$) when related to sound pressure level at a given distance r, is independent of the *emitted* energy but dependent on the distance. Mathematically, it can be expressed as in the Equation 2.9:



(a) (b)

Figure 2.10: Sound source emitting a spherical wave at the distance: (a) r equal two times the inter-aural distance; (b) r equal ten times the inter-aural distance. As the sound source moves away from the listener, lower ILD is observed.

$$\frac{(\frac{dp}{dr})}{p} = -\frac{kr^{-2}}{kr^{-1}} = -r^{-1} \tag{2.9}$$

Where p is acoustic pressure, r is the distance and k is a constant.

Other modes of head motion can lead to further enhancement of binaural differences like

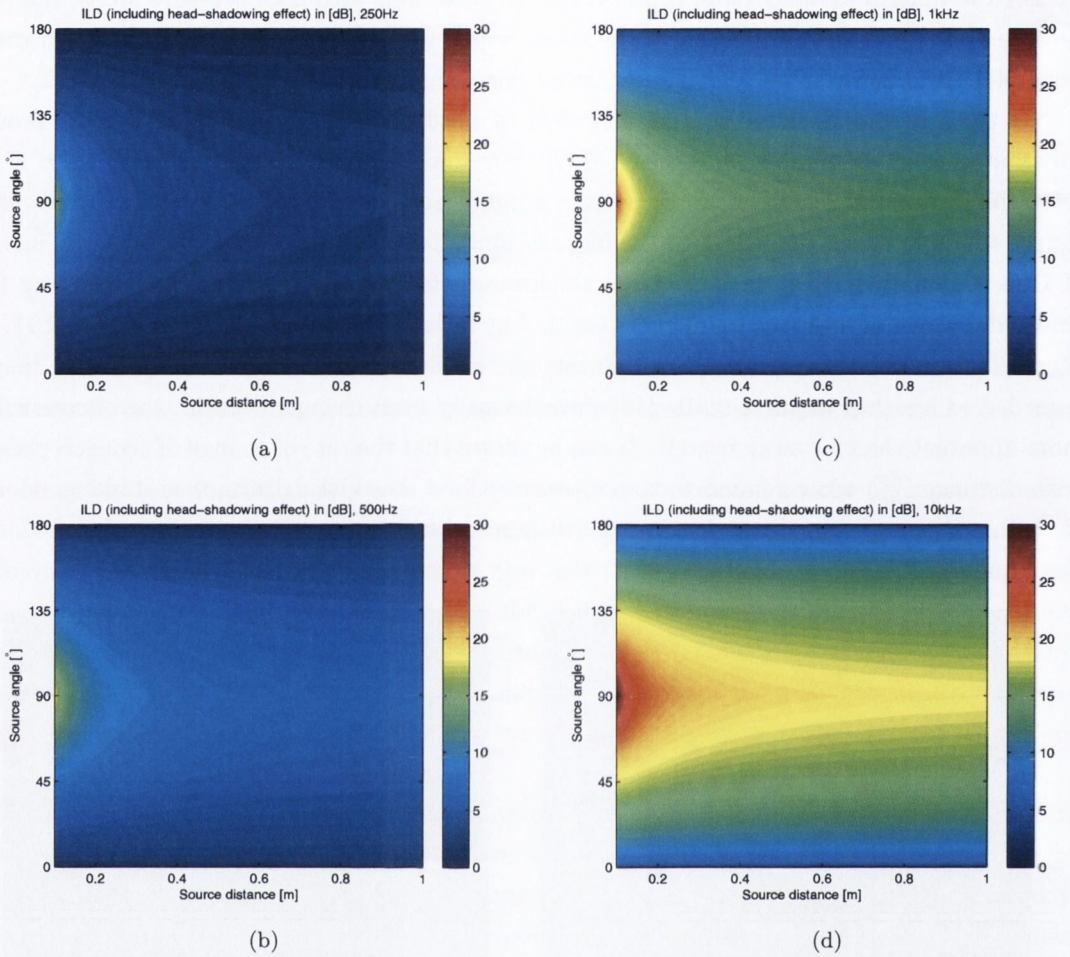


Figure 2.11: ILDs, including the shadowing effect of the head [87], resulting from sinusoidal waves of: (a) 250Hz; (b) 500Hz; (c) 1kHz; (d) 10kHz. Sound source distance varies 0.1m to 1m.

differences in phase, intensity and spectrum as well as cause an effect of acoustic parallax (which can be compared to the similar phenomenon encountered in vision) where directions of incidence are different for left and right ears [237].

2.3.2 Perception of Auditory Distance in Reverberant Environments

Reverberant spaces can offer additional facilitation when it comes to auditory distance judgements. This is in contrast with what has been said in terms of perception of direction in enclosures. On a macro level, the amount of directly emitted and reflected energy is very informative to human distance perception mechanisms. However, the ratio between these two types of energy varies across different points in rooms. The distance at which reflected energy is equal to the directly emitted energy is known as the critical distance of a room. Sabine's

approximation of critical distance in reverberant spaces is expressed as a square root of the ratio of the room volume V and the spatially averaged reverberation time RT_{60} [15]:

$$d_c \approx 0.057\sqrt{\frac{V}{RT_{60}}} \tag{2.10}$$

In general, within the critical distance of a room the direct sound energy decays roughly by $6dB$ per each doubling of the distance which follows from the inverse-square law. However, the diffuse energy fills the enclosure homogeneously and its level is approximately the same across the whole space. Therefore, beyond the critical distance, the diffuse properties of the room outbalance the direct field properties. This is illustrated in Figure 2.12.

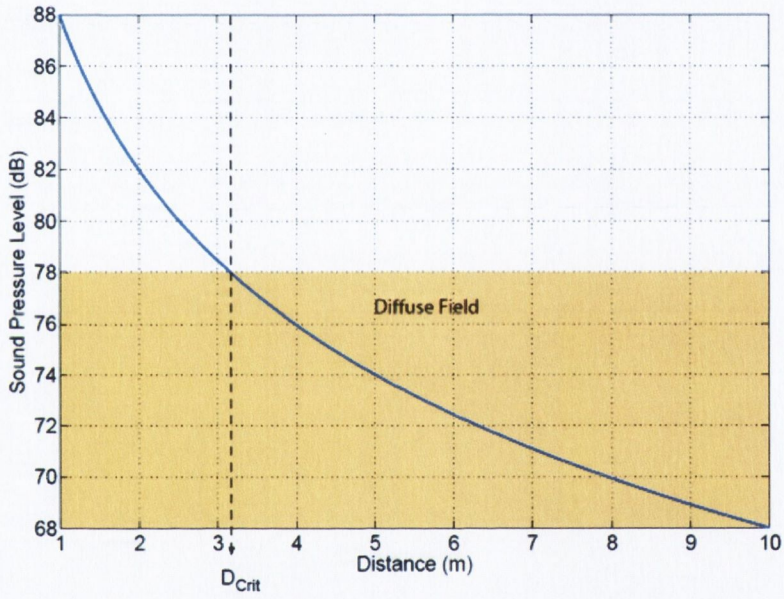


Figure 2.12: Typical acoustic energy distribution with distance in reverberant spaces, from [105]

On a micro level, the pattern of early reflections can be regarded as another important cue for the perception of auditory distance. This pattern is unique to each source-receiver configuration in a given environment (however, according to the Helmholtz reciprocity theorem [224], the source can have its location swapped with the receiver without affecting the pattern). Nevertheless, it is possible to observe some regularities in the way the pattern changes with increasing the distance between the source and the receiver. Figure 2.13 shows a collection of Room Impulse Responses from an enclosed space (middle size lecture hall) which main dimensions were $15.6m \times 5.85m \times 4.14m$. The responses were measured for source-receiver ranges of $1m$ (a), $2m$ (b), $4m$ (c) and $8m$ (d) respectively using exponentially swept sine tone technique [59]. The measured responses occupy the upper space of each plot. Below are their synthesised equivalents

that better depict occurrences of each reflection from the main bounding box of the room. The synthesis was performed using the Image Source Method [8] as described earlier in Section 2.2.1. As we can see, for close sources the gap between the direct sound and first reflections (often from the floor) is typically quite significant. The ratio of the direct sound energy and that of the subsequent reflections is high. However, as the source moves further away, both the gap between the direct sound and first reflections as well as their amplitude ratio diminish.

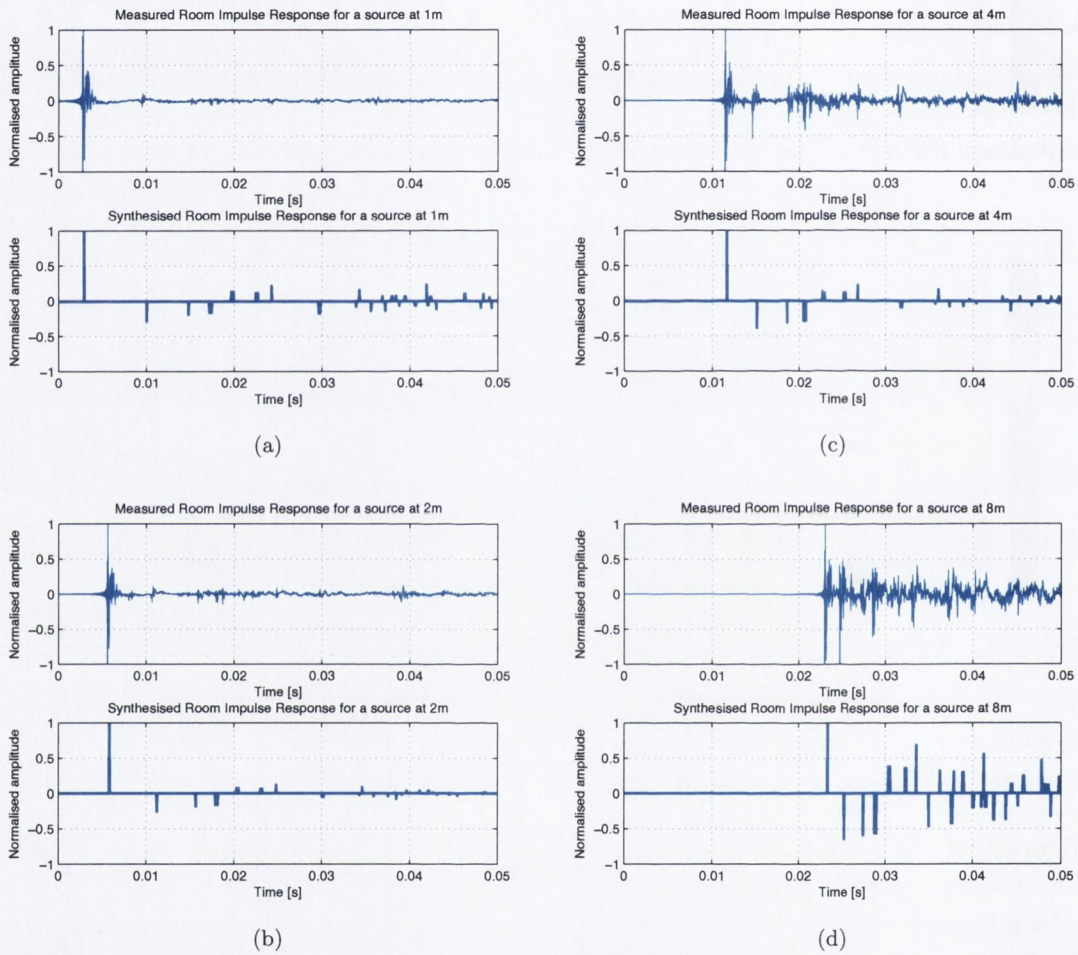


Figure 2.13: Measured and synthesised [8] Room Impulse Responses for a medium size lecture hall ($15.6m \times 5.85m \times 4.14m$) and source-receiver distances of: (a) 1m; (b) 2m; (c) 4m; (d) 8m

These patterns can be very well utilised in order to improve localisation accuracy in enclosed spaces. However, questions still remain as to the importance of the time-wise accuracy and the locatedness of the first reflections and whether or not can it positively contribute to the perception of sound source distance or perception of the scene depth as a whole.

2.3.2.1 Former Psychophysical Studies on Auditory Distance Perception

The problem of perception of auditory distance has undoubtedly received significantly less attention than other aspects of spatial hearing (e.g. discrimination of auditory direction). Nevertheless, there is still a significant amount of research done in the field of distance perception in the context of both anechoic and reverberant acoustic conditions. The study objectives as well as methodologies used vary greatly but in general they can be divided into those utilising real, virtual acoustics or both.

To start with, it has been shown in numerous studies that perception of auditory distance is not linearly dependant on the actual source distance. In most cases, distance is underestimated in both damped and reverberant spaces. For example, experiments by Nielsen [164] and Gardner [70] show this phenomenon in the context of speech stimuli under anechoic conditions.

In another study, Bronkhorst et al. [32] compare perception of distance in anechoic and echoic settings using virtual acoustics. The stimuli are presented binaurally and they differ in the number and amplitude of synthesised room reflections. The results show that in the damped environment, sources are consistently perceived to be closer than in the reverberant environment due to a high dry-to-reverberant energy ratio, as shown in Figure 2.14. The authors also demonstrate that it is possible to affect the perceived distance by increasing the number and amplitude of reflections. Similar results can be found in the work of Mershon et al. [146] where short reverberation times led to underestimation of source distance and long reverberation times produced the overestimating answers. Additionally, in this work the authors were also interested in the influence of other background sounds and their impact on subjective judgements of distance. Interestingly, they found that high level broadband background noise creates the impression of a closer sound source.

Zahorik [240] claims that the underestimation of auditory distance can be reasonably well approximated by a power function. Similarly, Bronkhorst et al. [32] suggest that the decreasing slope of the distance perception curve shows the effect of "acoustic horizon".

Direct comparison between distance perception in real and virtual environments has been done in the studies by Rychtarikova et al. [196] and Chan et al. [37]. Both in their experiments used binaural room simulation for the creation of their Virtual Acoustic Environments. In the first study, a good agreement was found at distances of $1m$, however at $2.4m$, the accuracy degraded significantly. Similarly, Chan et al. demonstrated the underestimation of the source distance in virtual reverberant environments, more so than in the case of the real environment.

Both sound intensity level and direct-to-reverberant energy ratio have been already identified as important distance cues. In another publication, Zahorik refers to them as primary distance cues [239]. In fact, changes of only these two cues are capable of creating different impressions of distance for the same subject. However, in [145] Mershon argues that only the latter can constitute an absolute distance cue whereas the former is only relative. In other words, greater reverberation is usually associated with greater perceived distances, but variations in amplitude

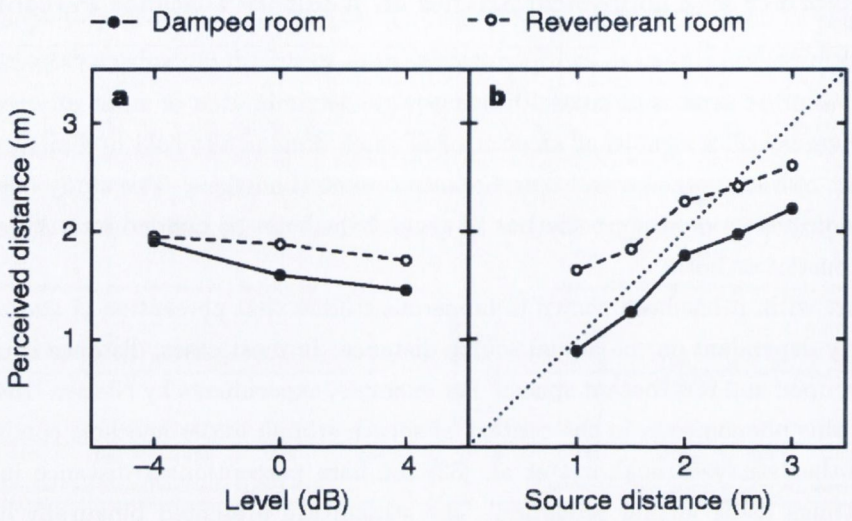


Figure 2.14: Perception of distance in damped and reverberant environments, as obtained by Bronkhorst et al. [32]

as large as 20dB not necessarily produce coherent answers. It was also experimentally verified by Bronkhorst et al. [32] who discovered that the amplitude effect on perceived distance is as small as $< 0.5m$ distance decrease with $8dB$ amplitude increase.

In the studies by Waller [226] and Ashmead at al. [13] it was shown that one of the factors improving distance perception is the listener movement in the virtual or real space. It is therefore important to account for any listener's movements (or lack thereof) in the experimental design. Similarly, for binaural reproduction of virtual acoustic environments, small, subconscious head movements may lead to improvement of distance perception by providing enhanced ILD and ITD cues. Thus, it is important that the sound field transformations reflect well that small changes of position and orientation of the listener's head.

Brungart & Scott [35] argue that familiarity also plays an extremely important role in the distance judgements. They show that fully voiced speech samples can be perceived at significantly different distances than their whispered versions, despite their objective amplitudes being the same. It may suggest that some sort of *semiosis* or experience-based assessment occurs at higher, cortical levels of auditory information processing. Results of their study are shown in Figure 2.15.

Lastly, in one more study Zahorik [238] has shown that presence of other modalities in the experimental protocol also affect the perception of auditory distance. In his experiment the standard deviations obtained in the presence of visual stimuli were significantly smaller than when the audio was presented in isolation. His results are illustrated in Figure 2.16.

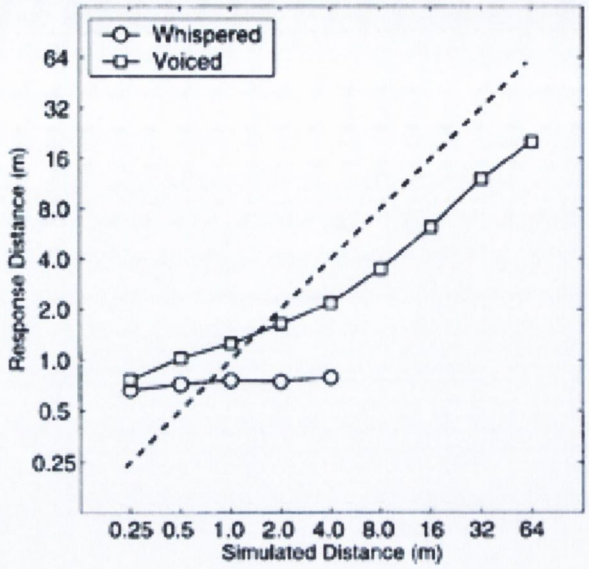


Figure 2.15: Differences in perception of voiced and whispered speech, as obtained by Brungart & Scott [35]

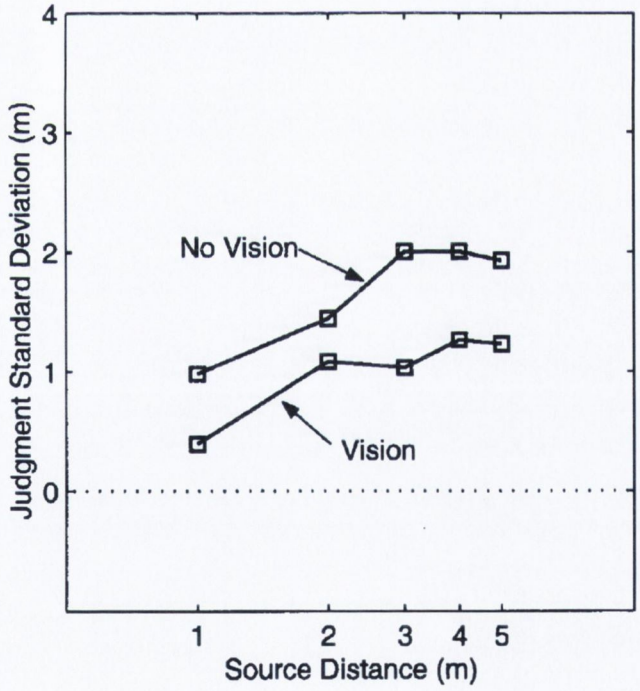


Figure 2.16: Differences in standard deviations for auditory distance judgements with and without visual stimuli, as obtained by Zahorik [238]

2.4 Localisation of a Moving Sound

One more aspect of spatial sound perception that needs to be taken into account, especially in the context of real-time rendering of auditory scene is perception of a moving sound. The types of motion can in general be divided into advancing/retreating motion, angular motion at a constant distance or combinations of the above. In dynamic auditory scenes (including real-time applications) the first type of motion can be quite easily simulated with changing the parameters responsible for auditory distance perception as described in previous sections, mainly: sound amplitude, dry-to-reverberant energy ratio and pattern of early reflections (for enclosed spaces) and high-frequency energy absorption (for long distances). Additionally, a physically correct results can be obtained if one simulates the Doppler effect, e.g. wavelength elongation with a retreating sound source or wavelength shortening with an advancing sound source.

Sources moving on an arc and at a constant distance are generally easier to render but it is still interesting to investigate what perceptual effect the motion imposes on the angular discrimination of the source location. Although it is not fully understood yet, it has been hypothesised that motion information is processed separately from other aspects of auditory information, such as stationary location, and is based on the assessment of the angular velocity of the source. In its favour are physiological studies on the hearing system of cats, which have proven the existence of neurons reacting in a different way when a sound source is in motion [211]. Also interesting are clinical studies where neurologically impaired subjects were completely unable to perceive auditory motion but performed well in judging the stationary location [122].

Another hypothesis attempting to explain a perception of sound motion was based on the analysis of instantaneous sound source position at times t and $t + \Delta t$ from which, what was believed, the motion can be inferred [149]. However, it was not confirmed in the research by Perrot et al. [174] who argued that a human listener is able to distinguish between two sources, accelerating and decelerating, moving on the same arc and in the same time span.

In the research on the perception of auditory motion the notion of finding a Minimum Audible Motion Angle (MAMA), as opposed to the Minimum Audible Angle encountered before, takes a central place [169]. MAMA is a minimum angle that the moving sound must traverse before it can be distinguished from a static sound or a sound travelling in the opposite direction. We know that for low angular velocity values MAMA is only slightly higher than (static) Minimum Audible Angle (MAA) [38, 89]. On the other hand, for rapidly moving sounds, MAMA extends to infinity and, for very slow movements, converges asymptotically to MAA (gets closer, but never reaches its value). It is then expected that localisation blur of moving sources may grow with angular velocity as well.

Some attempts have also been made to directly evaluate the perception of dynamic sound sources in virtual reality environments. For example, according to [85] a moving virtual sound source is localised with the same accuracy as a static phantom image source. Also, as shown by

the author in [82] the initial differences in localisation blur of a static source (e.g. due to variable accuracy of a rendering system) seem to vanish as soon as the sound source is put in motion. However, this study was performed for headphone listening only and further investigations are required to verify this statement. Therefore, it is still an open question what limitations on the perception of dynamic sound sources the human hearing system exert, especially in the context of virtual reality and immersive environments.

2.5 Conclusions

In this chapter basic background information concerning the spatial sound perception was introduced. First, localisation in the horizontal and vertical planes was discussed and human discrimination abilities in this regard shown based on several well regarded studies. Two localisation mechanisms - Interaural Level Difference (ILD) and Interaural Time Difference (ITD) - were presented and limitations of their usability discussed. Next, spectral shaping of the ear signals was elaborated on, as another important cue to localisation. This led to a notion of a Head-Related Transfer Function that model acoustic filtering of the impinging sound waves due to their interaction with physical features of human body. It was identified that due to angle-dependant filtering process humans are able to localise sound sources in the elevation.

The spatial sound localisation was subsequently extended to include the effect of enclosures. It was identified that a reverberant sound field is a superposition of directly emitted and reflected waves that alter left and right ear signals. These alterations in general make the process of angular source discrimination more difficult but it is heavily dependant on the spectral and temporal content of audio. However, localisation in the diffuse sound field is still possible due to the law of the first wave front reaching the listener's ears.

Next, in-depth analysis of distance perception mechanisms was performed. Based on the physical aspects of sound propagation as well as former studies in the subject it was identified that perception of distance may rely mainly on the monaural cues and to less extent on the binaural cues. The monaural cues which are commonly referred to when talking about distance perception, are sound amplitude, direct-to-reverberant energy ratio, pattern of early reflections and high-frequency attenuation of the waves from distant sources. Here, a significant role has been given to the reverberant environment that, contrary to the directional localisation, produces additional cues as to the sound source distance. There are however still numerous questions remaining, pertaining mainly to the importance of early reflections in the process of auditory distance assessment. It was identified that the general pattern and relative amplitudes of early reflections change in a predictable way when altering the distance between the sound and the receiver. However, it is still unclear whether the directional accuracy of the direct sound and early reflections can have any impact on the accuracy of subjective distance judgements.

For this reason, it is proposed to perform a series of subjective experiments that will investigate the above problem in isolation form other aspects of auditory distance rendering. This

will be the topic of Chapter 5 of this thesis. Before then, however, it is important to review different sound field rendering methodologies with a particular focus on the techniques that allow for rendering of spatial audio sources with arbitrary level of accuracy. Since in general better spatial resolution of sources comes with a higher computational demands, the problem seems to be particularly important from the point of view of real-time audio rendering e.g. for video games or virtual reality applications.

3

Spatial Reproduction of Sound Fields

One of the requirements imposed on any spatial audio system, as far as the faithful and immersive listening experience is concerned, is the correct reconstruction of auditory localisation cues discussed in the last chapter. Human abilities and limitations in this regard have been already presented and now we shall review different methodologies that have been proposed in literature in a hope to identify techniques that are particularly suited for real-time auditory scene synthesis.

This chapter starts with the very basics of virtual (or phantom) image creation in the case of 2-channel stereophonic sound field. Then, methods of recording stereophonic sound images are reviewed with a particular focus on the conservation of spatial audio cues as discussed in Chapter 2.

2-channel stereophony constitutes a very good introduction to multichannel stereophonic systems which are commonly referred to as "surround sound" systems. In the next section, strengths and weaknesses of the most popular surround sound layout is discussed. Then, the method allowing for expansion of multichannel stereophony principles to arbitrary 2- and 3-D arrays of loudspeakers with the use of Vector Base Amplitude Panning (VBAP) is introduced.

On the opposite side, there are methodologies aiming at "holophonic" reconstruction of a sound field within some defined listening area. This is in contrast with the notion of producing phantom sources on an imaginary area defined by loudspeakers in multichannel stereophony. Holophony, which in fact can be compared to holography in the visual domain, means that changes of acoustic pressure and velocity are correctly recreated at all points of the reconstruction volume. However, in multichannel audio technology there are at least two different well-known approaches to acoustic holophony that are based on different principles.

The first technique, which is called Wave Field Synthesis (WFS), is concerned with recreating acoustic wave fronts with the use of secondary sources. In order for these wave fronts to to be reformed correctly for a broad frequency range, large and densly positioned arrays of loudspeakers are used. This technique will be presented in Section 3.2.

The second technique, called Higher Order Ambisonics (HOA) looks at a sound field as a superposition of plane waves. Thus, the objectives of HOA can be stressed as the reproduction of plane waves with the use of loudspeaker configurations. The plane wave decomposition of a sound field may be truncated at arbitrary order which results in sound fields recreated with varying accuracy. Thus, Higher Order Ambisonics method is "asymptotically holophonic" which means that the higher the order of truncation is, the closer to the *real* holophonic reproduction we get. The HOA approach will be the subject of Section 3.4

The discussion about HAO will be preceded with the introduction to First Order Ambisonics (FOA or simply Ambisonics) as a basic method for decomposition of a sound field into one omnidirectional and three directional components (Section 3.3) that provide a general description of the sound field in all three spatial dimensions. FOA is important from the practical point since it is the easiest way to physically record real sound fields. This is because of the wide availability of recording equipment (microphones and processors) that allow for sound field recording directly into the Ambisonic format. As we shall see, recording of HOA is much more problematic and requires a very large number of microphone capsules arranged in a strictly defined way. To mitigate the problem with HOA recording, a method for synthesising higher order components from first order recordings will be discussed in Section 3.4.5.

Next, binaural audio reproduction techniques will be presented in Section 3.5. In general, in binaural techniques the left and right ear signals are already pre-filtered with corresponding Head Related Transfer Functions (Chapeter 2, Section 2.1.3) so it is a requirement that they are presented to the listener without cross-talk. Although techniques exist that allow for loudspeaker reproduction of binaural signals, the best way to assure left and right ear signal separation is headphone reproduction. That is why headphone listening will be of primary focus in this section.

Binaural playback can be also realised with the use of "virtual loudspeaker feeds" which will be the subject of discussion in Section 3.5.1. The method has been proven appropriate in real-time audio where relative source-listener locations change dynamically. In the normal situation, frequent change of HRTFs would be required that poses several difficulties from the practical point of view. In particular, very large datasets of HRTFs would be required in order localise sources at continuously changing locations. The virtual loudspeaker approach deals with the problem by using only limited number of HRTFs to describe all possible sound locations and forming phantom images at all the intermediate points. However, on the downside, multiple filtering processes are required simultaneously. Luckily, a method for optimisation of virtual loudspeaker reproduction will be also discussed in Section 3.5.2.

Finally, some concluding remarks will be presented in Section 3.6.

3.1 Stereophony

3.1.1 2-channel Stereophony

We call an audio system "stereophonic" if it incorporates more than one channel of audio information and where the acoustic scene is created at and in between loudspeakers. A directional imaging in stereophony can be achieved by the means of level and time difference panning. In other words, to create a virtual sound source between two loudspeakers, it is necessary to ensure that the combination of signals impinging at the listener's left and right ears give rise to ILD, ITD or both as explained previously in Chapter 2. Depending on which phenomenon is exploited, we are talking about intensity-based stereophony or time-based stereophony. Thus, the phantom images in stereophony are typically created by the means of amplitude panning, time-shifting or both. However, regardless of the method used, the correct image can be only created for one particular point in space, called the "sweet spot". The sweet spot creates with the left and right loudspeakers an isosceles triangle, as presented in Figure 3.1. Coming closer to one of the loudspeakers perturbs level differences and affects fine phase relations between two signals. Also, signal coming from the closest loudspeaker is perceived earlier causing the stereo image to collapse toward this loudspeaker. This is known as the precedence effect (see Section 2.2.2).

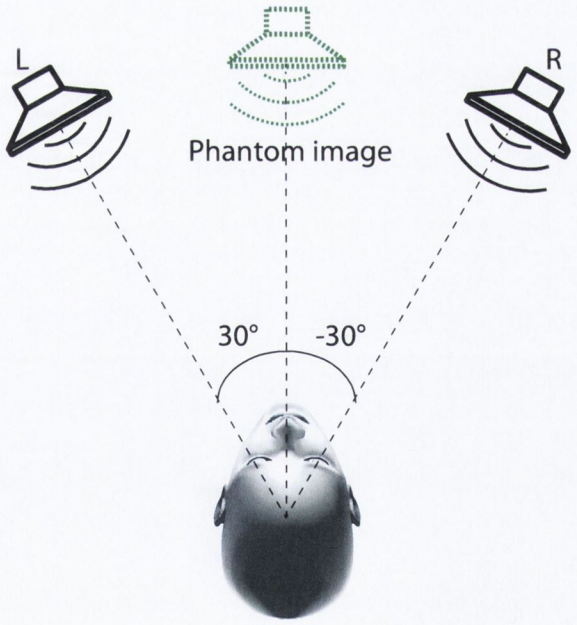


Figure 3.1: 2-channel setup for stereophonic sound reproduction

There are at least a few different ways of obtaining a stereophonic playback. The first is to use the aforementioned amplitude and/or time-shift panning for phantom image creation. In this method, a source is usually recorded within the critical distance of a room (or in the

Phantom source position	ΔA		ΔT
30°	$15dB$	or	$1.12ms$
20°	$5.5dB$	or	$0.44ms$
10°	$2.5dB$	or	$0.2ms$
0°	$0dB$	or	$0ms$

Table 3.1: Amplitude or time differences required in order to shift a phantom source by a specific angle. Values as obtained experimentally by Simonsen [204]

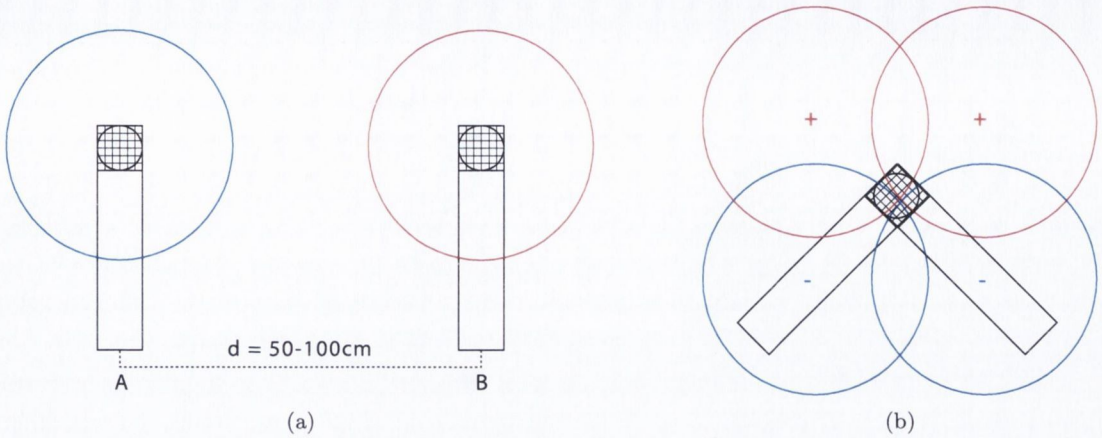
anechoic chamber). Then, the monophonic audio signal is fed into the left and right loudspeaker for reproduction. If the gain of both channels is equal as well as there are no time differences at the sweet spot, the phantom image is created in between two loudspeakers. On the other hand, if amplitude and/or phase differences are applied to left and right channels, the phantom source will shift toward left or right loudspeaker. Amplitude differences ΔA and time differences ΔT required to shift the phantom image by particular angle (assuming the standard 2-channel stereophonic listening setup as in Figure 3.1) are collected in Table 3.1 [204].

Signals for the left and right loudspeakers can be also recorded using arrays of microphones. First significant advancements in this regard were incepted independently by Harvey Fletcher [65] in the United States and by Alan Blumlein [29] in the United Kingdom in the early 1930s.

Fletcher's method uses two separated omnidirectional microphones in order to pick up time and phase differences of the incoming sound waves. The separation is usually in a range of 0.5 - 1m. The phase difference in turn creates a directional stereophonic image when reproduced over the loudspeakers. This spaced microphone technique is commonly referred to as the time-difference stereophony or A-B. One of the problems with this method is that physical spacing of the microphones raises a risk that some frequencies are recorded out-of-phase and destructive interference may occur. In practice, it is advisable to manipulate the spacing in order to check whether the tonal balance of the reproduced stereophonic sound is satisfactory. Another weakness is that A-B recordings are not mono-compatible. A diagram showing a typical A-B configuration can be seen in Figure 3.2(a).

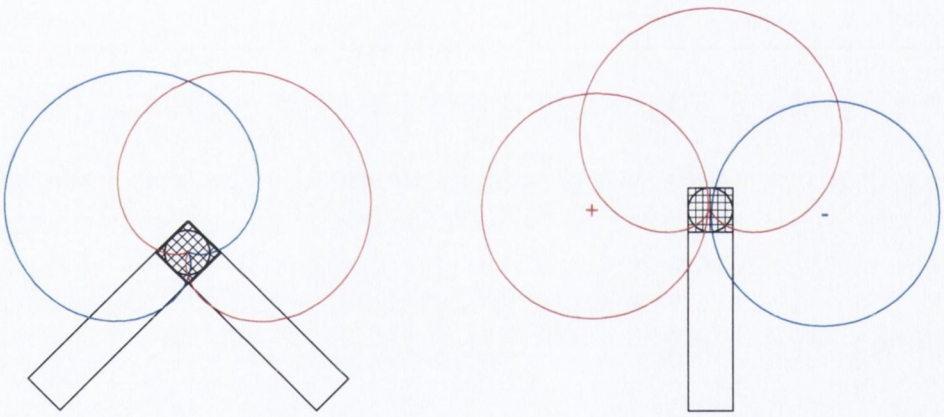
In contrast, Blumlein's approach has started a whole group of so-called coincident microphone techniques. His experiments were mainly concerned with picking up the difference in amplitude of incoming sound waves. For this reasons, two directional microphones are used in close vicinity but pointing in different directions. In its original form, the coincident microphone technique named after Blumlein uses two bi-directional microphones at 90° angle. The amplitude differences are then assured by the on-axis response of one of the microphones and the simultaneous off-axis response of the other one. This configuration is presented in Figure 3.2(b).

There are two popular variations of the original Blumlein's technique. First of them is known as the XY coincident pair [56]. In this configuration, two crossed cardioid microphones are used instead of bi-directional ones. Because of the rejection zone at the back of the array, less of the



(a) (b)

Figure 3.2: Two diagrams showing opposite approaches to stereophonic imaging: (a) Fletcher's spaced AB pair which aims at capturing timing differences between two microphone signals; (b) Blumlein's coincident pair which aims at capturing amplitude differences between two microphone signals



(a) (b)

Figure 3.3: Two popular variations of a coincident microphone setup: (a) XY; (b) Mid-Side (M-S)

reflected acoustic energy (ambience) is recorded. The second one is the Mid-Side configuration [56]. In this technique, a cardioid microphone (Mid or M) is facing forward (optionally it can be replaced with an omnidirectional one) and a bi-directional microphone (Side or S) is directed to the sides, so that its rejection zone is directly in the front. In M-S, the stereophonic images is created by the means of *matrixing* of the M and S signals because in order to derive the loudspeaker signals with this technique, a simple decoding matrix is required:

$$\begin{aligned} L &= M + gS \\ R &= M - gS \end{aligned} \tag{3.1}$$

Where L and R are the loudspeaker signals, M is the Mid microphone input, S is the Side microphone input and and g is a positive scalar gain. We see that the right loudspeaker signal is formed by summing the Mid microphone pickup with the phase-inverted Side microphone pickup. Results obtained with the M-S technique can be in general compared to the results achievable with the XY technique. Summation of M and S microphone signals forms cardioid-like patterns pointing to the left and to the right hand side of the sound stage. The parameter g can be used optionally to adjust the ratio of the M and S input signals. In practice, since the S microphone usually has its rejection zone toward the source direction, it will result in changing the ratio between the direct sound picked up by the Mid and reverberant energy (ambience) picked up by the Side microphones. The XY as well as M-S configurations are illustrated in Figure 3.3(a) & (b).

Somewhere in between spaced and coincident microphone techniques are near-coincident configurations. One of the most popular of them is the ORTF (Office de Radiodiffusion Télévision Française). In this arrangement two cardioid microphones are used as in XY. However, the two cardioids are now spaced $17cm$ apart and form a 110° angle instead of 90°. Generally, with this technique it is possible to obtain more spacious stereophonic images than in the case of XY at the same retaining good definition and locatedness of the sources in the sound stage. It can be attributed to the fact that the small distance of the microphones attempts to mimic the natural spacing of the ears whereas the angular separation creates amplitude changes which to some extent mimic interaural level differences. The ORTF setup is presented in Figure 3.4.

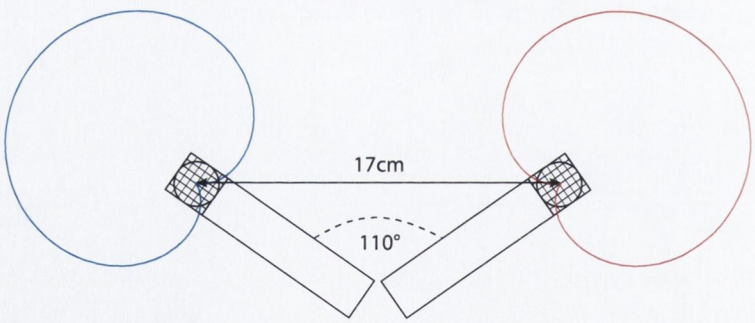


Figure 3.4: Near-coincident ORTF microphone configuration

3.1.2 Multichannel Stereophony

2-channel stereophony has developed into multi-loudspeaker formats which can be broadly classified as "surround sound". Over the last thirty years, the development of various surround

formats by companies such as *Dolby* and *DTS* has brought immersive audio experience to mass consumerism level.

Loudspeaker layouts in multichannel stereophony such as *ITU 5.1* (5 surround loudspeakers typically accompanied by a sub-woofer carrying a low frequency effects channel called "LFE", as specified in [101]) are in general capable of presenting stable phantom image sources for a single, centrally seated listener. These images are reliably created between loudspeakers in the front, with some lateral enhancement from the rear of the array. This can create a feeling of surrounding sound because sources can be reproduced from many directions at the same time. The *ITU* recommended configuration is presented in Figure 3.5, where channels are abbreviated as follows: $L = 30°$ (Left), $R = -30°$ (Right), $C = 0°$ (Centre), $Ls = 110 \pm 10°$ (Left surround), $Rs = -110 \pm 10°$ (Right surround).

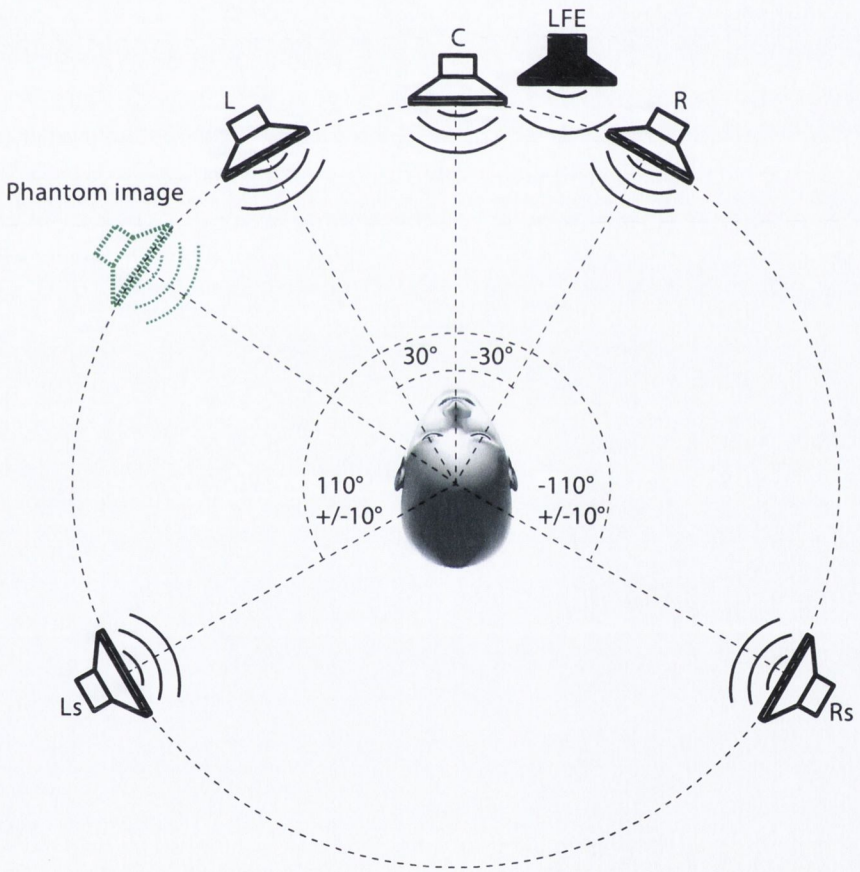


Figure 3.5: Recommended 6-channel configuration for the *ITU 5.1* surround sound reproduction

Multichannel stereophonic techniques are commonly used in cinemas, theatres, installation spaces and at home. Different encoding and decoding schemes support different numbers of loudspeakers as well as their layouts (e.g. "with height" configurations are also supported) and provide diversity of quality and fidelity of recreated sound fields [53, 54]. However, the

biggest disadvantage of all phantom source based systems is that the listener is restricted to the aforementioned "sweet spot". In other words, they do not support listener movements, and even small changes of listening perspective causes the stereo image to collapse, favouring one (or several) of the loudspeakers. This can be attributed mainly to the precedence effect that outbalances often more subtle intensity and timing differences. That is why, e.g. in cinemas, in order to enlarge the listening area as well as to provide a more uniform coverage, the number of surround loudspeakers carrying same channel information is usually increased. Also, a commonly used enhancement technique is to keep all the important parts (such as dialogues or vocals in music) in mono in the center channel only or as a phantom source between the left and right loudspeakers. The latter approach has the advantage that the surround material will not loose important parts of the mix if the playback is done over an incomplete loudspeaker layout.

Typically in multichannel stereophony, monophonic sound sources are amplitude-panned in a pairwise manner. So, it makes the positioning of phantom sources in multichannel stereo very similar to 2-channel mode, with the exception that now pairs of loudspeakers are selected, based on the desired location of the phantom image. Pairwise Constant Power Panning (PCPP) scheme is usually a good choice since it avoids energy fluctuations with the panned source, however other schemes are available as well, good review of which can be found e.g. in [232]. The use of more loudspeakers at a time is also possible and it can result in wider or more diffuse acoustic images.

One thing to note at this occasion is that in pairwise panning techniques, the number of loudspeakers contributing to a single phantom image creation usually varies. For example, only one loudspeaker is needed for reproduction, whenever the virtual sound source location corresponds to the location of the loudspeaker. In practice, it can results in at least two problems. First of all, localisation blur of the sound source will vary depending on whether the reproduction is over one or two (or more) loudspeakers at a time. The difference in localisation will be more dramatic if the spacing between the loudspeakers is significant [184]. For setups like the *ITU 5.1*, the loudspeaker spacing varies across different panning angles so one can intuitively expect variability of localisation blur as well. This problem will be devoted more attention later on in this thesis, when discussing possible strategies of sound field rotations in Chapter 6. Secondly, different number of contributing loudspeakers change the way emitted acoustic waves interfere with each other at the listening point causing audible tonal changes with the panned sources. This is known as "comb filtering effect" and has been already thoroughly investigated in the literature (e.g. [133]). What is important, human listeners are able to detect these tonal changes and assess virtual sources as less natural [105].

Finally, it is also possible to record for a particular multichannel layout using microphone arrays, similarly as in the case of the 2-channel stereophony. A good overview of multichannel recording techniques can be found e.g. in [220]. Although, recording for multichannel reproduction is perfectly applicable for, e.g. surround music production or linear motion picture soundtracks, the possibilities to alter individual sound source properties *post factum* are rather

limited. That is why the importance of multichannel recording techniques to real-time and especially interactive spatial sound reproduction seems to be rather low and it will not be elaborated on in this thesis. Instead, let us proceed to the techniques that have a very broad use in variety of virtual reality applications.

3.1.3 Vector Base Amplitude Panning

Vector Base Amplitude Panning (VBAP) is a vector formulation of the pairwise panning law [182]. In this form, pairwise panning can be applied to arbitrary 2-D loudspeaker arrangements with equidistant loudspeakers. Moreover, layouts with height are also supported and in the case of elevated phantom sources, triplets of reproducing loudspeakers can be selected instead.

In 2-D VBAP, the listening position constitutes a vector base with two unit length vectors $\mathbf{l_1}$ and $\mathbf{l_2}$ pointing toward two contributing loudspeakers. Spatial positioning of a phantom source is achieved by multiplying the unit vectors by adequate gain factors g_i where i is the loudspeaker index. Linear combination of the vectors $\mathbf{l_1}$ and $\mathbf{l_2}$ should create another vector $\mathbf{p}$ pointing in the desired direction of the phantom source. This is shown in Figure 3.6.

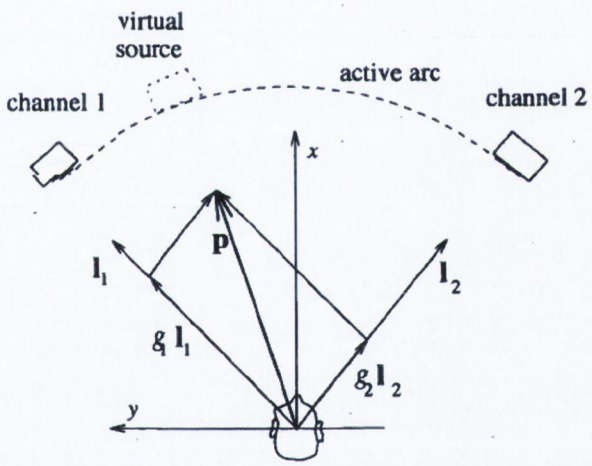


Figure 3.6: 2-channel stereophonic configuration reformulated with vectors, from [182]

When the number of equidistant loudspeakers is increased to three, then the vector $\mathbf{p}$ is calculated from the loudspeaker unit length vectors $\mathbf{l_1}$, $\mathbf{l_2}$ and $\mathbf{l_3}$ instead. This allows for panning in two dimensions (azimuth and elevation). In this situation, the active panning arc that was used on the horizontal plane (the curve joining two loudspeakers on which the sound source can be positioned) becomes a fragment of a sphere. This situation is shown in Figure 3.7.

So, in a generalised 3-D case, the phantom source vector $\mathbf{p}$ can be expressed as:

$$\mathbf{p} = g_1\mathbf{l_1} + g_2\mathbf{l_2} + g_3\mathbf{l_3} = \mathbf{gL} \tag{3.2}$$

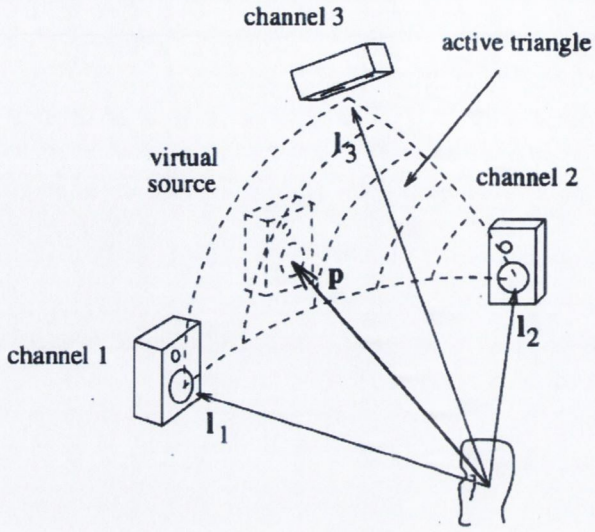


Figure 3.7: 3-channel stereophonic configuration reformulated with vectors, from [182]

where $\mathbf{g} = [g_1 \quad g_2 \quad g_3]$ and $\mathbf{L} = [\mathbf{l}_1 \quad \mathbf{l}_2 \quad \mathbf{l}_3]^T$. In order to obtain gain coefficients $\mathbf{g}$ for a particular phantoms source direction described by $\mathbf{p}$ we have to perform:

$$\mathbf{g} = \mathbf{p}\mathbf{L}^{-1} \tag{3.3}$$

VBAP theory can be reformulated for any number of speakers and their arbitrary spatial configurations. In order to position the sound source correctly, it is first necessary to select a correct active arc or fragment of a sphere defined by a pair or a triplet of loudspeakers. Then, the phantom image is created using appropriate loudspeakers. When an arbitrary loudspeaker configuration is used and when distances to the sweet spot are not equal, it is important to calibrate the system so that the initial time delays and level differences from all the loudspeakers does not exist in the sweet spot.

To sum up, VBAP can be considered as a method of extending a pairwise stereophony to another spatial dimension (using loudspeaker triplets) and arbitrary loudspeaker arrangements. Thus, similarly to other panning methods, the localisation blur with VBAP varies depending on the phantom source position and the loudspeaker separation. Also, problems with tonal colouration occur when changing the number of contributing loudspeakers. Some methods have been proposed in order to minimise the latter effect, e.g. in [183] where source is rendered at multiple close locations so that one loudspeaker never reconstructs the virtual source. The obvious disadvantages of this method is however the increased number of channels per virtual source and deterioration of localisation accuracy by increasing the source's spread.

3.2 Wave Field Synthesis

Unlike stereophony, Wave Field Synthesis (WFS), which was first proposed by Berkhout in 1988 [22], attempts to recreate the actual sound field within the large listening area. This is the holophonic approach to sound field reconstruction which attempts to reconstruct acoustic pressure as well as its gradient at each point inside the listening space. That is why localisation of sources is implicit (since the pressure field creates necessary ILD and ITD cues) and also stable with the listener's movements. So, the immediately recognised advantage over the stereophony is that it is no longer restricted to the "sweet spot". The wavefronts are also reconstructed with their correct curvature that reflects the physical distance to the sound source.

The approach in Wave Field Synthesis is to attempt to recreate original wave fronts of the original sound sources. The theoretical basis behind it is contained in the Huygens principle, which states that the wave front of the primary source can be reconstructed by substituting it with an infinite number of secondary sources. The Huygens principle is illustrated in Figure 3.8.

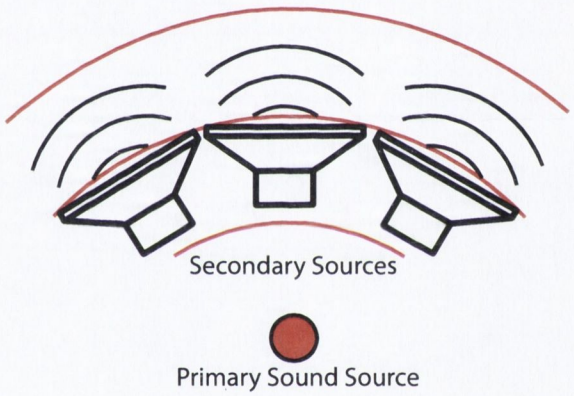


Figure 3.8: An illustration of the Huygens principle where primary source wavefronts are reconstructed by secondary sound sources

A theoretical formulation of the Wave Field Synthesis was introduced by Berkhout [22] by the means of a mathematical expression known as the Kirchhoff-Helmholtz integral. This theorem states that the sound pressure can be calculated at any listening point within a source-free volume V if both the sound pressure and its gradient are known on the surface S enclosing the volume V. Under such conditions, the pressure field can be calculated as:

$$P(\mathbf{r},\omega) = \frac{1}{4\pi} \oint \left[P(\mathbf{r_s},\omega) \frac{\delta}{\delta \mathbf{n}} \left(\frac{e^{-jk|\mathbf{r}-\mathbf{r}_s|}}{|\mathbf{r}-\mathbf{r}_s|} \right) - \frac{\delta P(\mathbf{r}_s,\omega)}{\delta \mathbf{n}} \left(\frac{e^{-jk|\mathbf{r}-\mathbf{r}_s|}}{|\mathbf{r}-\mathbf{r}_s|} \right) \right] dS \tag{3.4}$$

where $P(\mathbf{r},\omega)$ is the Fourier transformed pressure at an arbitrary point $\mathbf{r}$ within the surface S and at the angular frequency ω, r_s is the positional vector of a point source outside the surface

S, k is the wavenumber and $\mathbf{n}$ is the surface normal.

So, in other words, if each point on the surface S is substituted by a secondary sound source, by using the principles of the acoustic wave propagation, it is possible to calculate the pressure and its gradient at any point in the listening space. The illustration of the Kirchhoff-Helmholtz theorem is presented in Figure 3.9.

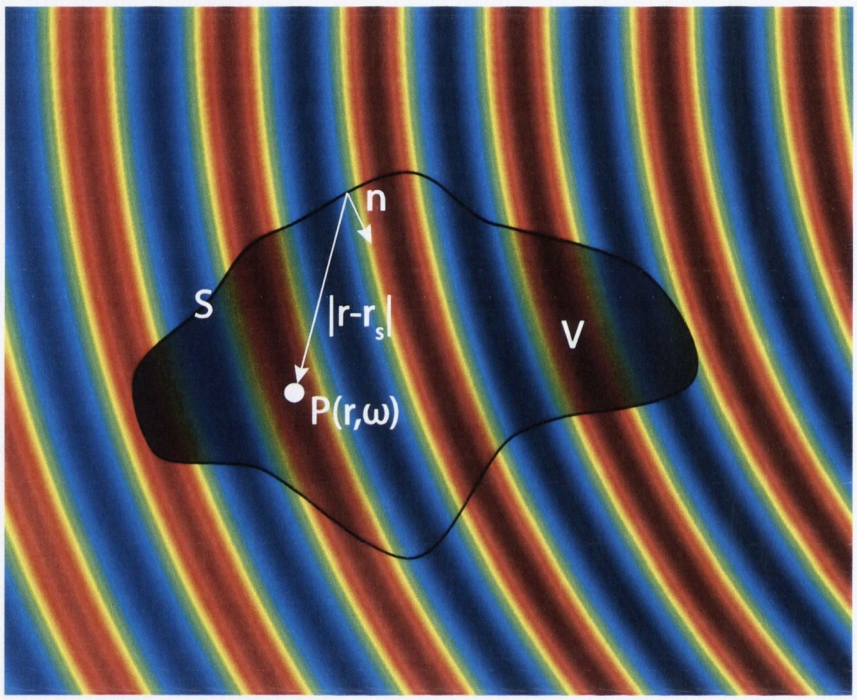


Figure 3.9: Illustration of the Kirchoff-Helmholtz integral theorem

We can see that indeed the equation 3.4 contains the terms for both acoustic pressure at the surface and its gradient. That is why, the reconstruction of the pressure field $P(\mathbf{r}, \omega)$ would normally require contributions of both monopole and dipole secondary sources arranged around the whole surface.

More so, in theory, a volumetric sound field reproduction would require an infinite number of secondary sources evenly distributed across the whole surface area. This restriction can be relaxed if we consider that the reproduction needs only be accurate up to a particular frequency dictated by our perceptual limitations (usually around $20kHz$). But still, for most practical applications the volumetric reproduction is rather infeasible and that is why a very common simplification is to consider a horizontal plane reproduction instead. Under such degradation, the Kirchhoff-Helmholtz integral simplifies to Rayleigh I (Equation 3.5) & II (Equation 3.6) integrals that describe the pressure field as a result of action of monopole & dipole secondary sources respectively [30]:

$$P(\mathbf{r},\omega) = \rho_a c \frac{jk}{2\pi} \iint \left[V_n(\mathbf{r}_S,\omega) \frac{e^{-jk|\mathbf{r}-\mathbf{r}_s|}}{|\mathbf{r}-\mathbf{r}_s|} \right] dydz \tag{3.5}$$

$$P(\mathbf{r},\omega) = \frac{jk}{2\pi} \iint \left[P(\mathbf{r}_S,\omega) \frac{1+jk|\mathbf{r}-\mathbf{r}_s|}{jk|\mathbf{r}-\mathbf{r}_s|} \cos\phi \frac{e^{-jk|\mathbf{r}-\mathbf{r}_s|}}{|\mathbf{r}-\mathbf{r}_s|} \right] dydz \tag{3.6}$$

where $V_n(\mathbf{r}_S,\omega)$ is the acoustic velocity in the surface normal direction and $\rho_a c$ is the specific acoustic impedance of the air. What we infer from Rayleigh I & II integrals is that the use both monopole and dipole sources is redundant and it suffices to use only monopole sources reproducing pressure gradient signals (as in formula Rayleigh I) [163].

In fact, practical implementation would also necessitate spatial discretisation of the secondary sources. That would modify the Rayleigh I equation to:

$$P(\mathbf{r},\omega) = \rho_a c \frac{jk}{2\pi} \sum_{i=1}^{N} \left[V_n(\mathbf{r}_i,\omega) \frac{e^{-jk|\mathbf{r}-\mathbf{r}_i|}}{|\mathbf{r}-\mathbf{r}_i|} \right] \Delta y \Delta z \tag{3.7}$$

where i denotes the i^{th} secondary monopole source N is the total number of secondary sources and $\Delta y \Delta z$ denotes the spacing between sources in the horizontal and vertical directions[1].

From the above discussion, assuming a horizontal line array of reproducing sources (which is the most common and the most practical scenario) as well as the source-array-listener geometry as presented in Figure 3.10, the driving signals for each of the loudspeakers can be calculated as [143]:

$$Q_i(\mathbf{r}_i,\omega) = S(\omega) \frac{cos\beta}{G_i(\beta,\omega)} \sqrt{\frac{jk}{2\pi}} \sqrt{\frac{|x_l - x_i|}{|x_l - x_m|}} \frac{e^{-}jk|\mathbf{r}_i - \mathbf{r}_m|}{\sqrt{|\mathbf{r}_i - \mathbf{r}_m|}} \tag{3.8}$$

Where $S(\omega)$ is the frequency domain representation of the source signal and $G_i(\beta,\omega)$ is the gain correction due to non-spherical source radiation. The generalised time-domain driving signals can be obtained by applying inverse Fourier transform to Equation 3.8, which yields:

$$q_i[k] = g_i(h[k] * s[k]) * \delta[k - \kappa] \tag{3.9}$$

[1]By convention in WFS the z-axis is usually the depth axis. However, in order to be consistent in this work, we use the co-ordinate system where positive z is pointing upward, x is pointing forward and y is pointing leftward (from the listener's perspective). For clarity, the co-ordinate system used is shown in Figures 3.10 and 3.11

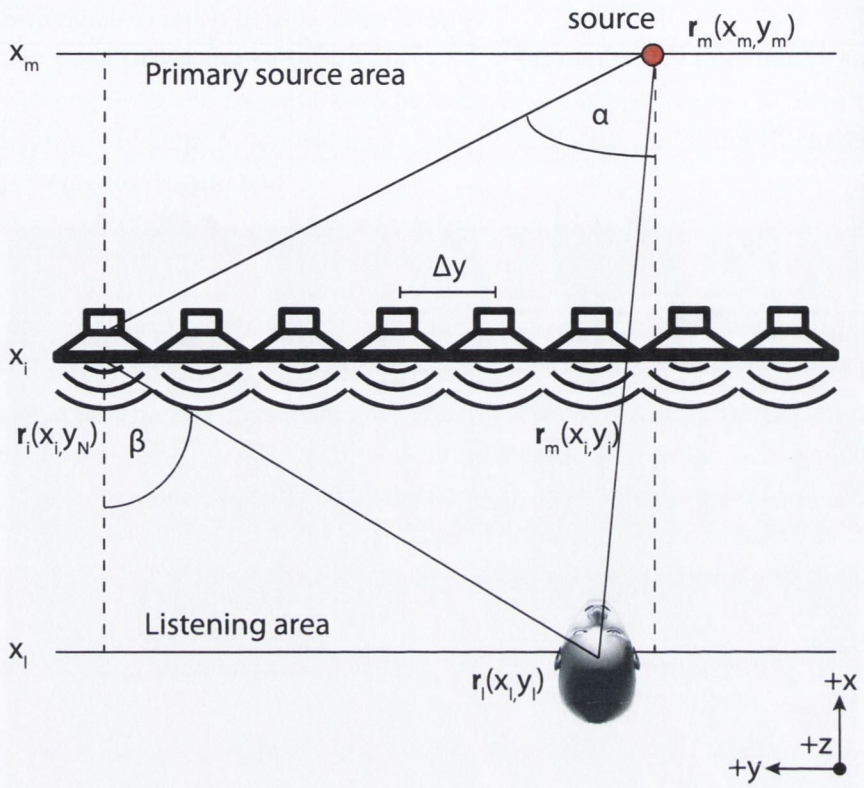


Figure 3.10: Geometry used for calculation of the horizontal-only driving functions, [143]

where g_i is the gain factor, $h[k]$ is the inverse Fourier transform of the $\sqrt{\frac{jk}{2\pi}}$ term (a filter causing $+3dB$ per octave boost) and $\delta[k-\kappa]$ is κ-samples time delay.

The physical impossibility of implementing an infinite number of secondary sources causes some artefacts. First of all, due to array truncation the listening area can indeed be large but still somewhat restricted. Secondly, the finite number of loudspeakers imposes restrictions in terms of the highest frequency that is possible to reproduce without artefacts. Large loudspeaker separation leads to the effect known as "spatial aliasing". It means that above a particular frequency known as the *spatial aliasing frequency*, the actual waveform is not reconstructed correctly.

An acoustic wave incidence angle as well as the angle of listening position on the receiving side influence the spatial aliasing frequency and can be calculated from the following equation:

$$f_a = \frac{c}{\Delta y|\sin(\phi_m) - \sin(\phi_l)|} \tag{3.10}$$

where f_a denotes spatial aliasing frequency in $[Hz]$, Δy is the loudspeaker spacing in $[m]$, ϕ_m

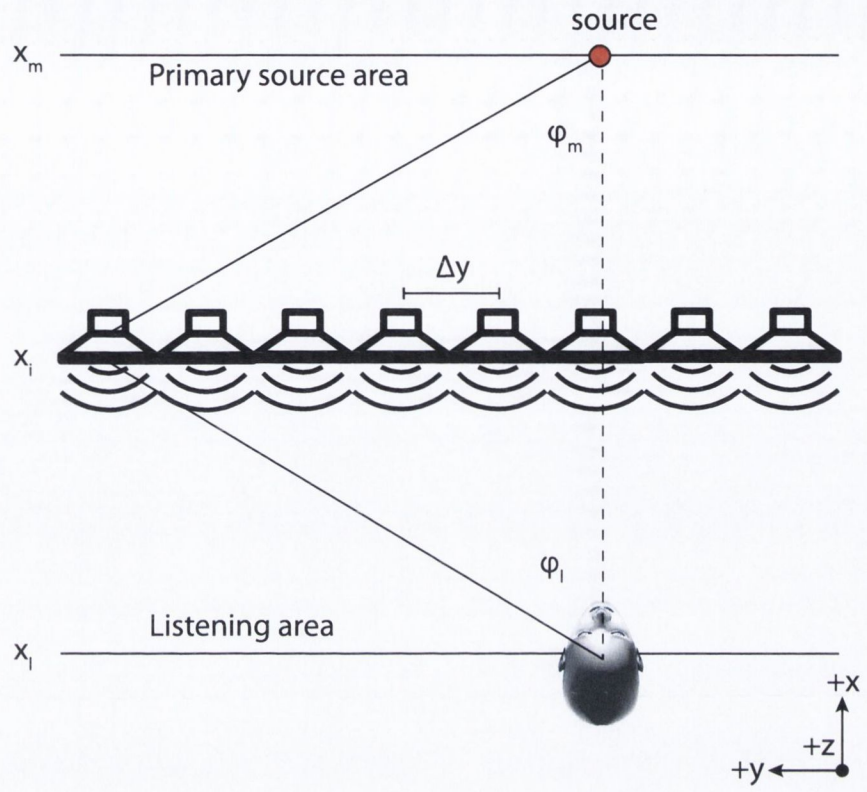


Figure 3.11: Angles used in spatial aliasing calculation

and ϕ_l are the maximum angles on the source and reproduction sides respectively, as illustrated in the Figure 3.11.

From the Equation 3.10 and its illustration in Figure 3.11 it can be deduced that the spatial aliasing cannot really be avoided unless we consider a theoretical case of an infinitely remote virtual sound source and infinitely distant listening point. In such a scenario, the incident wave will be a plane wave, i.e. perpendicular to all the loudspeakers in the array. For all other situations, when dealing with finite distances, it is usually recommended to consider the worst case scenario when a reconstructed sound source is localised on the array and the listening distance is very close. Then, the lowest possible spatial aliasing frequency can be calculated based solely on the loudspeaker spacing:

$$f_{a_{min}} = \frac{c}{2\Delta y} \tag{3.11}$$

Thus, when the spacing between speakers is, for example, equal to $10cm$ the spatial aliasing frequency is approximately $1.7kHz$. So, in this case, wave fields will be reconstructed correctly only below this frequency. However, Varheijen [163] suggests that the degradation is not severe

until the aliasing frequency drops below $1.5kHz$. In all other cases, there have been numerous methods proposed to deal with this problem, e.g. by using a hybrid system combining WFS at low frequencies and stereophony at mid- and high frequencies [237].

In conclusion, WFS (as well as its hybrid derivatives), offers a noteworthy approach to acoustic holophony, especially as long as the wide-area, faithful sound field reproduction is concerned. Large audience playback like in sound reinforcement, movie theatres and installations are definitely in the range of the its optimal applicability. However, discrete channels and loudspeakers needed are usually way too numerous in order for the system to be considered an optimal for immersive real-time applications, like video games. Its implementation also imposes many practical problems from the point of view of budget and physical space, so it is quite unlikely that the system targets at mass consumer market any time soon.

However, before we ultimately abandon the idea of acoustic holophony we shall investigate the other aforementioned approach to sound field reconstruction which is based on superposition of plane waves.

3.3 Ambisonics

Ambisonics is deeply rooted in the principles of the early Blumlein's work, discussed formerly in Section 3.1. In its simplest form, it can be thought of as a spatial extension of the coincident microphone techniques, particularly the combination of a Blumlein pair and the Mid-Side technique. It is centred around the idea that carefully aligned pressure and pressure-gradient microphones are capable of probing acoustic pressure and its time derivative in the full three dimensional space surrounding them. By recording only four channels of information it is then possible to create acoustic images all around the listener. What is more, because of the uniform treatment of the information around the sensors, all the directions are given equal priority.

Due to the *matrixing* approach, the decoding stage is fully separated from the encoding stage, allowing for the material to be reproduced using custom loudspeaker layouts. What is noteworthy, at the time of its invention these intrinsic features of Ambisonics could already have been considered superior to its predecessor technologies, including quadraphony. In fact, with Ambisonics it was possible to overcome some of the problems of quadraphony, particularly when it comes to the lateral imaging. Moreover, it was also possible to include the height information [72]. Nevertheless, the commercial success of Ambisonics was rather dubious in contrast to the pan-potted multi-channel stereophony.

However, over the years Ambisonics has grown into a considerable area of research providing a cohort of invaluable tools for studies on topics such as sound field reconstruction, the directivity of sound sources, Head Related Impulse Responses, auditory distance coding or perception of spatial sound. This growth has to be in huge part assigned to its elegant and rigorous mathematical scaffold, first introduced by British academics Gerzon, Fellget and Burton [63,72]. These foundations led to the development of the Higher Order Ambisonics (HOA) concept in

the 1990s, which is very central in this thesis. As will be explained later in Section 3.4, HOA can be used to reproduce sound fields with arbitrary levels of accuracy and as such, it constitutes a very attractive tool for a researcher testing the impact of the audio rendering system on certain perceptual aspects of sound.

3.3.1 Encoding Ambisonics

Let us imagine that it is possible to place four ideal microphones at exactly the same point in space. First is the omnidirectional microphone whose pick-up amplitude we reduce by $3dB$. The remaining are the bi-directional microphones facing forward, leftward and upward respectively. With this configuration we record four audio signals that are referred to as B-Format Ambisonics - the format that is conventionally used to store and manipulate Ambisonic recordings. The aforementioned hypothetical recording setup is depicted in Figure 3.12.

However, since it is a physical impossibility to arrange sensors in the way required by the B-Format standard, in practice other configurations are used in conjunction with the necessary post-processing in order to come out with the B-Format signals. One such solution has been proposed and implemented by Gerzon and Craven using near-coincident sub-cardioid microphone arrays arranged on a surface of the tetrahedron [43, 73, 74]. Tetrahedral microphones are commercially available from e.g. *SoundField Ltd.* [209] or *Core Sound LCC* [41], which are presented in Figure 3.13.

Tetrahedral microphones are capable of recording so-called A-Format Ambisonic signals. The B-Format signals can be derived from them using the following conversion:

$$\begin{aligned} W &= 0.5(LF + LB + RF + RB) \\ X &= 0.5((LF - LB) + (RF - RB)) \\ Y &= 0.5((LF - RB) - (RF - LB)) \\ Z &= 0.5((LF - LB) + (RB - RF)) \end{aligned} \tag{3.12}$$

where LF, LB, RF and RB refer to left-front, left-back, right-front and right-back sub-cardioid capsules of the tetrahedral array.

Thus, one way for spatial recordings to be encoded into the Ambisonics domain is to record the acoustic events using the tetrahedral microphone arrays with further post-processing according to the Equation 3.12. However, it is also noteworthy, that B-format spherical characteristics can be described using simple analytical functions:

$$\begin{aligned} W(\phi, \theta) &= \tfrac{1}{\sqrt{2}} \\ X(\phi, \theta) &= \cos(\phi)\cos(\theta) \\ Y(\phi, \theta) &= \sin(\phi)\cos(\theta) \\ Z(\phi, \theta) &= \sin(\theta) \end{aligned} \tag{3.13}$$

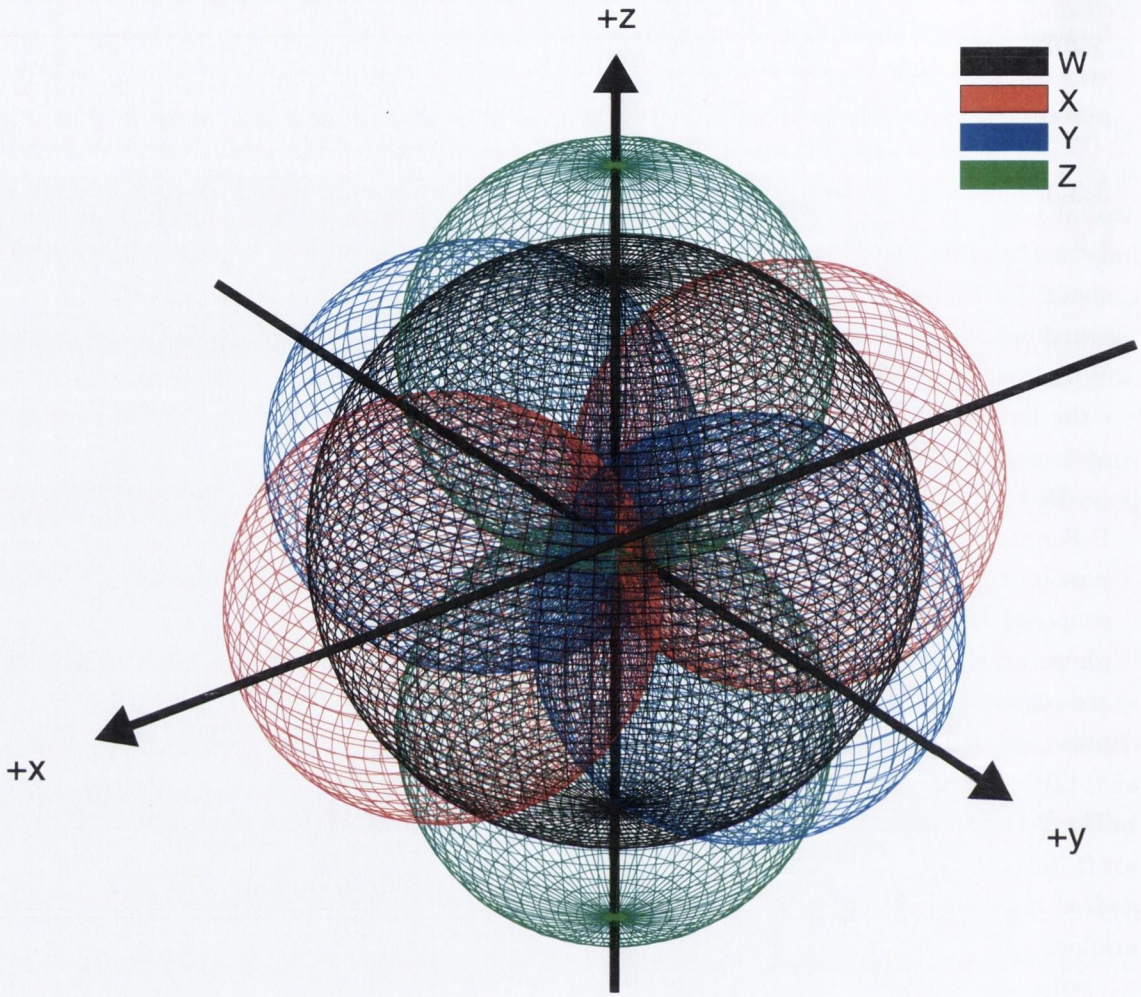
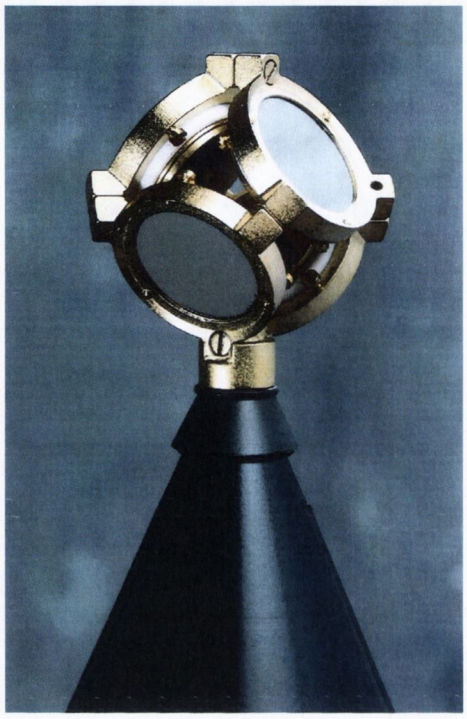


Figure 3.12: Spatial configuration of four microphone pick-up patterns required to obtain B-Format Ambisonic signals. W is the omnidirectional component whose pick-up amplitude is reduced by $3dB$. All three remaining microphones are bi-directional, directed along the positive x, y and z axis respectively.

where is ϕ, $[0 \leq \phi < 2\pi]$ is the azimuth angle (counted from the positive x direction in the anticlockwise manner) and θ, $[-\frac{\pi}{2} \leq \theta < \frac{\pi}{2}]$ is the elevation angle (counted from the equator toward positive z direction - upward, and the negative z direction - downward).

The coordinate system used in this work comprises x, y and z axes pointing to the front, left and up respectively, ϕ is the azimuthal angle with the anti-clockwise rotation and θ is the elevation angle form the x-y plane. Spherical coordinate system used in this thesis is presented in Figure 3.14. The conversion between spherical and cartesian coordinates is defined as follows:



(a)

(b)

Figure 3.13: (a) Soundfield tetrahedral capsule (from [209]); (b) Core Sound TetraMic (from [41])

$$\begin{aligned} x &= \|\vec{r}\| \cos(\phi) \cos(\theta) \\ y &= \|\vec{r}\| \sin(\phi) \cos(\theta) \\ z &= \|\vec{r}\| \sin(\theta) \end{aligned} \tag{3.14}$$

The encoding process can be then applied by multiplying the anechoically acquired audio source material with the B-Format spherical patterns for the desired incidence angle. This process is fairly straightforward and also allows for the close-microphoned sources (using more relaxed definition of *anechoic*) to be encoded into Ambisonic sound field representation *post factum* by applying:

$$\begin{aligned} w(t) &= s(t) \times \frac{1}{\sqrt{2}} &&= s(t) \times W(\phi_S, \theta_S) \\ x(t) &= s(t) \times \cos(\phi_S) \cos(\theta_S) &&= s(t) \times X(\phi_S, \theta_S) \\ y(t) &= s(t) \times \sin(\phi_S) \cos(\theta_S) &&= s(t) \times Y(\phi_S, \theta_S) \\ z(t) &= s(t) \times \sin(\theta_S) &&= s(t) \times Z(\phi_S, \theta_S) \end{aligned} \tag{3.15}$$

where $w(t)$, $x(t)$, $y(t)$ and $z(t)$ are the B-Format signal vectors, $s(t)$ is the pressure signal of

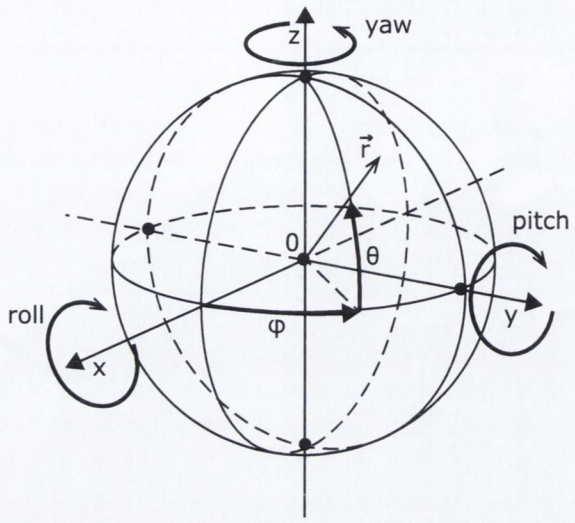


Figure 3.14: Spherical coordinate system used in this thesis, from [47]

a sound source to be encoded into direction given by ϕ_S and θ_S.

3.3.2 Sound Field Transformations

In Ambisonics, B-Format signals provide an orthogonal decomposition of the auditory scene. Because of that, Ambisonically encoded sound sources are added (or mixed) together by summing their corresponding $w(t)$, $x(t)$, $y(t)$ and $z(t)$ channels. We will show now that whole 3-D auditory scenes encoded into their B-Format representations can be subjected to various transformation operations including mirroring, focusing, rotating, tilting and tumbling. Importantly, the transformations will affect all the sources in the scene and there is no need to process individual sources. What it means in practice is that the desired spatial effect can be often achieved with the minimum use of computational resources.

Before we talk about each one of the aforementioned transformations in more detail, it is useful to introduce a vector notation of the B-format signals. Let us define $\mathbf{b}$ being the column vector containing the current samples of the B-Format channels $w[n]$, $x[n]$, $y[n]$ and $z[n]$:

$$\mathbf{b} = \begin{bmatrix} w[n] \\ x[n] \\ y[n] \\ z[n] \end{bmatrix} = \begin{bmatrix} s[n] & \times & W \\ s[n] & \times & X \\ s[n] & \times & Y \\ s[n] & \times & Z \end{bmatrix} = \begin{bmatrix} s[n] & \times & \frac{1}{\sqrt{(2)}} \\ s[n] & \times & \cos(\phi_S)\cos(\theta_S) \\ s[n] & \times & \sin(\phi_S)\cos(\theta_S) \\ s[n] & \times & \sin(\theta_S) \end{bmatrix} \tag{3.16}$$

Then each sound field transformation can be written in the form:

$$\mathbf{b_T} = \mathbf{T}\mathbf{b} \tag{3.17}$$

where $\mathbf{b_T}$ is a column vector with the transformed B-Format signals and $\mathbf{T}$ is a 4×4 transformation matrix.

Mirroring is probably the simplest form of a B-Format transformation and it simply concerns inverting the phase of certain B-Format signals in order for the source(s) to be "mirrored" against xy, yz, xz planes or the coordinate system's origin. To perform this action we need to construct a matrix $\mathbf{T_M}$ with non-zero terms along its diagonal. For example, in order to mirror sound sources against the xz plane, the phase of $\mathbf{y}$ signal needs to be inverted which can be achieved with the following transformation matrix $\mathbf{T_M}$:

$$\underbrace{\begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix}_M}_{\mathbf{b_M}} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\mathbf{T_M}} \underbrace{\begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix}}_{\mathbf{b}} \tag{3.18}$$

where we have simplified the notation of time domain B-Format signals to simply lowercase w, x, y and z.

Focusing, (sometimes also referred to as *zooming* or *forward dominance*) was proposed by Gerzon and Barton [76] as a method of emphasising the sound content at the front of the array and attenuating the content at the back. The original method was developed based on Lorentzian transformations and can be applied using the following transformation matrix:

$$\underbrace{\begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix}_{FD}}_{\mathbf{b_{FD}}} = \underbrace{\begin{bmatrix} \frac{1}{2}(\lambda + \lambda^{-1}) & \frac{1}{\sqrt{8}}(\lambda + \lambda^{-1}) & 0 & 0 \\ \frac{1}{\sqrt{2}}(\lambda + \lambda^{-1}) & \frac{1}{2}(\lambda + \lambda^{-1}) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\mathbf{T_{FD}}} \underbrace{\begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix}}_{\mathbf{b}} \tag{3.19}$$

where the frontal scene is emphasised using the gain factor λ and the rear scene is attenuated by its inverse factor $\frac{1}{\lambda}$. Anderson [10] refers to this transform simply as *dominance* and besides that proposes three variations thereof (*focus*, *push* and *press*). It is sufficient to say that they differ in a way the front scene is modified with respect to the rear scene.

One of the most useful transformations from the point of view of a real-time manipulation of the sound field is *rotation* around the z, y and x axis (in this thesis referred to as *yaw*, *pitch* and *roll* respectively, as introduced in Figure 3.14). The matrices required to perform these rotations by a given angle α are given by:

$$\mathbf{R_z} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) & 0 \\ 0 & \sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{3.20}$$

$$\mathbf{R_y} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\alpha) & 0 & -\sin(\alpha) \\ 0 & 0 & 1 & 0 \\ 0 & \sin(\alpha) & 0 & \cos(\alpha) \end{bmatrix} \tag{3.21}$$

$$\mathbf{R_x} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & 0 & \sin(\alpha) & \cos(\alpha) \end{bmatrix} \tag{3.22}$$

where $\mathbf{R_z}$, $\mathbf{R_y}$ and $\mathbf{R_x}$ denote the yaw, pitch and roll respectively. After the transformation, the B-Format signals need to be eventually decoded for a given loudspeaker layout. This process is explained next.

3.3.3 Decoding Ambisonics

Decoding Ambisonics consists in finding signal feeds for loudspeakers arranged on a surface of the imaginary sphere in the directions given by the angles Φ_i and Θ_i, where i denotes the i^{th} loudspeaker. In theory, Ambisonics allows for the use of arbitrary loudspeaker arrangements. In practice, best results can be achieved using uniform distributions (e.g. using vertices or faces of Platonic solids) and diametrically opposite pairs [47, 72].

In general, loudspeaker feeds can be calculated as a weighted sum of all the Ambisonic channels (w, x, y and z). Weighting coefficients are obtained by evaluating the B-format functions from Equation 3.13 at the given angles Φ_i and Θ_i:

$$\begin{aligned} l_{w_i} &= \tfrac{1}{\sqrt{2}} \\ l_{x_i} &= \cos(\Phi_i)\cos(\Theta_i) \\ l_{y_i} &= \sin(\Phi_i)\cos(\Theta_i) \\ l_{z_i} &= \sin(\Theta i) \end{aligned} \tag{3.23}$$

Then, the i^{th} loudspeaker feed can be calculated as

$$\mathbf{l_i} = \frac{1}{N}\left[l_{w_i}w + l_{x_i}x + l_{y_i}y + l_{z_i}z\right] \tag{3.24}$$

where N is the total number of loudspeakers in the array. However, sometimes it is more convenient to define the $N \times 4$ decoding matrix $\mathbf{D}$ which, when multiplied with the column vector $\mathbf{b}$ from the Equation 3.16, delivers another column vector $\mathbf{g}$ with N loudspeaker gains:

$$\underbrace{\begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_i \\ \vdots \\ g_N. \end{bmatrix}}_{\mathbf{g}} = \underbrace{\frac{1}{N}\begin{bmatrix} l_{w_1} & l_{x_1} & l_{y_1} & l_{z_1} \\ l_{w_2} & l_{x_2} & l_{y_2} & l_{z_2} \\ \vdots & \vdots & \vdots & \vdots \\ l_{w_i} & l_{x_i} & l_{y_i} & l_{z_i} \\ \vdots & \vdots & \vdots & \vdots \\ l_{w_N} & l_{x_N} & l_{y_N} & l_{z_N} \end{bmatrix}}_{\mathbf{D}} \underbrace{\begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix}}_{\mathbf{b}} \tag{3.25}$$

This process is sometimes referred to as decoding through projection and works for regular loudspeaker layouts (e.g. regular polygons or based on Platonic solids).

Another method of constructing the decoding matrix emerges as soon as we appreciate that at the centre of the loudspeaker array we expect that the $\mathbf{b}$ signals are reconstructed correctly by N loudspeakers. Using the previously introduced notation, we can write that as:

$$\mathbf{b} = \mathbf{L}\mathbf{g} \tag{3.26}$$

where $\mathbf{L}$ is the $4 \times N$ matrix with the spherical harmonic representation of the loudspeaker locations and $\mathbf{g}$ is the column vector with the loudspeaker gains. It follows that

$$\mathbf{g} = \mathbf{L}^{-1}\mathbf{b} \tag{3.27}$$

but only if the matrix $\mathbf{L}$ is invertible, which is only possible when the number of loudspeakers is equal to the number of the B-Format channels. When $N > 4$ we can approximate the inversion using Moore-Penrose pseudo-inverse, which gives $\mathbf{D} = \mathbf{L}^{\dagger}$ where

$$\mathbf{L}^{\dagger} = \mathbf{L}^{\mathbf{T}}(\mathbf{L}\mathbf{L}^{\mathbf{T}})^{-1} \tag{3.28}$$

However, care must be taken when using this method with irregular loudspeaker arrangements as resulting decoders may be unstable (e.g. due to numerical instability [221]). Also, when decoding for arrangements with more loudspeakers than ambisonic channels, decoding becomes an underdetermined problem and may lead to less than optimal solutions [221]. For regular arrangements though and whenever the number of B-Format channels is equal to the number of the loudspeakers, the pseudoinverse method should return equivalent gains as the method of projection.

Whatever the method used, the decoding process can be viewed as (beam)forming of a virtual microphone that points to the reproduced sound source's direction. Similarly, the "sensitivity" of the microphone at a given loudspeaker angle Φ_i and Θ_i will determine the extent to which the given loudspeaker will contribute to the sound field reproduction. This situation is illustrated in the Figure 3.15. On the left hand column, there are 2-D virtual microphone polar patterns for a sound source encoded at 25°. The right hand column contains the unwrapped gain curves for a pantophonic loudspeaker system. Loudspeaker gains can be read on the y axis. As an example, on the x axis, angular locations have been marked for 8 loudspeakers arranged in a regular, octagonal shape.

One can note that the final microphone directivity characteristic will depend on how each of the B-Format channels contributes to the beamforming process. For example, the supercardioid virtual microphone characteristic results from a straightforward, unaffected combination of the pressure and velocity components of the B-format signals and when we account for the fact that the W component of a sound field is by convention attenuated by $3dB$ with respect to the directional components. In the literature, this type of decode is referred to as *velocity* (e.g. [105,234]) or *basic* decode [47]. However, other characteristics of the virtual microphone(s) are possible as well. In general, by changing the ratio between the pressure to velocity signals, we can form an infinite number of directional characteristics ranging from purely omnidirectional (all velocity components reduced to 0) to bi-directional (the omnidirectional component reduced to 0). This can be achieved by adjusting the directivity parameter d in the following equation being a modification of Equation 3.24:

$$l_i = \frac{1}{N}\left[\sqrt{2}(2-d)l_{w_i}w + d(l_{x_i}x + l_{y_i}y + l_{z_i}z)\right] \tag{3.29}$$

In practice, there are three most popular decoder types, for the reasons that will be discussed shortly. They are the velocity (or basic) decode ($d = \frac{1}{\sqrt{2}}$), energy decode ($d = \frac{1}{\sqrt{3}}$) and in-phase (or cardioid) decode ($d = 0.5$). It is worth noting that the in-phase (or cardioid) decode is a special kind of decode in which there are no out-of-phase signals (played back by the opposite loudspeakers of the array).

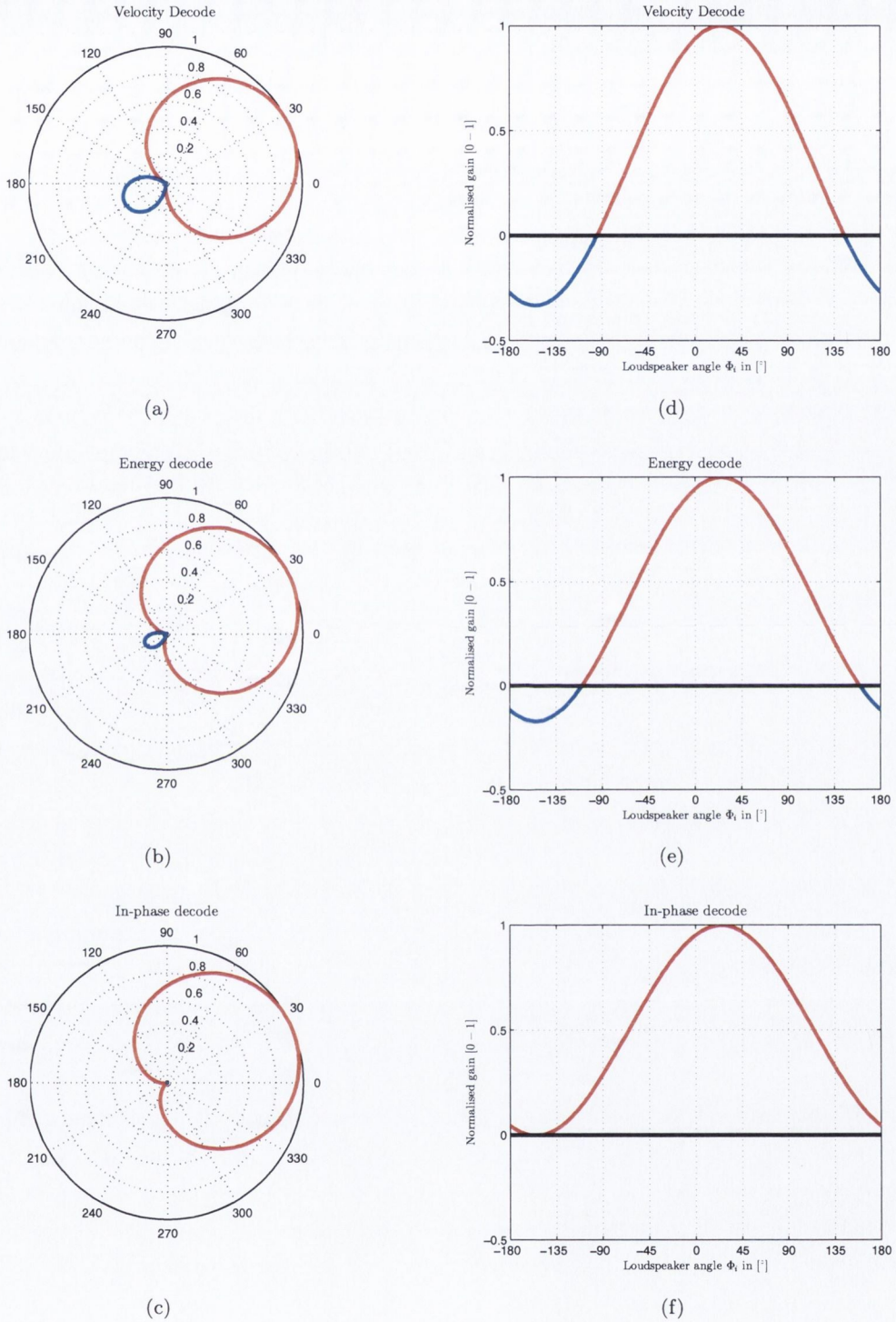


Figure 3.15: Virtual microphones resulting from the sound source at 25° and three different decoder types (velocity, energy and in-phase). Red colour denotes positive gains whereas blue colour denotes negative gains. Loudspeaker gains can be read from the right hand column curves for a given angle Φ_i.

Before the different decoder types can be fully understood, it is necessary to explain the Gerzon's criteria for a decoder to be considered as Ambisonics. For this reason we will refer to Gerzon's localisation theory that would provide some useful guidelines in the decoder design process.

3.3.3.1 Psychoacoustic decoder optimisation

Every acoustic field can be characterised with two quantities: acoustic pressure and particle velocity. Craven [42] pointed out that the pressure signal from any sound source can be reconstructed correctly at the ear, regardless of whether the direction of the virtual source and the loudspeaker(s) coincide. It is because pressure is a scalar quantity that is not associated with any direction. However, this will not induce correct ILD unless both the virtual source and the loudspeaker coincide. Nevertheless, the ILD can be simulated using two or more loudspeakers. The measure of how well they perform in achieving this goal can be obtained by assessing the energy of the signals that they generate.

On the other hand, acoustic velocity is a vector quantity whose direction depends on the direction of oscillatory motion of the air particles. That is why, acoustic velocity can be only reconstructed correctly if the virtual source and the reproducing loudspeaker are in the same direction. However, this motion can also be simulated using two or more loudspeakers.

In Chapter 2 we explained two main phenomena (ITD and ILD) that help us determine the direction of the impinging sound wave. It was emphasised that the phase difference at the ears is very informative at low frequencies (i.e. when the wavelengths are greater than the diameter of the head), however quickly become ambiguous when the wavelengths become short. In contrary, intensity differences are more effective at high frequencies due to the shadowing effect of the head. Accordingly, the acoustic energy should be regarded as a good estimator of localisation for mid and high frequencies and acoustic velocity should be effective in determining low frequency sounds locations.

Based on this elementary psychoacoustic knowledge, Gerzon [76] formulated a set of objective parameters or predictors of localisation in multi-loudspeaker systems. These are known as energy and velocity vectors.

Energy and velocity vectors are calculated for a given set of loudspeaker gains in a multi-channel audio system. We can distinguish the vector's components in the x, y and z directions respectively. However, for simplicity, in this discussion we will restrict ourselves to horizontal only reproduction, so that the energy vector is defined as:

$$\mathbf{e} = [e_x \quad e_y] \tag{3.30}$$

$$e_x = \sum_{i=1}^{N} \frac{g_i^2 \cos(\Phi_i)}{P_e} \tag{3.31}$$

$$e_y = \sum_{i=1}^{N} \frac{g_i^2 \sin(\Phi_i)}{P_e} \tag{3.32}$$

$$P_e = \sum_{i=1}^{N} g_i^2 \tag{3.33}$$

where e_x and e_y are the vector components in the x and y directions respectively, N is the total number of loudspeakers in the array and g_i is the real gain of the i^{th} loudspeaker located at the horizontal angle Φ_i. The physical meaning of P_e can be considered as a total energy of the system. The magnitude or norm of the energy vector, defined as

$$||\mathbf{e}|| = \sqrt{e_x^2 + e_y^2} \tag{3.34}$$

can be thought of as the measure of energy concentration in a particular direction. The direction of the maximum energy concentration is given by:

$$\phi_e = \arctan\left(\frac{e_y}{e_x}\right) = 2\arctan\left(\frac{||\mathbf{e}|| - e_x}{e_y}\right) \tag{3.35}$$

Similarly, velocity vectors are defined as:

$$\mathbf{v} = [v_x \quad v_y] \tag{3.36}$$

where

$$v_x = \sum_{i=1}^{N} \frac{g_i \cos(\Phi_i)}{P_v} \tag{3.37}$$

$$v_y = \sum_{i=1}^{N} \frac{g_i \sin(\Phi_i)}{P_v} \tag{3.38}$$

$$P_v = \sum_{i=1}^{N} g_i \tag{3.39}$$

The magnitude or norm of the velocity vector, defined as

$$||\mathbf{v}|| = \sqrt{v_x^2 + v_y^2} \tag{3.40}$$

can be thought of as a ratio of the net acoustic velocity from the N loudspeakers that simulate a sound source in the direction ϕ_S and the velocity that would have resulted from the single sound source in this direction [42]. It is important to remember that while the sign of the gains squared in the energy vectors is always positive that in the the velocity vectors the sign is preserved and can be negative as well. The practical implications of this fact are that the norm of the velocity vector can be adjusted by using out-of-phase loudspeakers "pulling" the pressure from the diametrically opposite direction. For physical sources, the magnitude of the velocity vector is always 1 but for a virtual source, because of the possible out-of-phase components, the magnitude of the velocity vector can be greater than 1.

The velocity vector direction defined as

$$\phi_v = \arctan\left(\frac{v_y}{v_x}\right) = 2\arctan\left(\frac{||\mathbf{v}|| - v_x}{v_y}\right) \tag{3.41}$$

simply indicates the net direction of air particle oscillations.

Energy and velocity vectors has proven to be useful in predicting the high and low frequency localisation in multi-loudspeaker systems and have been used extensively as a tool in designing decoders. Vector directions are good predictors of perceived angles of low and mid-high frequency sources and the length of each vector is a good predictor of the "quality" or "goodness" of localisation. Based on this theory, Gerzon [77] defined some useful guidelines that can be utilised not only for the optimal Ambisonic decoder design but also for the general design of sound source panners utilising multichannel loudspeaker configurations, i.e. (after [42]):

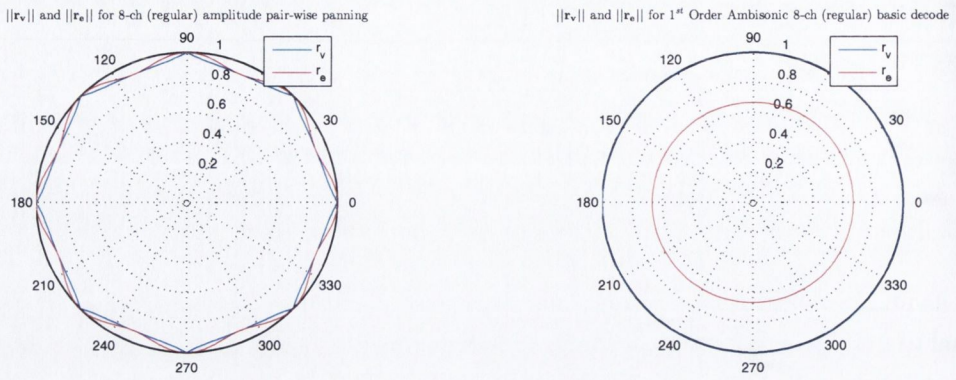
1. Reproduced energy P_e should be substantially independent of the panning angle

2. The velocity and energy vector directions ϕ_v and ϕ_e should be closely matched
3. The velocity and energy vector directions ϕ_v and ϕ_e should be reasonably close to the panning angle ϕ_S
4. Velocity vector length $||\mathbf{v}||$ should be close to unity
5. Energy vector length $||\mathbf{e}||$ should be as large as possible

As an example, for amplitude-panned sources, magnitudes of energy and velocity vectors are equal to unity whenever virtual sound sources are at the locations of the reproducing loudspeakers and drop at in-between loudspeaker locations. On the other hand, for two loudspeakers at angles $\pm\Phi$ creating a phantom image at the point of their symmetry the magnitudes of energy and velocity vectors are at their minimum. It can be shown that at in-between loudspeaker locations the calculation of the magnitudes of energy and velocity vectors simplifies to $||\mathbf{v}|| = ||\mathbf{e}|| = cos(\Phi)$. Nevertheless, for a given multichannel configuration, pair-wise amplitude panning already optimises the energy vectors and by using more than two loudspeakers to reproduce a single phantom source one can only deteriorate the energy vector norm [42]. In favour of the amplitude panning is also the fact that it is easy to preserve the energy with the panned virtual sources and thus, the perceived loudness (e.g. Pairwise Constant Power Panning).

On the other hand, in Ambisonics, usually more than two loudspeakers contribute to sound field reconstruction at all times. That is why, the magnitudes of energy vectors are lower than in the case of the pairwise panning. However, for the regular decoders using diametrically opposed pairs of loudspeakers it is true that their lengths can be made independent of the panning angle ϕ_S so the same quality of localisation can be generally assured in all panning directions. The comparison of magnitudes of energy and velocity vectors for Pairwise Constant Power Panning and Ambisonics using identical octagonal array of loudspeakers is presented in Figure 3.16.

One of the methods for optimising the performance of Ambisonic renderings is to realise that different decoder types can be used more effectively if they operate in their optimal frequency bands. The approach is to optimise the velocity vectors at the low frequency range and the energy vectors at the mid and high frequency range. Decoders that utilise different decoding schemes at particular frequency ranges are known as shelf-filter decoders or dual-band decoders [117]. In such decoders, usually a linear-phase shelf-filters are used with a cross-over point set in the range of $400Hz$ - $700Hz$, as illustrated in Figure 3.17 [105]. Up to the cross-over point the omnidirectional and directional B-Format components are decoded using the velocity scheme. Above the cross-over point however, the ratio between the omnidirectional component and all the directional components is changed by $+3dB$ allowing for the energy decode to apply. In the Figure 3.17 the red curve ($G_W(f)$) denotes the magnitude response of the shelf-filter acting on the W component of the B-Format stream and the blue curve ($G_X(f)$) denotes the magnitude response of the shelf-filters acting on the directional sound field components (i.e. X and Y).



(a) (b)

Figure 3.16: Comparison of magnitudes of velocity and energy vectors $||\mathbf{v}||$ and $||\mathbf{e}||$ for PCPP (a) and Ambisonics (b) using identical, octagonal loudspeaker array.

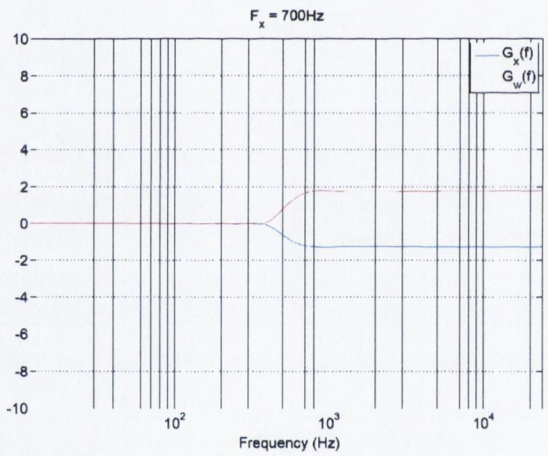


Figure 3.17: Shelf-filters used for dual-band Ambisonic decoding, from [105]

Another very important application of the Gerzon's localisation criteria can be found in the field of decoder design for standard and non-standard, non-uniform loudspeaker layouts (like *ITU 5.1*). The optimisation process in general looks into ways of finding the best possible decoder weighting coefficients that can be applied the B-Format signals for a given loudspeaker configuration so that the energy and velocity vectors are optimised for all panning angles. The problem is well documented in the literature and has been investigated by Gerzon [75, 79], Wiggins [234] or Benjamin, Heller and Lee e.g. [20] amongst others.

Nevertheless, even under optimal decoding conditions, the localisation in Ambisonics generally suffers from significant blur. It is of little recompense that this blur is (for regular loudspeaker configurations) the same at loudspeaker and in-between loudspeaker locations. Better

spatial localisation of sound sources can be achieved in Ambisonics only if the virtual microphones pointing at the loudspeaker locations have sharper directivity characteristics. Gerzon in [72] used the example that for a perfect sound field reconstruction an infinite number of virtual microphones, all with Kronecker-delta directivity function would be required to point at all possible directions on a unity sphere. Of course, this purely theoretical concept cannot be physically realised. However, the idea of forming more directive virtual microphones is perfectly valid and leads to the notion of Higher Order Ambisonics (HOA). The virtual microphone beam forming can be mathematically realised rather easily using methods of approximation directivity functions by truncated series expansion in the spatial domain. The functions that happen to be the most useful in the process are called *spherical harmonics*. To understand how it works we will next look at the process of approximating one- and multi-dimensional functions using finite series expansions which will subsequently lead us to the spherical harmonics decomposition of functions describing real sound fields.

3.4 Higher Order Ambisonics

We have already shown by the means of Kirchhoff-Helmholtz integral (Equation 3.4) that at any listening point within a source-free volume, the pressure field can be calculated if both the sound pressure and its gradient are known on the surface enclosing this volume. This important theorem is rudimentary in WFS that was examined earlier in this work.

Here, we investigate an alternative holophonic approach that considers two- or three-dimensional sound fields as being the superposition of plane waves. To set a basis for this work, we first, we look at some basic theory about approximation of one- and two-dimensional functions using finite series expansion. Then we introduce the spherical harmonics decomposition of a sound field and, ultimately, Higher Order Ambisonics.

To start with, it is good to appreciate that the truly holophonic approach would require that plane wave sources are contributing to the sound field from all possible directions. Their relative contribution could be expressed with some continuous function $p(\phi, \theta)$ which is defined on a surface of the unity sphere and where ϕ and θ horizontal and vertical angles respectively.

Now, let us assume a source-free volume inside an imaginary sphere of a radius 1. As a result of a sound source acting outside the volume, at particular time t we will observe an acoustic pressure distributed over the surface of the sphere. Mathematically, we can also write this pressure function in terms of spherical co-ordinates ϕ and θ as $p(\phi, \theta, t)$.

If we take a snapshot of this function at a particular point in time t_1 i.e. $p_{t_1}(\phi, \theta)$ then it turns out that, this function can be expressed in terms of an infinite series of some weighted basis functions which we can write as:

$$\tilde{p}_{t_1}(\phi, \theta) = \sum_{k=1}^{\infty} c_k b_k(\phi, \theta) \tag{3.42}$$

where $\tilde{p}_{t_1}(\phi, \theta)$ is the reconstructed function, c_k are weighting coefficients and $b_k(\phi, \theta)$ describes the set of some basis functions used for the reconstruction. However, a sum of infinite number of terms is a computational impossibility. That is why the process of a perfect reconstruction is in practice replaced with the process of approximation using a finite number of terms. To understand this better, it is useful to look at the approximation of arbitrary 1-D functions first.

To start with, it is necessary to decide on what set of basis functions can be used. In general, functions that are linearly independent, orthogonal and also orthonormal should be considered as a good choice for this task. The first condition simply means that it should not be possible to obtain any of the basis function from the combination of other basis functions. Orthogonality property means that if we integrate a product of any two of the functions we will end up with either zero if they are the same or a constant if they are different:

$$\int_a^b f_m(x) f_n(x) dx = \begin{cases} 0 & \text{for } n \neq m \\ c & \text{for } n = m \end{cases} \tag{3.43}$$

where c is some constant. A similar term is also used in vector analysis when we talk about the dot product of two vectors. Additionally, Orthonormality is simply assumed when c always equals to 1.

Here, we use the basis functions known as Legendre polynomials, more detailed definition of which can be found in Appendix A. First six Legendre polynomials of up to and including order 5 are presented in Figure 3.18.

For example, Figure 3.19 shows the approximation of a function $f(x) = e^{-x} sin(10x)$ using 1,2,3,4 and 5 Legendre basis functions respectively and the mathematical derivations of these results can be found in Appendix A.

Although a perfect reconstruction of a function $f(x)$ would often necessitate an infinite number of terms to be summed, the acceptable approximation can yet be achieved with only few first terms. We can see that the fourth order approximation is already very close to the original function. Of note, that the low-order approximations are band-limited, i.e. it is impossible to reconstruct the fine details of the function $f(x)$.

However, because the ultimate goal is to be able to approximate multi-dimensional functions (e.g. 2-D pressure distribution defined on the surface of a sphere), we need to extend our considerations to include the vertical component as well. For this reason, an extended set of basis functions needs to be employed. Therefore, let us introduce spherical harmonics basis functions which form a set of orthogonal basis functions and which are defined on a surface of the unitary sphere.

Spherical harmonics combine together two sets of orthogonal basis functions: associated Leg-

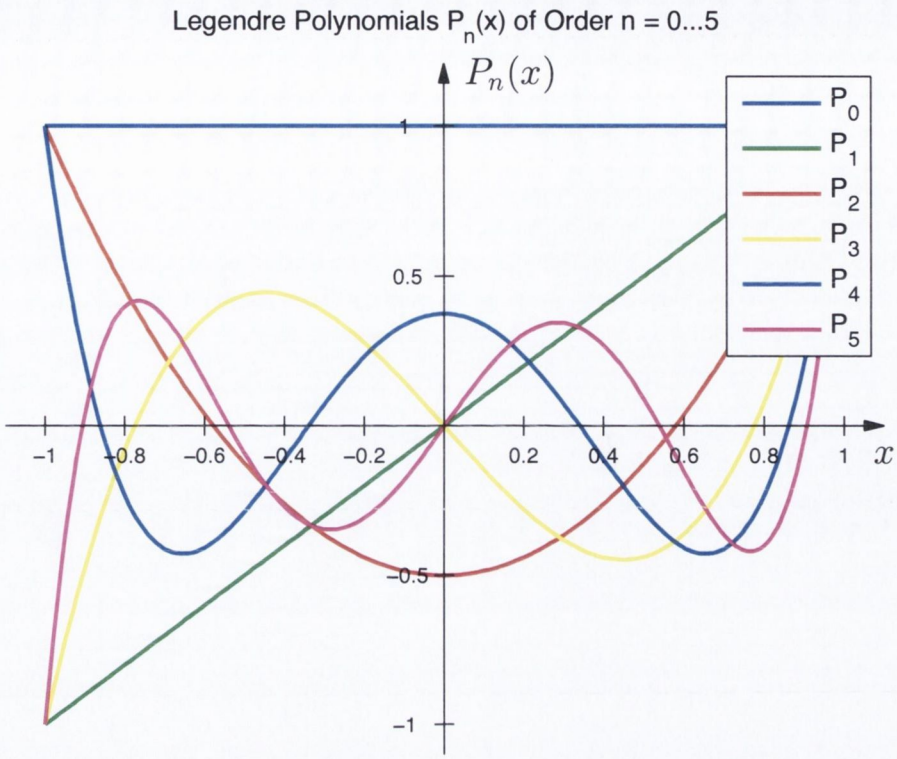


Figure 3.18: Legendre Polynomials $P_n(x)$ of orders $n = 0...5$

endre polynomials and trigonometric functions (sines and cosines). For reference, the associated Legendre polynomials are defined in the Appendix B. First six Associated Legendre Polynomials of up to and including order 2 are presented in Figure 3.20.

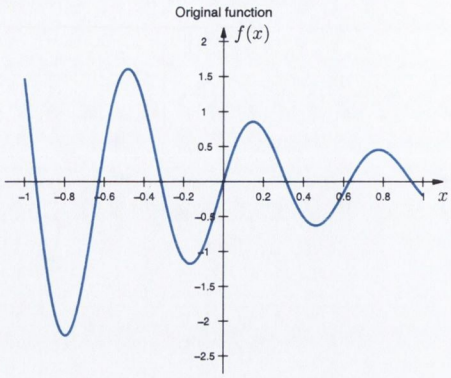
It is convenient to think of spherical harmonics in terms of spherical coordinates using longitudinal and attitudinal angles ϕ and θ respectively. In this thesis, the following convention is used in order to define the domain of spherical harmonic functions:

$$Y_n^m(\phi,\theta), \text{where } 0 \leq \phi < 2\pi, -\tfrac{\pi}{2} \leq \theta < \tfrac{\pi}{2}, n \in R^+, -n \leq m \leq n \tag{3.44}$$

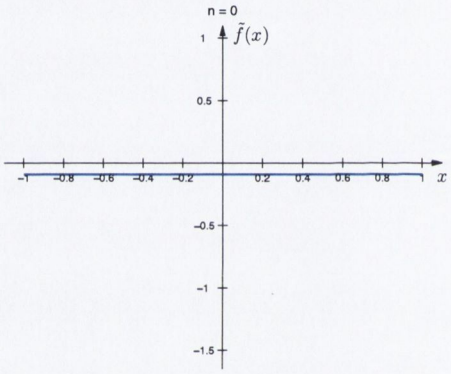
Thus, the spherical harmonic function of order n and degree m can be constructed using the following equation:

$$Y_n^m(\phi,\theta) = \begin{cases} A_n^m \cos(m\phi) P_n^m \cos(\theta) & \text{if } m > 0 \\ A_n^m \sin(m\phi) P_n^m \cos(\theta) & \text{if } m < 0 \\ \frac{1}{\sqrt{2}} A_n^0 P_n^0 \sin(\theta) & \text{if } m = 0 \end{cases} \tag{3.45}$$

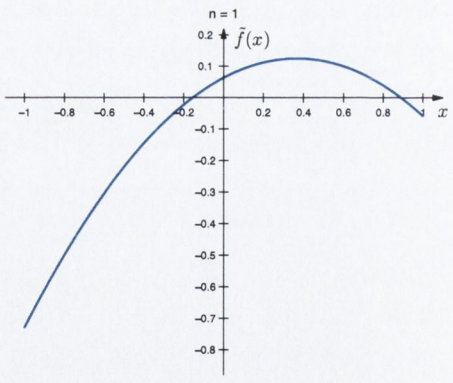
where n is the order and m is the degree of the spherical harmonic and P_n^m are the associated



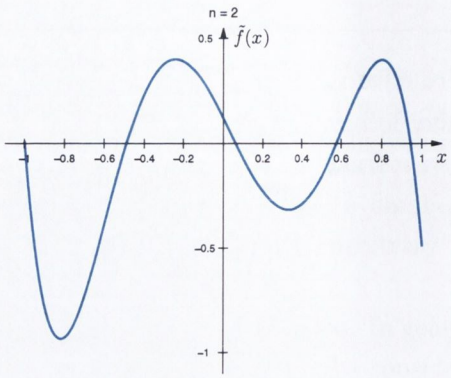
(a) Original function



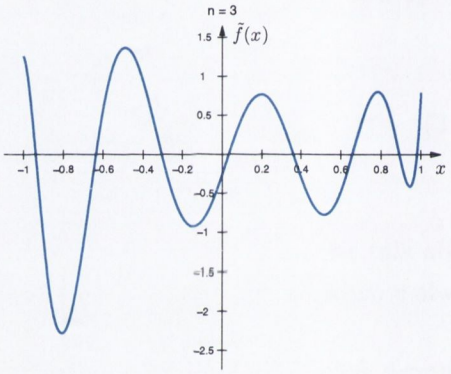
(b) n = 0



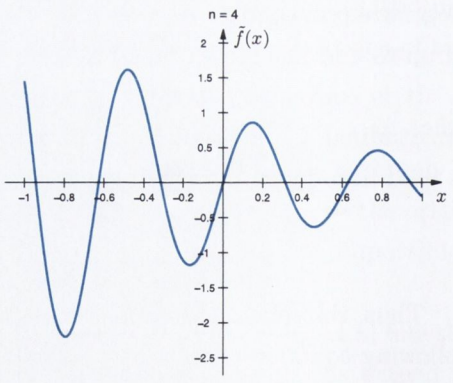
(c) n = 1



(d) n = 2



(e) n = 3



(f) n = 4

Figure 3.19: Legendre polynomial approximation of a function $f(x) = e^{-x}sin(10x)$. The higher the order n, the more polynomials are used and the more accurate the reconstruction is.

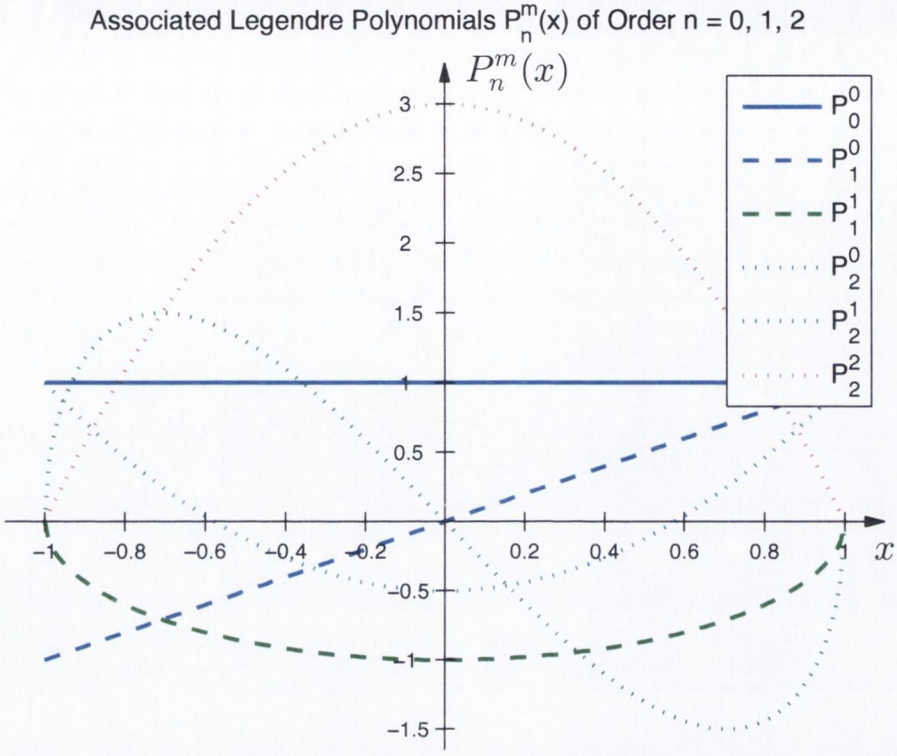


Figure 3.20: Associated Legendre Polynomials $P_n^m(x)$ of orders $n = 0, 1, 2$

Legendre functions. The term A_n^m is used to normalise each function Y_n^m and in this thesis is calculated using the following formula:

$$A_n^m = \sqrt{\frac{2n+1}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}} \tag{3.46}$$

For each order n there are $(2n+1)$ spherical harmonics and the total number of harmonics up to and including specific order n equals $(n+1)^2$. First 25 harmonics of orders from 0 to 4 are presented in Figures 3.21 and 3.22. The visualisation is done using either a colour map on a unit radius sphere (Figure 3.21) or as a displacement in the angular direction (Figure 3.22).

Of note for the future considerations, there are three specific types of these harmonics. *Zonal harmonics* results from $m = 0$ and reduce the Y_n^m function to a scaled associated Legendre polynomial. They divide the unit sphere into zones parallel to the equator. Importantly, rotating these function against the z axis (yaw) has no effect. *Sectoral harmonics* are functions of a form $Y_{|m|}^m$. These are particularly important since they form a set of basis function for horizontal only approximation in circular coordinates. All the remaining functions are referred to as *tesseral*

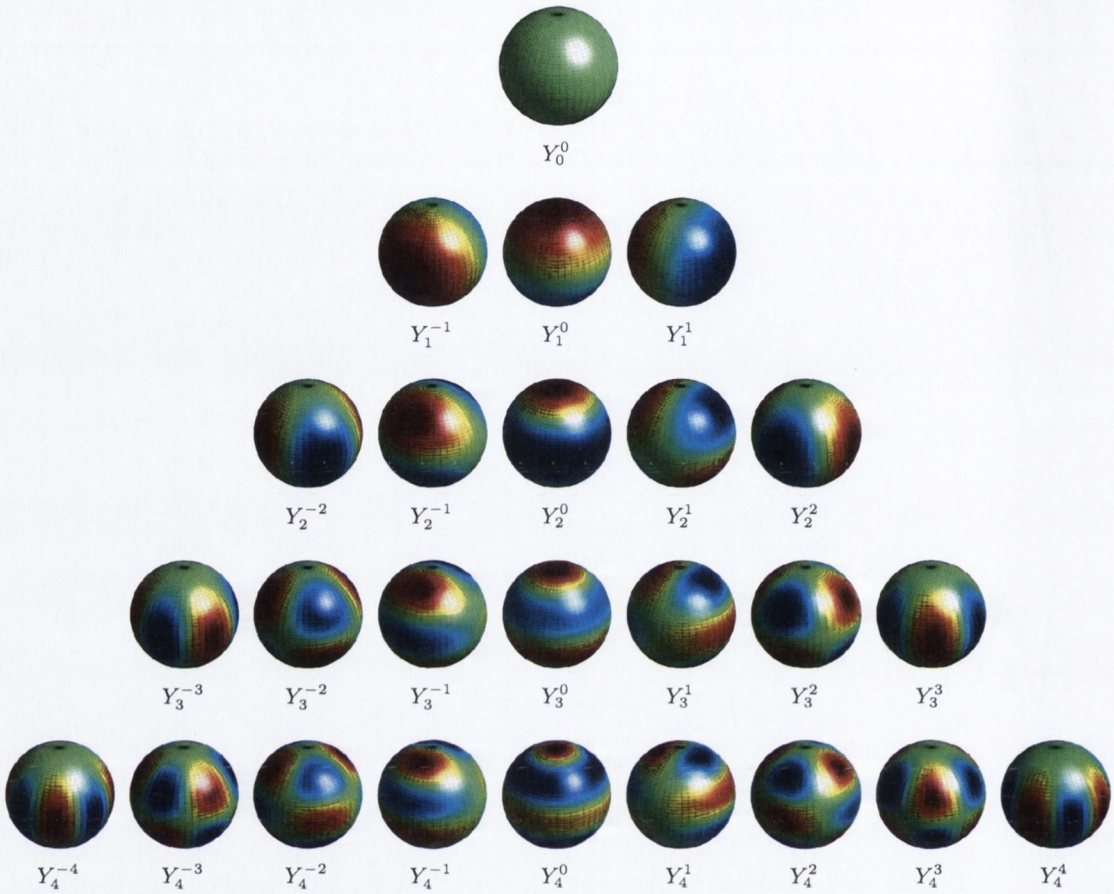


Figure 3.21: Spherical Harmonics Y_n^m of order 0 to 4. Function $Y_n^m(\phi, \theta)$ is visualised as a color map on a unit sphere.

harmonics.

Figure 3.23 shows the approximation process of a function $f(\phi, \theta) = \frac{1}{2}(cos^2(4\theta) - sin^3(3\phi) + 4)^2$ with increasing the order of reconstruction from 0 to 20. The exact mathematial procedure used in order to come up with these results is explained in more detail in the Appendix B. Reconstructed functions are either visualised using the shape deformation ((a) - (l)) or colour maps ((m) - (x)). We can see that the insufficient number of the spherical harmonics used leads to visible artefacts. These artefacts are manifested as a reduction of detail of the original function shape. In other words, low order reconstruction band-limits the function in the spatial domain in an analogous way the low sampling rate band-limits the reconstruction of audio signals and affects its high frequency content.

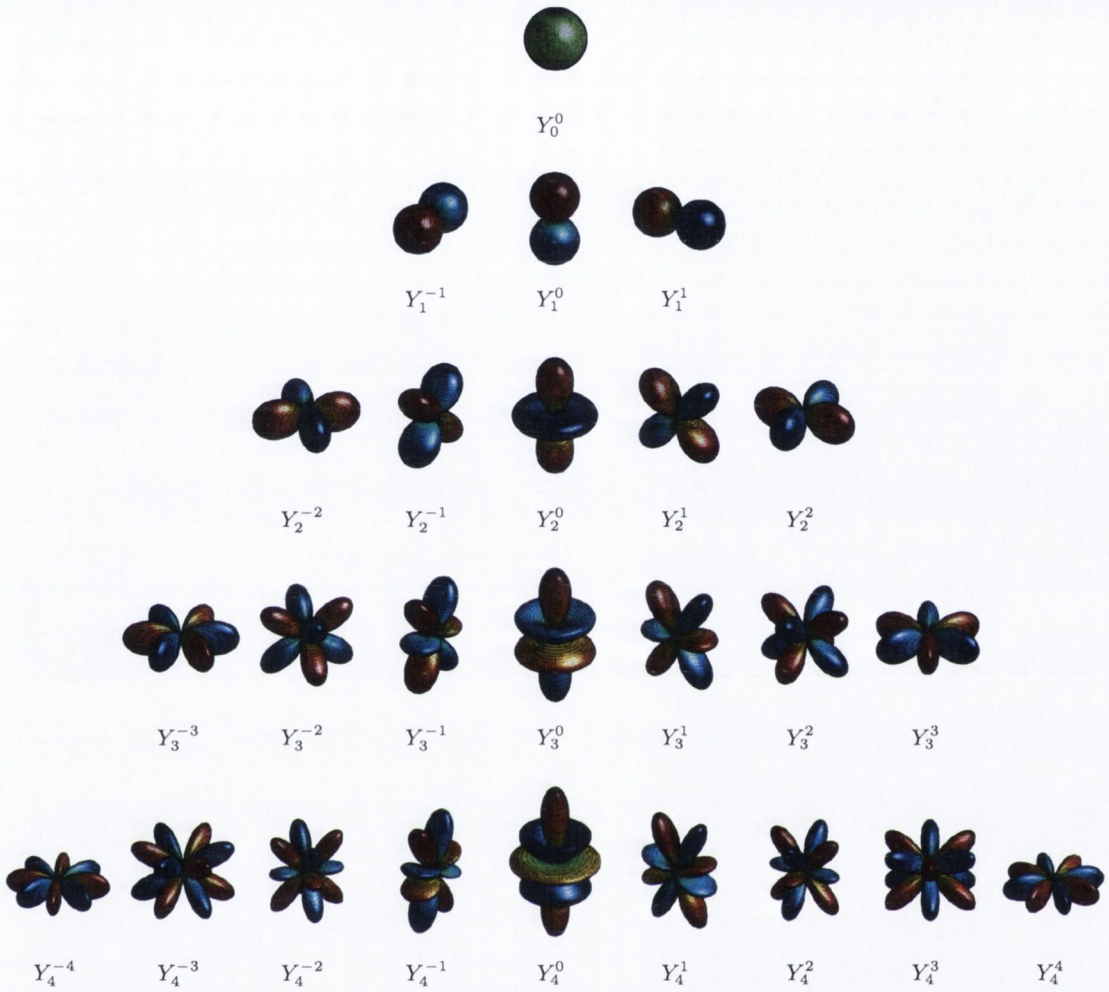
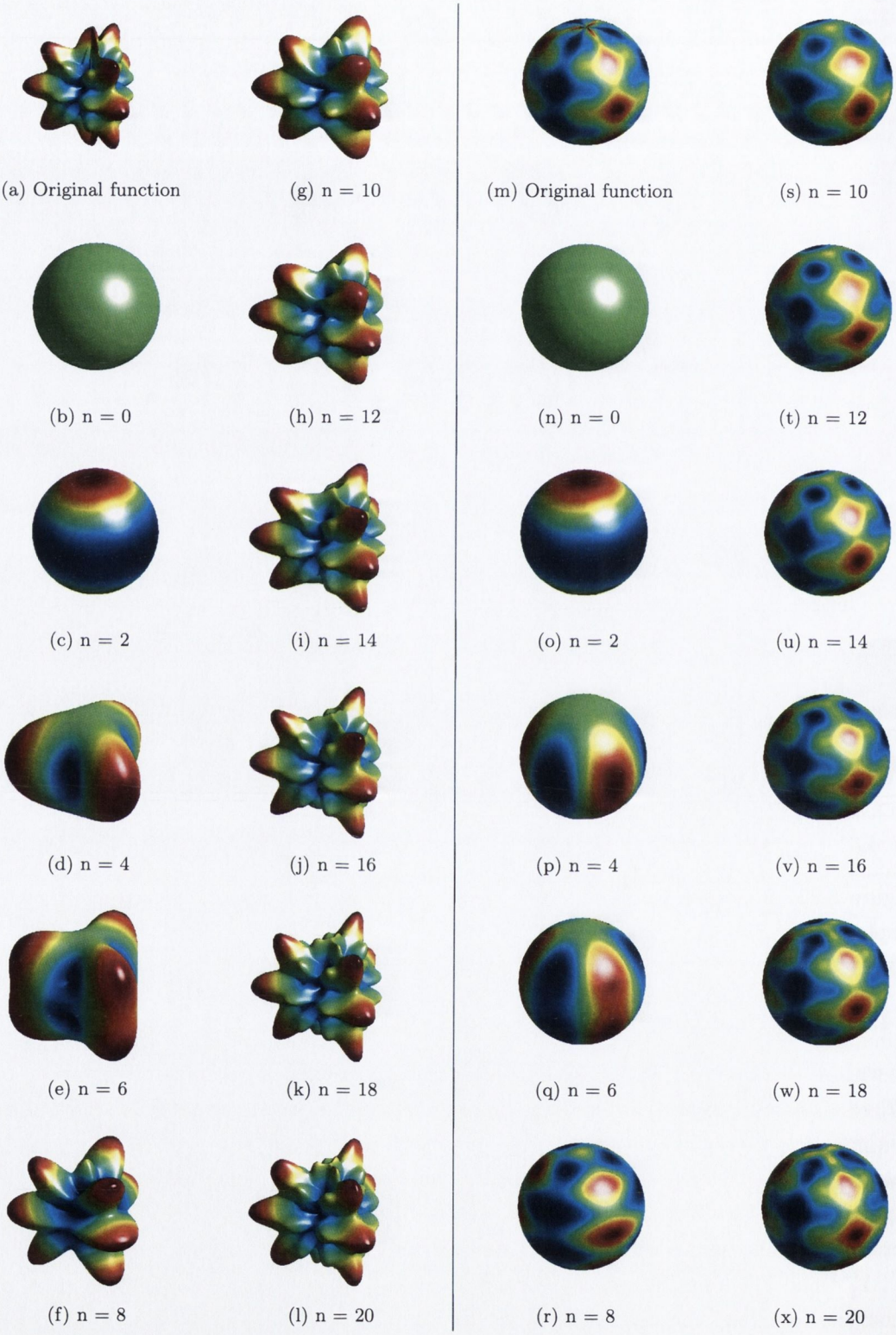


Figure 3.22: Spherical Harmonics Y_n^m of order 0 to 4. Function $Y_n^m(\phi, \theta)$ is visualised as a displacement of sphere surface.

3.4.1 Spherical Harmonics Decomposition of a Sound Field

As we have discussed earlier, Ambisonics was originally developed by Gerzon, Barton and Fellgett [72] as a unified system for recording, reproduction and transmission of surround sound. The mathematical theory behind Ambisonics is in fact based on the decomposition of the the sound field measured at single point in space into spherical harmonic functions as defined earlier in Equation B.6. One has to realise that the first four polar patterns W, X, Y and Z used to obtain B-Format signals w, x, y and z are in fact identical to spherical harmonic functions Y_0^0, Y_1^1, Y_1^{-1} and Y_1^0. Thus, it seems to be a reasonable assumption that it should be possible and beneficial to extend the channel count of Ambisonically encoded sound field by incorporating the higher order spherical harmonic functions Y_n^m.



(a) Original function (g) n = 10 (m) Original function (s) n = 10

(b) n = 0 (h) n = 12 (n) n = 0 (t) n = 12

(c) n = 2 (i) n = 14 (o) n = 2 (u) n = 14

(d) n = 4 (j) n = 16 (p) n = 4 (v) n = 16

(e) n = 6 (k) n = 18 (q) n = 6 (w) n = 18

(f) n = 8 (l) n = 20 (r) n = 8 (x) n = 20

Figure 3.23: Approximation of a function $f(\phi, \theta) = \frac{1}{2}(cos^2(4\theta) - sin^3(3\phi) + 4)^2$

The term *Higher Order Ambisonics* started to appear in the scientific literature in the 1990s in the work of Bamford [16], Daniel [49] or Poletti [178] to name just a few. However, its mathematical foundations were established much earlier because in the 1970s - independently in their papers, Michael Gerzon [72] and Cooper and Shiga [40] first introduced the idea which adapted the Fourier decomposition into the spatial domain of reproduced sound. Essentially, they realised that it is possible to approximate the pressure distribution over a unitary sphere and inside its volume by using a linear of some orthogonal basis functions in the same way as the periodic 1-D function can be decomposed into a series of weighted sine and cosine components.

Cooper and Shiga wrote with respect to the description of a sound source in the azimuthal domain:

> "Since the function $S(\theta)$ is guaranteed to be a periodic function in its dependence upon the azimuthal variable, it may always be represented by a Fourier series either in trigonometric form
>
> $$\begin{aligned} S(\theta) = a_0 &+ a_1 cos\theta + a_2 cos2\theta + \ldots \\ &+ b_1 sin\theta + b_2 sin2\theta + \ldots \end{aligned}$$
>
> or in complex-exponential (phasor) form
>
> $$\begin{aligned} S(\theta) = a_0 &+ c_1 exp(j\theta) + c_2 exp(j2\theta) + \ldots \\ &+ c_1 exp(-j\theta) + c_2 exp(-j2\theta) + \ldots \end{aligned}$$
>
> "

On the other hand, in Gerzon's work we can read:

> "(...) Any function on the sphere is expressible, in a unique manner, as the sum of a zeroth harmonic, a first harmonic, a second harmonic, ..., an n^{th} harmonic,..."

In this way, spherical harmonics have been fostered into the field of multi-channel audio reproduction as an effective and efficient tool for arbitrarily accurate reconstruction of 3-D sound fields.

To look at the problem in more detail, the theoretical foundations of Higher Order Ambisonics are centred around the idea that any arbitrarily complex sound field can be approximated as a superposition of plane waves. A plane wave can be thought of as an acoustic wave coming from an infinitely remote sound source so that its wave fronts are parallel to each other and the inverse-square law no longer applies. In most practical situations though, a plane wave can be assumed so long as its wavelength is much smaller than the distance R from the emitting source:

$$\lambda << R \tag{3.47}$$

Technically, the approximation process is similar to that of the 1-D functions (using the Legendre polynomials) or 2-D functions (using spherical harmonics expansion) although now the approximated function describes a distribution of acoustic pressure inside some volume. Most commonly in Ambisonics, this function $p(\phi, \theta, r)$ is written in terms of the spherical coordinates ϕ, θ and r compliant with the spherical coordinate system shown in the Figure 3.14. Pressure field resulting from a plane wave with amplitude A characterised with the wavenumber k and incidence angle ϕ_S can be obtained directly from the solution to the wave equation and in the 2-D case, is formulated as:

$$p(kr, \phi) = Ae^{jkr\cos(\phi - \phi_S)} \tag{3.48}$$

This solution can also be expressed in terms of the so-called Fourier-Bessel infinite series expansion [155] for any point $\vec{r}$ in the 3-D space, defined by the spherical coordinates ϕ, θ and r as:

$$p(\vec{r}) = \sum_{n=0}^{\infty} j^n j_n(kr) \sum_{m=-n}^{n} B_{nm}(\phi_S, \theta_S) Y_{nm}(\phi, \theta) \tag{3.49}$$

where $B_{nm}(\phi_S, \theta_S)$ are the coefficients resulting from the multiplication of the amplitude of a plane wave coming from the direction (ϕ_S, θ_S) and the spherical harmonics functions evaluated at these angles:

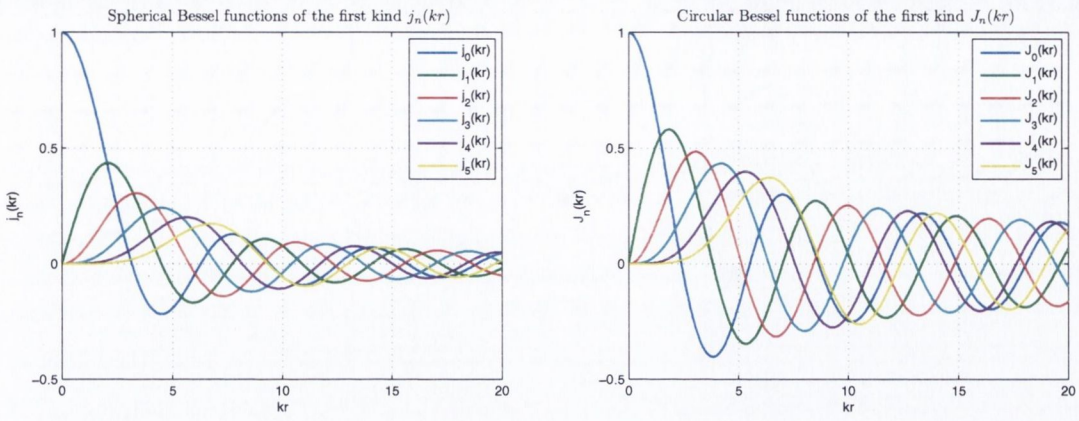
$$B_{nm}(\phi_S, \theta_S) = AY_{nm}(\phi_S, \theta_S) \tag{3.50}$$

In this relation $Y_{nm}(\phi, \theta)$ are the spherical harmonics coefficients calculated using Equation B.6, j_n are the spherical Bessel functions presented in Figure 3.24(a) and $j = \sqrt{-1}$

In 2-D case, the above series expansion must be modified for the use with the circular Bessel functions $J_n(kr)$ (Figure 3.24(b)) as follows:

$$p(\vec{r}) = Aj_0(kr) + 2A\sum_{n=0}^{\infty} j^n J_n(kr)cos(n(\phi - \phi_S)) \tag{3.51}$$

Equation 3.51 can be further expanded to yield



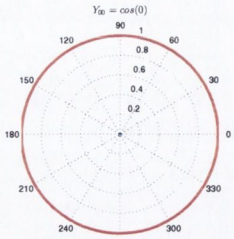
(a) (b)

Figure 3.24: (a) Spherical Bessel functions of the first kind $j_n(kr)$; (b) Circular Bessel functions of the first kind $J_n(kr)$

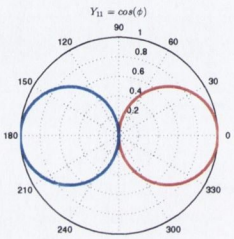
$$
\begin{aligned}
p(\vec{r}) &= AY_0^{0(2D)} j_0(kr) + 2\sum_{n=0}^{\infty} j^n j_n(kr) A cos(n\phi_S) cos(n\phi) \\
&\quad + 2\sum_{n=0}^{\infty} j^n j_n(kr) A sin(n\phi_S) sin(n\phi) \\
&= B_0^0 j_0(kr) + 2\sum_{n=0}^{\infty} j^n j_n(kr) B_n^{n(2D)}(\phi_S) Y_n^{n(2D)}(\phi) \\
&\quad + 2\sum_{n=0}^{\infty} j^n j_n(kr) B_n^{-n(2D)}(\phi_S) Y_n^{-n(2D)}(\phi)
\end{aligned}
\tag{3.52}
$$

in which we can now clearly identify the spherical harmonics coefficients diminished to two dimensions: $Y_n^{n(2D)}$ and $Y_n^{-n(2D)}$. These coefficients result from reducing the subset of the full 3-D spherical harmonic functions $Y_n^{m(3D)}$ and $Y_n^{-m(3D)}$ where $|m| = n$ to two dimensions. First 9 circular harmonics (up to and including order 4) with corresponding generating functions are presented in Figure 3.25.

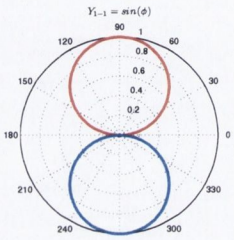
The reconstruction process according to Equation 3.52 is shown in Figure 3.26 and 3.27 using 2-D and 3-D visualisation methods respectively. We can clearly see that the higher the order of series truncation, the better the reconstruction is. Also, assuming the same order of truncation, due to the nature of the Bessel functions, plane waves with lower wavenumbers are reconstructed correctly over the wider area (or volume in the 3-D case). This phenomenon is illustrated in the Figure 3.28 where the 8^{th} order reconstruction is applied to plane waves with frequencies of $100Hz$, $250Hz$, $500Hz$ and $1000Hz$.



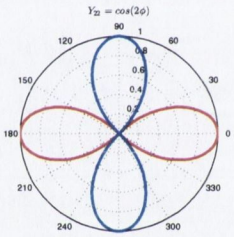
(a) $Y_0^0 = cos(0)$



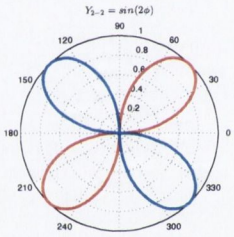
(b) $Y_1^1 = cos(\phi)$



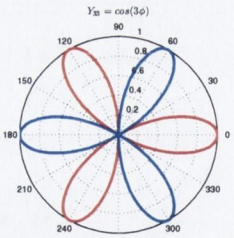
(f) $Y_1^{-1} = sin(\phi)$



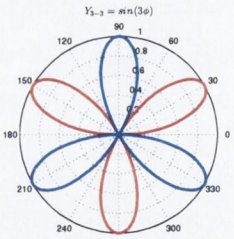
(c) $Y_2^2 = cos(2\phi)$



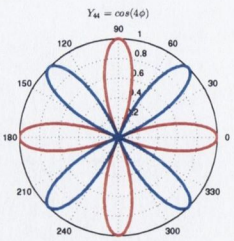
(g) $Y_2^{-2} = sin(2\phi)$



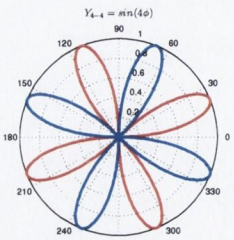
(d) $Y_3^3 = cos(3\phi)$



(h) $Y_3^{-3} = sin(3\phi)$

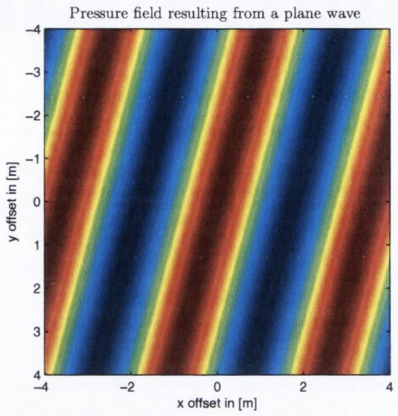


(e) $Y_4^4 = cos(4\phi)$

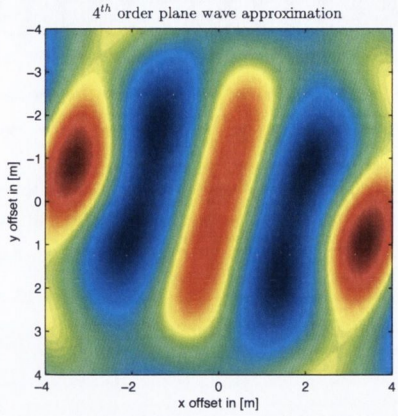


(i) $Y_4^{-4} = sin(4\phi)$

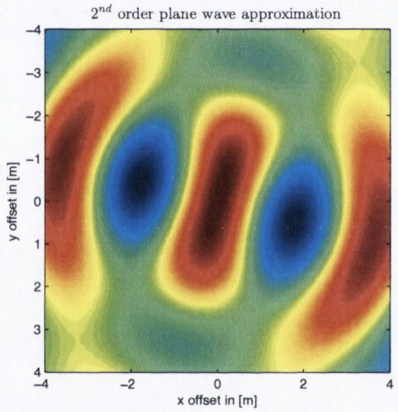
Figure 3.25: Circular harmonic functions Y_{nn}^{2D} and Y_{n-n}^{2D} of order 0 to 4.



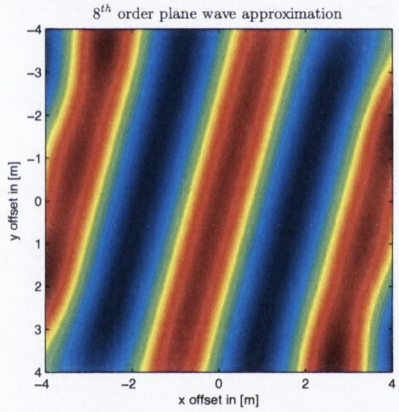
(a) Original plane wave



(c) 4^{th} order approximation



(b) 2^{nd} order approximation



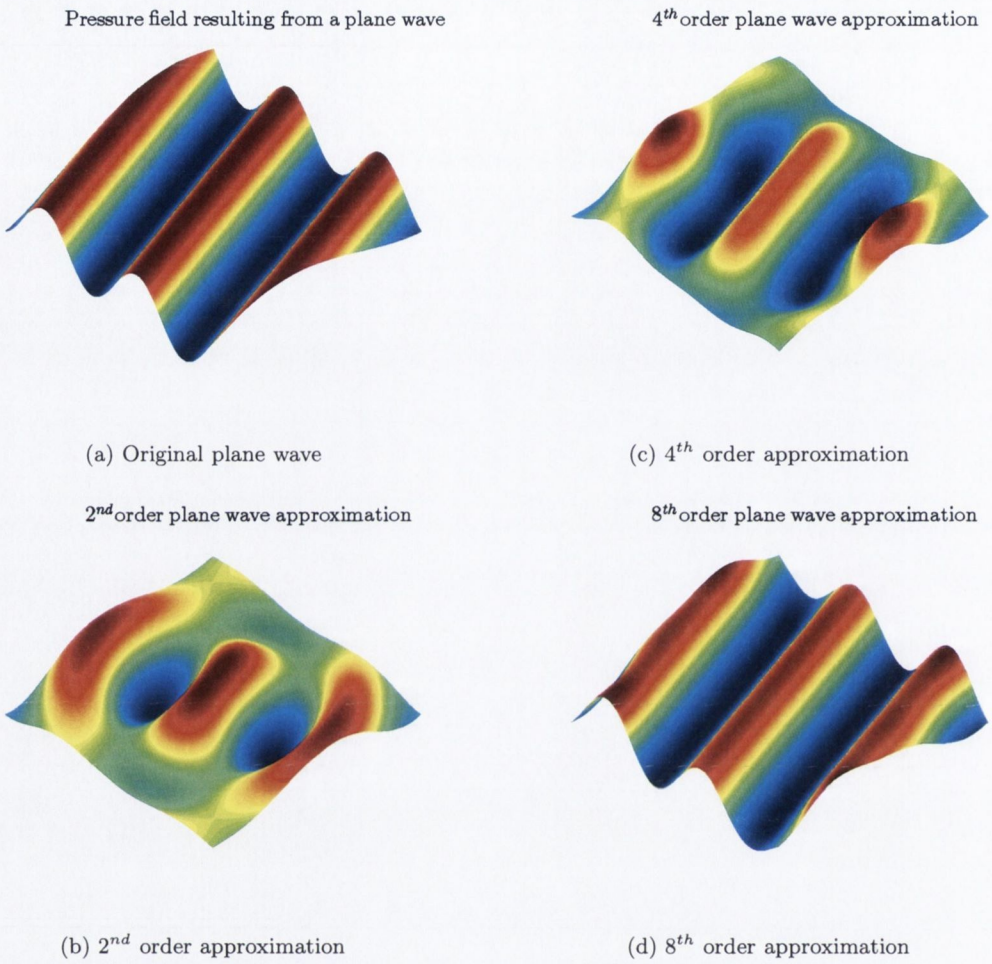
(d) 8^{th} order approximation

Figure 3.26: Approximation of a pressure field resulting from a plane wave with frequency $f = 100Hz$ and the incidence direction $\phi_S = -15°$, visualised as a colour map on a 2-D surface

3.4.2 HOA Sound Field Rotation

We have already shown in Section 3.3.2 how the rotation matrices $\mathbf{R_z}$, $\mathbf{R_y}$ and $\mathbf{R_x}$ can be used in order to rotate 1^{st} order Ambisonic sound fields around the z-, y- and x-axis respectively. Now, we will extend our discussion to higher orders of reproduction.

In general, a transformation matrix applying rotation to zeroth, first and higher order components is *block-diagonal*, i.e. it consists of sub-matrices $\mathbf{R^0}$, $\mathbf{R^1}$, $\mathbf{R^2}$,..., $\mathbf{R^n}$ along its diagonal acting on spherical harmonic components of orders 0, 1, 2,...,n respectively. It means that lower order components of the matrix $\mathbf{R}$ have no effect on the computation of rotated higher order spherical harmonic components. A general form of the matrix $\mathbf{R}$ is presented in Equation 3.53 (after [84]):



(a) Original plane wave

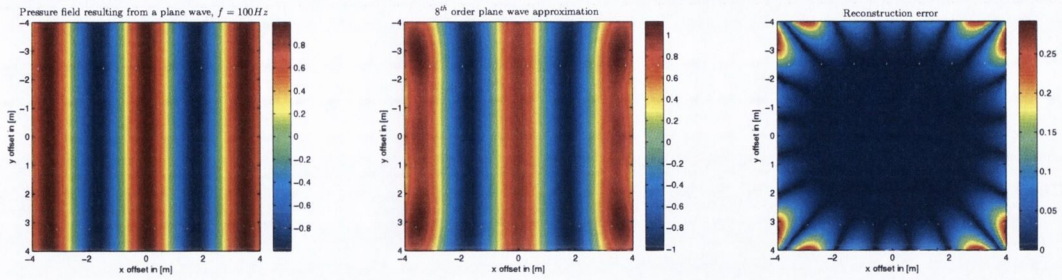
(c) 4^{th} order approximation

(b) 2^{nd} order approximation

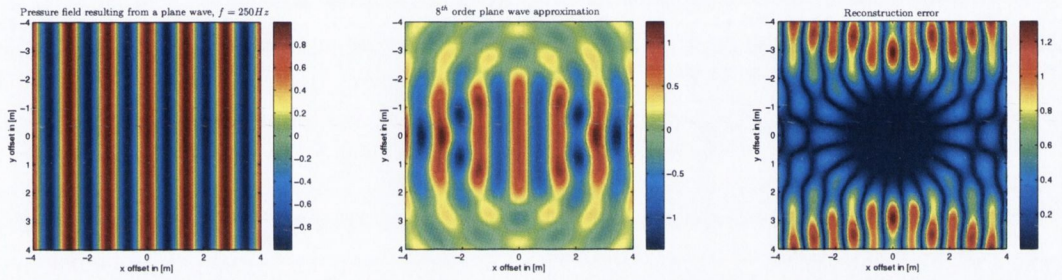
(d) 8^{th} order approximation

Figure 3.27: Approximation of a pressure field resulting from a plane wave with frequency $f = 100Hz$ and the incidence direction $\phi_S = -15°$, visualised as a displacement on a 2-D surface

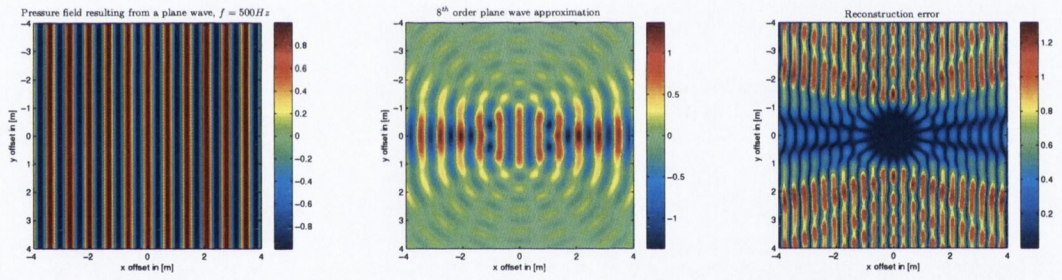
$$\mathbf{R} = \left[\begin{array}{c|ccc|ccccc|c} X & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \hline 0 & X & X & X & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & X & X & X & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & X & X & X & 0 & 0 & 0 & 0 & 0 & \dots \\ \hline 0 & 0 & 0 & 0 & X & X & X & X & X & \dots \\ 0 & 0 & 0 & 0 & X & X & X & X & X & \dots \\ 0 & 0 & 0 & 0 & X & X & X & X & X & \dots \\ 0 & 0 & 0 & 0 & X & X & X & X & X & \dots \\ 0 & 0 & 0 & 0 & X & X & X & X & X & \dots \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}\right] = \begin{bmatrix} 1 & 0 & 0 & \dots \\ 0 & R^1 & 0 & \dots \\ 0 & 0 & R^2 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{3.53}$$



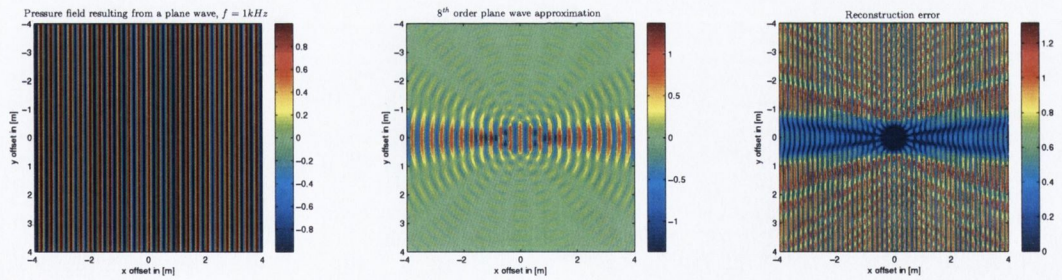
(a) Original $100Hz$ plane wave (e) 8^{th} order approximation (i) Reconstruction error



(b) Original $250Hz$ plane wave (f) 8^{th} order approximation (j) Reconstruction error



(c) Original $500Hz$ plane wave (g) 8^{th} order approximation (k) Reconstruction error



(d) Original $1kHz$ plane wave (h) 8^{th} order approximation (l) Reconstruction error

Figure 3.28: Spatial reconstruction error of 8^{th} order pressure field approximations for plane waves at: (a) $f = 100Hz$; (b) $f = 250Hz$; (c) $f = 500Hz$; (d) $f = 1kHz$ and the incidence direction $\phi_S = 0°$. As the frequency gets higher the area of correct reconstruction gets smaller.

Many different algorithms have been proposed in literature to obtain higher order terms for rotation matrices $\mathbf{R_z}(\alpha)$, $\mathbf{R_y}(\beta)$, $\mathbf{R_x}(\gamma)$ [25, 39, 102, 103, 113, 177] e.g. by direct recursion. The easiest to derive are higher order sub-matrices of the matrix $\mathbf{R_z}(\alpha)$ (yaw). It can be shown [245], that the higher order terms of a sub-matrix $\mathbf{R_z^n}(\alpha)$ are fairly straightforward to derive since the dependency on the vertical component (elevation) is lost. Thus, in general rotation of horizontal-only parts can be broken down into the following matrix multiplication [245]:

$$\begin{bmatrix} \cos n(\phi+\alpha) \\ \sin n(\phi+\alpha) \end{bmatrix} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}^n \begin{bmatrix} \cos m\phi \\ \sin m\phi \end{bmatrix} \tag{3.54}$$

where ϕ is the original angle of some encoded sound source. Therefore, the rotation matrix $\mathbf{R_z}(\alpha)$ will take on a general form of:

$$\mathbf{R_z}(\alpha) = \left[\begin{array}{c|ccc|cc|cc|c} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \hline 0 & \cos\alpha & -\sin\alpha & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & \sin\alpha & \cos\alpha & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \dots \\ \hline 0 & 0 & 0 & 0 & \cos 2\alpha & -\sin 2\alpha & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & \sin 2\alpha & \cos 2\alpha & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cos\alpha & -\sin\alpha & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & \sin\alpha & \cos\alpha & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}\right] \tag{3.55}$$

Obtaining matrices $\mathbf{R_y}(\beta)$ and $\mathbf{R_x}(\gamma)$ is less trivial and quickly becomes mathematically involved when higher order spherical harmonics are employed. However, the problem can be simplified as soon as we realise that any arbitrary 3-D rotation $\mathbf{R}(\alpha, \beta, \gamma)$ can be expressed in terms of a matrix multiplication of the three component matrices $\mathbf{R_z}(\alpha)\mathbf{R_y}(\beta)\mathbf{R_x}(\gamma)$. In fact, any arbitrary 3-axis rotation can be also expressed as a combination of 2-axis rotations. For example, Zotter [245] proposes that the full 3-axis rotation is performed using only $\mathbf{R_z}(\alpha)$ and $\mathbf{R_y}(\mathbf{90°})$, i.e.

$$\mathbf{R}(\alpha, \beta, \gamma) = \mathbf{R_z}(\alpha + \mathbf{45°})\mathbf{R_y}(\mathbf{90°})\mathbf{R_z}(\beta + \mathbf{180°})\mathbf{R_y}(\mathbf{90°})\mathbf{R_z}(\gamma + \mathbf{45°}) \tag{3.56}$$

Pre-computed rotation matrices $\mathbf{R_y(90°)}$ of orders 1 to 21 can be found on-line in [245].

3.4.3 Decoding Higher Order Ambisonics

Similarly as in the case of the First Order Ambisonics (Section 3.3.3) in order for a plane wave to be reconstructed by a loudspeaker array we must ensure that

$$\mathbf{b} = \mathbf{L} \cdot \mathbf{g} \tag{3.57}$$

where

$$\mathbf{b} = s\mathbf{Y}_{\phi_\mathbf{S}\Theta_\mathbf{S}} = s[Y_0^0(\phi_S\theta_S), Y_1^{-1}(\phi_S\theta_S), \ldots, Y_n^m(\phi_S\theta_S)]^T \tag{3.58}$$

is a column vector with the HOA B-format signals according to the ACN convention [9]:

$$\mathbf{b} = s[W, Y, Z, X, V, T, R, S, U, Q, O, M, K, L, N, P, \ldots]^T \tag{3.59}$$

and s is the current pressure sample of the source signal from direction (ϕ_S, θ_S). $\mathbf{L}$ is the loudspeaker *re-encoding* matrix that encodes each loudspeaker direction into the spherical harmonics domain:

$$\mathbf{L} = \begin{bmatrix} Y_0^0(\Phi_1,\Theta_1) & Y_0^0(\Phi_2,\Theta_2) & \ldots & Y_0^0(\Phi_i,\Theta_i) & \ldots & Y_0^0(\Phi_N,\Theta_N) \\ Y_1^{-1}(\Phi_1,\Theta_1) & Y_1^{-1}(\Phi_2,\Theta_2) & \ldots & Y_1^{-1}(\Phi_i,\Theta_i) & \ldots & Y_1^{-1}(\Phi_N,\Theta_N) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Y_n^m(\Phi_1,\Theta_1) & Y_n^m(\Phi_2,\Theta_2) & \ldots & Y_n^m(\Phi_i,\Theta_i) & \ldots & Y_n^m(\Phi_N,\Theta_N) \end{bmatrix} \tag{3.60}$$

and $\mathbf{g}$ is the column vector with loudspeaker gains:

$$\mathbf{g} = [g_1, g_2, \ldots, g_i, \ldots, g_N]^T \tag{3.61}$$

where g_i is the i^{th} loudspeaker gain from direction (Φ_i, Θ_i). The loudspeaker gain vector $\mathbf{g}$ can be obtained again by finding the decoding matrix $\mathbf{D}$ which is the inverse of the re-encoding matrix $\mathbf{L}$:

$$\mathbf{g} = \mathbf{Db} = \mathbf{L}^{-1}\mathbf{b} \tag{3.62}$$

However, to invert $\mathbf{L}$ we need the matrix to be a square which is only possible if 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 pseudoinverse of $\mathbf{L}$ where:

$$\mathbf{D} = \mathbf{L}^{\dagger} = \mathbf{L}^{\mathbf{T}}(\mathbf{L}\mathbf{L}^{\mathbf{T}})^{-1} \tag{3.63}$$

Again, the process of decoding HOA can be seen as (beam)forming of virtual microphones pointing in the direction of the reproduced sound source. However, the more spherical harmonic components used, the more directive the virtual microphones become. Theoretically, as was pointed out earlier, the directivity of the virtual microphones tend to the Dirac delta function as the order of spherical harmonics decomposition tends to infinity. So, Higher Order Ambisonic systems lead to better spatial definition of sound sources (sources become more point-like as the order goes higher) but this is at the cost of number of loudspeakers required for the reproduction.

In general, for correct reproduction it is recommended that the number of loudspeakers in the array is equal or higher than the number of Ambisonic channels. In the case of 3-D reproduction, the minimum number of loudspeakers can be calculated as:

$$N \geq (n+1)^2 \tag{3.64}$$

where N is the number of loudspeakers and n is the order of Ambisonic reproduction. For pantophonic listening, this equation modifies to:

$$N \geq 2n+1 \tag{3.65}$$

However, in order for the reproduction array to meet the requirements of diametrically opposed pairs, the minimum number of loudspeakers in the 2-D case must equal to:

$$N = 2n+2 \tag{3.66}$$

In the 3-D case, Equation 3.64 still holds whenever the reproduction order is odd. Otherwise, additional loudspeaker is required to make the total number of loudspeakers even:

$$N = (n+1)^2 + 1 \tag{3.67}$$

In order to come up with different decoder types for HOA it is necessary to apply different correction gains k_n to different spherical harmonic orders $1, 2, 3, \ldots, n$. Then the Equation 3.29 becomes:

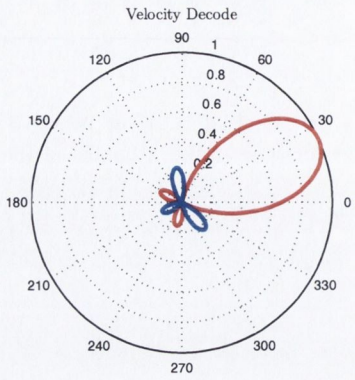
$$l_i = \frac{1}{N}\left[\sqrt{2}k_0 l_{w_i} w + k_1(l_{x_i}x + l_{y_i}y + l_{z_i}z) + \ldots, +k_n(\ldots)\right] \tag{3.68}$$

Daniel [47] proposed general closed-form expressions to calculate k_n's for all popular decoder types. These are summarised in Table 3.2. Polar patterns for the 3^{rd} order pantophonic systems together with the unwrapped loudspeaker gain curves are presented in Figure 3.29.

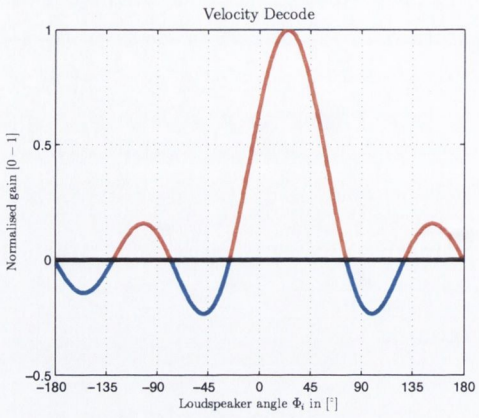
Decoder type	2-D/3-D	k_n
Velocity	2-D	1
	3-D	1
Energy	2-D	$\frac{\cos n\pi}{2n_{max}+2}$
	3-D	$P_n(r_E)$
In-phase	2-D	$\frac{n_{max}!^2}{(n_{max}+n)!(n_{max}-n)!}$
	3-D	$\frac{n_{max}!(n_{max}+1)!}{(n_{max}+n+1)!(n_{max}-n)!}$

(3.69)

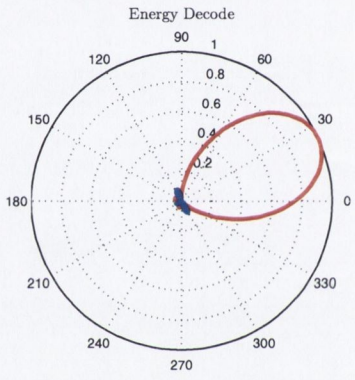
Table 3.2: Correction gains k_n for three different decoder types (velocity, energy, in-phase) in the case of pantophonic and periphonic reproduction. For the energy decode, P_n denotes the n^{th} order Legendre polynomial (see Appendix A) and r_E can be calculated as the largest root of the $P_{n_{max}+1}$ [47].



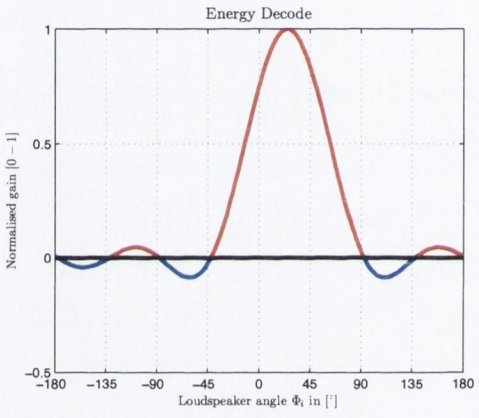
(a)



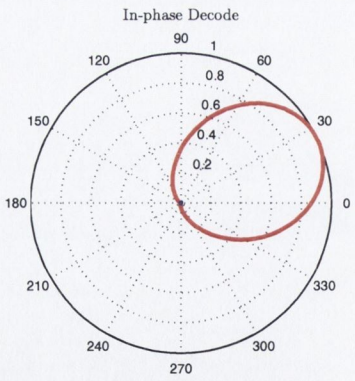
(d)



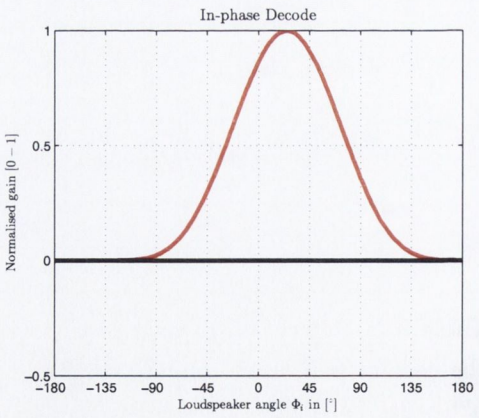
(b)



(e)



(c)



(f)

Figure 3.29: Virtual microphones resulting from the sound source at 25° and three different decoder types (velocity, energy and in-phase). Red colour denotes positive gains whereas blue colour denotes negative gains. Loudspeaker gains can be read from the right hand column curves for a given angle Φ_i.

3.4.4 Near-Field Problem

So far, in our discussion one of the main HOA theoretical assumptions was that the sound field reconstruction is achieved by the means of superposition of plane waves. This fact imposes the requirement for the reconstruction loudspeakers to be at an infinite distance from the centre listening position. Since this requirement is never to be fulfilled it is useful to analyse what errors are to be expected whenever the sound field is rendered by loudspeakers localised at a finite radius.

To begin with, we have already expressed a more relaxed definition of a plane wave which we will repeat here for convenience. An impinging wave can be considered as plane whenever its wavelength is significantly shorter than the distance from the emitting source, i.e.:

$$\lambda << R \tag{3.70}$$

Thus, simple intuitive analysis suggests that larger errors are to be expected in lower frequency bands where the wavelengths are relatively long. On the other hand, we have already shown that with increasing the wavenumber k, the area of correct reconstruction diminishes. Therefore, it seems that for high frequencies, the near-field effect of the loudspeakers is less apparent. However, at the same time they require a higher order of series truncation for the correct wide-area reproduction. On the contrary, for low frequencies, a correct wide-area reproduction is possible with much lower orders of truncation. However, the proximity effect of the loudspeaker array is much more influential. This is illustrated in Figure 3.30 and 3.31.

Reconstruction errors pertaining to the source/loudspeaker array proximity generally manifest themselves as incorrect wave front curvatures and excessive bass-boost (or insufficient bass boost whenever the source is intended to be rendered inside the array) [48, 76]. The first problem can be reduced to the following statement: in the real world scenarios, the ambisonically presented sound source will always appear to be located on the loudspeaker array and not at its original distance, as long as the wave front curvature is concerned. Thus, from the above it can be inferred that only when the array distance matches the sound source distance, the reconstruction inside the listening area is correct. In this case, the reference has to be made to the spherical source instead of the plane wave source which can be modelled using the following equation:

$$p(\vec{r}) = A\frac{\rho}{|\vec{\rho} - \vec{r}|}\frac{e^{-j|\vec{\rho} - \vec{r}|}}{e^{-jk\rho}} \tag{3.71}$$

where $\vec{\rho}$ is the spherical source positional vector, $\vec{r}$ are the points for which the pressure field is to be calculated (with respect to the origin), ρ is the distance from the sound source to the

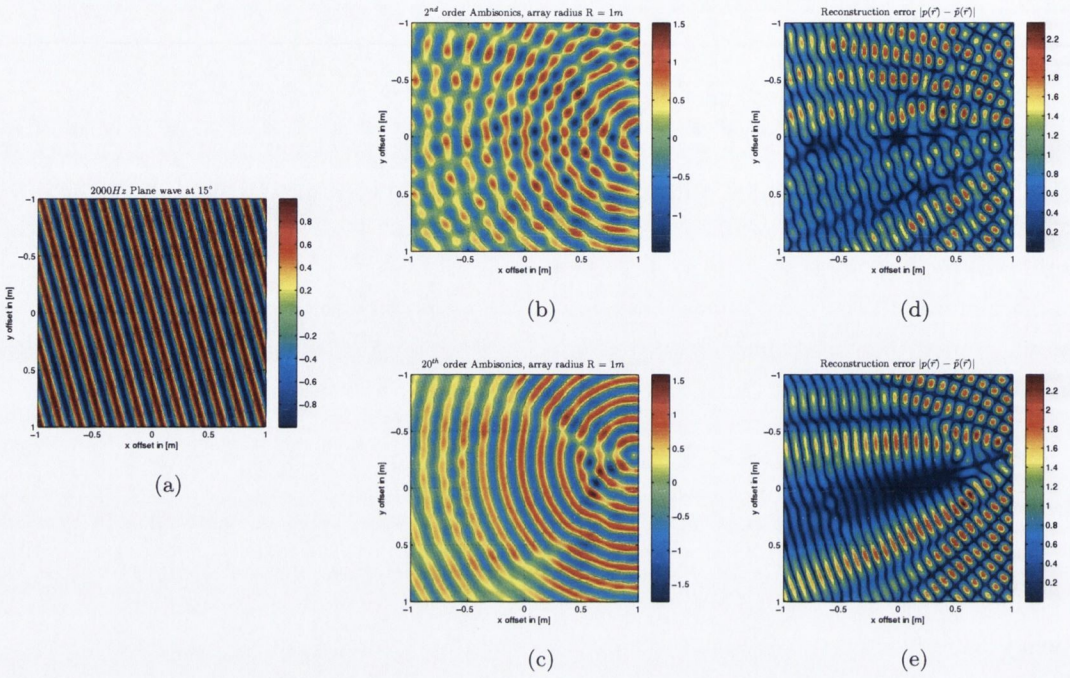


Figure 3.30: Ambisonic horizontal only reconstruction of a $2000Hz$ plane wave coming from the azimuth angle of 15°: (a) original wave field; (b) 2^{nd} order reconstruction; (c) 20^{th} Order reconstruction; (d) Error for he 2^{nd} order reconstruction; (e) Error for the 20^{th} order reconstruction;

origin and A and k are the amplitude the wavenumber respectively. Figure 3.32 illustrates the situation where the spherical sound source is acting beyond the Ambisonic array. We can see that when the array radius approaches the sound source distance, wave curvatures are reconstructed more accurately and the overall reconstruction error in the listening area diminishes.

As long as the wide-area reconstruction is concerned, the wave front curvature distortion may lead to severe off-centre localisation problems. Again, it is because the sound source appears as coming from a point on the loudspeaker array and not from its original far-field location [48] which obviously translates to its angular bias. In effect, angles subtended by each ear could be incorrect. This situation is illustrated in Figure 3.33.

Although this particular problem do not affect a centrally seated listener, the other problem with the excessive/insufficient bass energy certainly might. Figure 3.34 shows the excessive bass-boost issue resulting from a sound source rendered at $1m$ for different HOA components. Moreover, for laterally localised sources, the near-field effect may lead to excessive values of ILD and thus, misinformation as to the sound source distance. It is because for sources simulated closer to the side of the head the inverse square law could dominate the head shadowing effect in the overall ILD, as already explained in Chapter 2 Section 2.3. However, it has been shown by Wittek in [237], this particular distance cue is relatively weak and is usually easily overridden

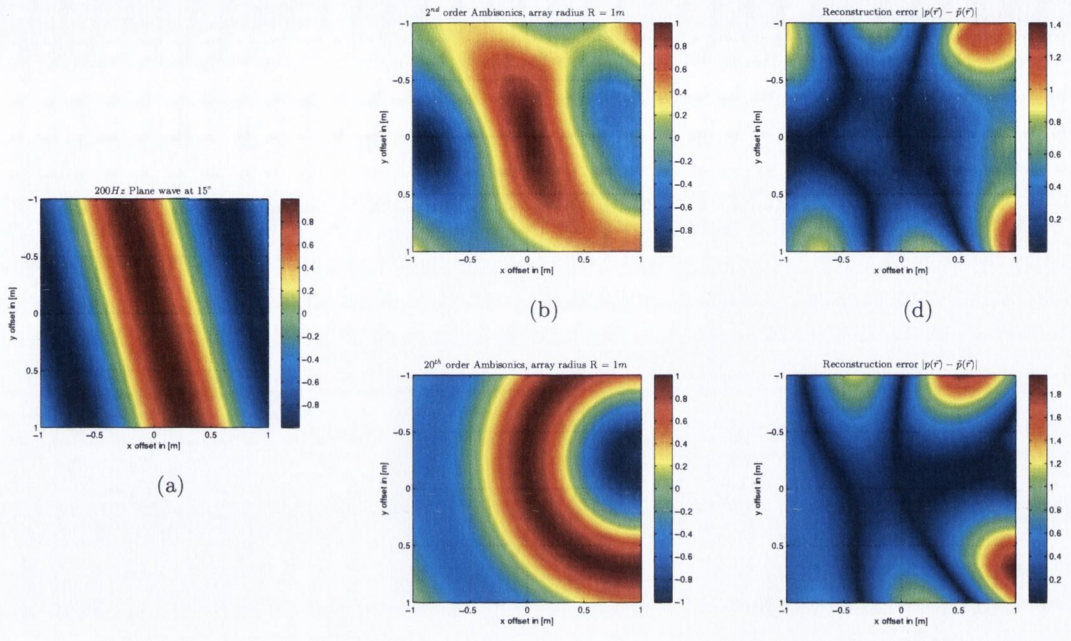


Figure 3.31: Ambisonic horizontal only reconstruction of a $200Hz$ plane wave coming from the azimuth angle of 15°: (a) original wave field; (b) 2^{nd} order reconstruction; (c) 20^{th} Order reconstruction; (d) Error for he 2^{nd} order reconstruction; (e) Error for the 20^{th} order reconstruction;

by stronger distance cues like amplitude or direct-to-reverberant energy ratio.

According to Daniel [48], the near-field effect of loudspeakers on the HOA components can be expressed mathematically in the form of a following transfer function:

$$F_n^R(\omega) = \sum_{n=0}^{n_{max}} \frac{(n_{max}+n)!}{(n_{max}-n)!n!} \left(\frac{-jc}{\omega R}\right), \qquad \omega = 2\pi f \tag{3.72}$$

where n is the current order of the HOA component and n_{max} is the general reconstruction order. Then, during the decoding stage, correction gains need to be applied to the decoding matrix $\mathbf{D}$ so that each HOA component is corrected by the factor $1/F_n^R(\omega)$. Then, the decoding gains, corrected for the near-field effect of the loudspeakers can be obtained from the modified decoding Equation 3.27:

$$\mathbf{g} = \mathbf{D} diag\left(\left[\dots \frac{1}{F_n^R(\omega)} \dots\right]\right) \mathbf{b} \tag{3.73}$$

Compensation for the loudspeaker near-field can be assessed in the Figure 3.35 where a

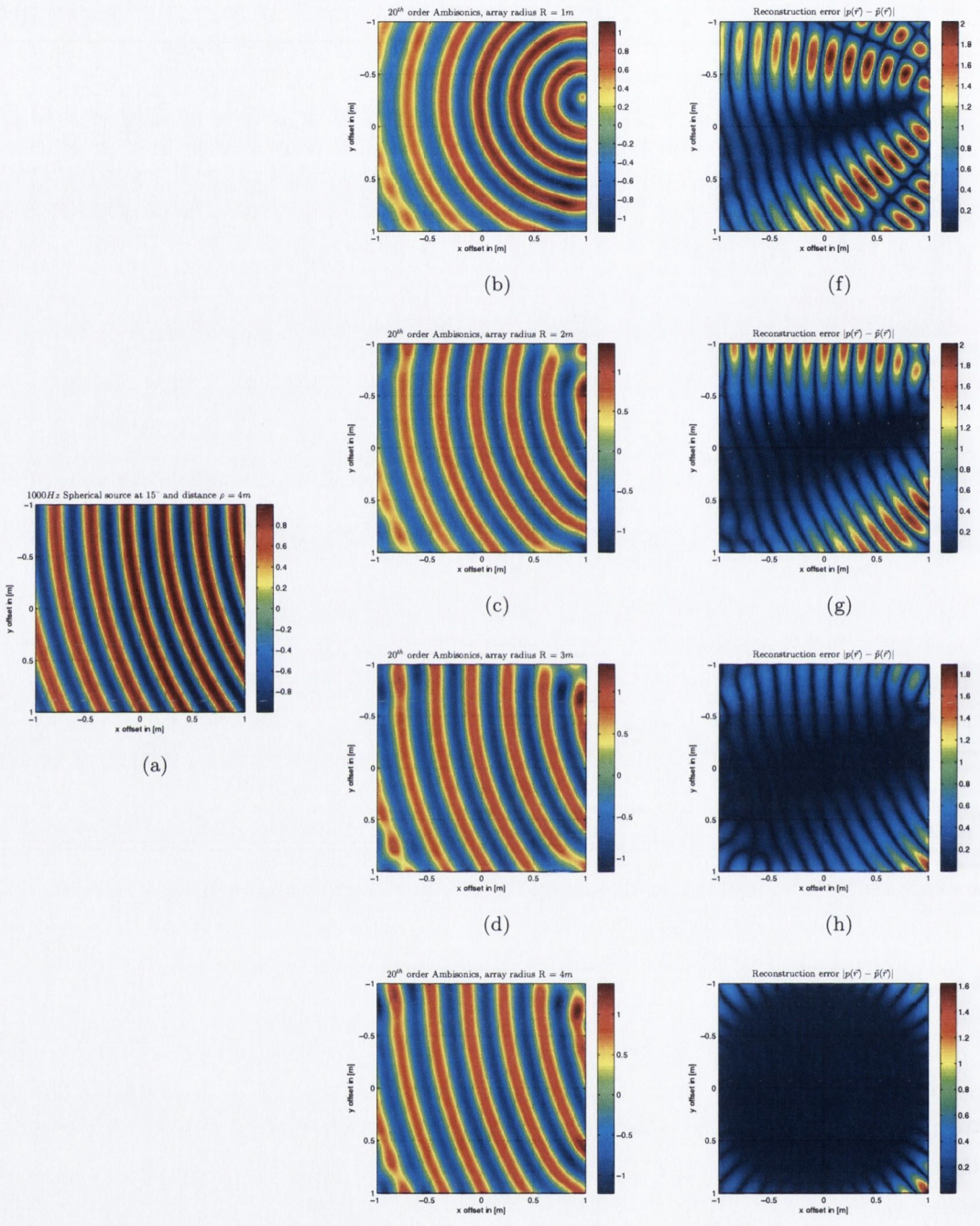
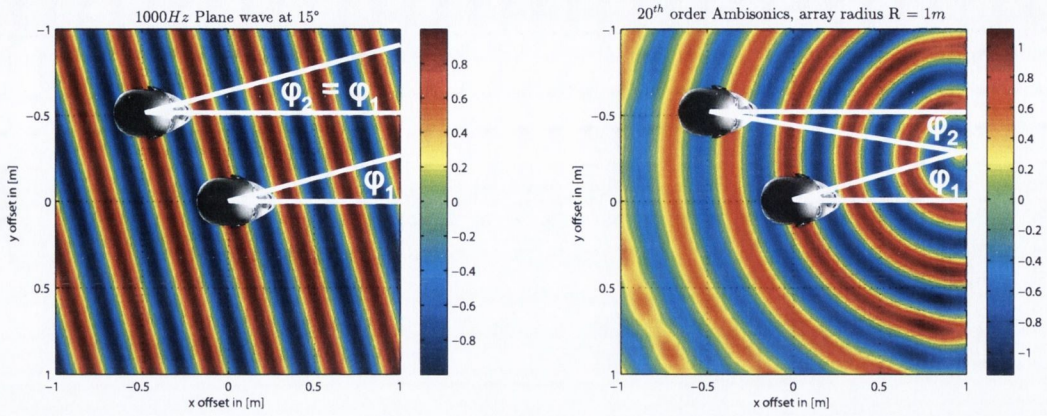


Figure 3.32: Reconstruction of a $1000Hz$ spherical sound source (a) localised at 15° azimuth and $\rho = 4m$ distance. Pressure field is approximated using 20^{th} order Ambisonics with 42 loudspeakers arranged in the diametrically opposed pairs with a radius of: (b) $1m$; (c) $2m$; (d) $3m$; (e) $4m$. Rightmost column ((f) - (i)) presents the reconstruction error in and around the listening area.



(a) (b)

Figure 3.33: Off-centre localisation error due to incorrect wave curvature.

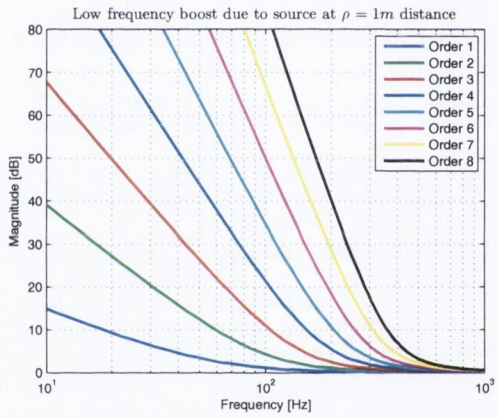


Figure 3.34: Low frequency boost due to spherical sound source at $1m$

$1000Hz$ plane wave is reconstructed using $20th$ order Ambisonics and the regular, circular loudspeaker array with the radius $R = 1m$.

So, near-field compensation would retrieve the correct shape of a plane wave while reproducing the sound field over an array at a finite distance R. However, in order to encode a sound source at a desired distance the near-field effect of the sound source must be also accounted for. There already exist software panners that allow for the reproduction of the near-field effect of the source [235] along with the directional information. Wiggins also investigated the encoding of distance information by the 1^{st} order SoundField ST350 and MKV microphones which can be subsequently utilised by the decoders [235]. In general, to correct for the proximity of both sound sources and the reproduction array at the same time, the inverse of the transfer function 3.72 can be rewritten as:

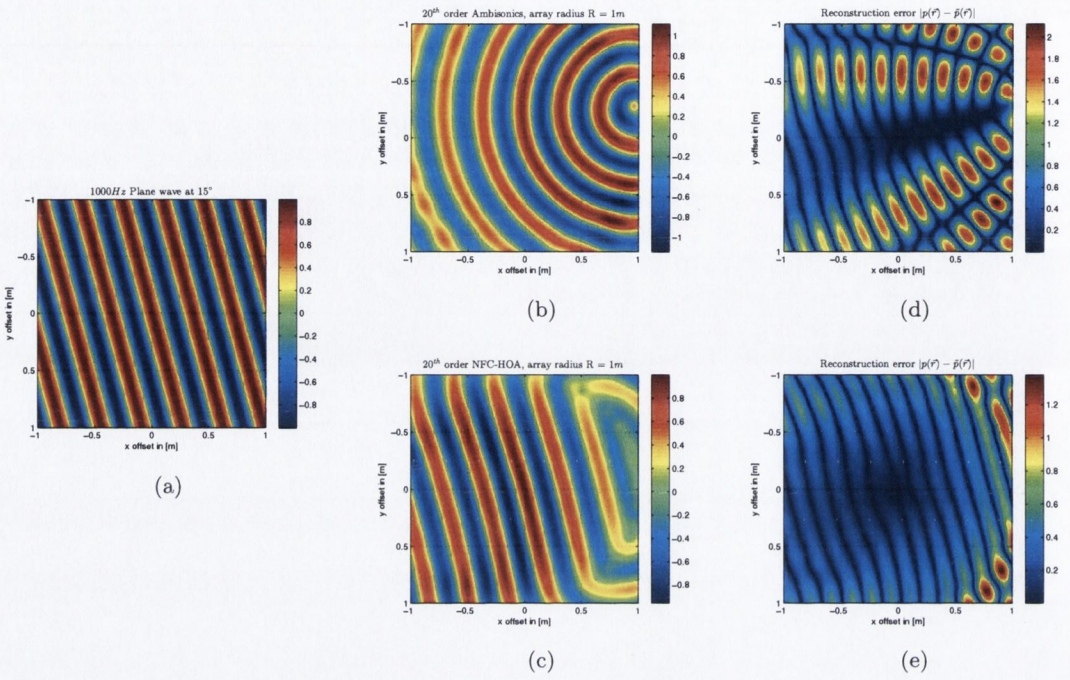
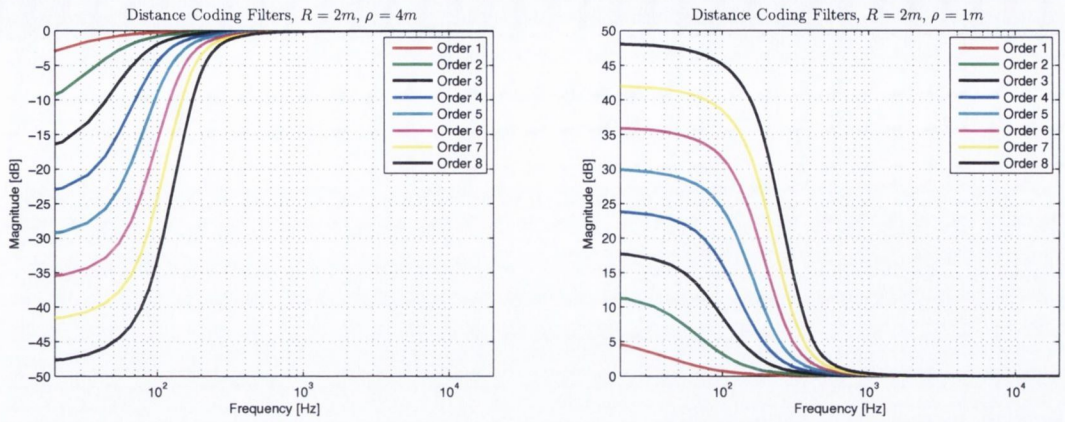


Figure 3.35: Basic reconstruction (b) and the near-field compensated reconstruction (c) of a $1000Hz$ plane wave (a) by the loudspeaker array at $R = 1m$. Rightmost figures (d) and (e) show that NFC-HOA significantly reduces the reconstruction error across the wide listening area.

$$H_n^{R,\rho}(\omega) = \frac{F_n^{\rho}(\omega)}{F_n^{R}(\omega)} \tag{3.74}$$

Daniel uses the term "near-field coding" or "near-field control" (NFC) when referring to "the combination of near-field effect (for a source distance ρ) and compensation (for a loudspeaker distance R)" [48]. By applying $H_n^{R,\rho}(\omega)$ to the individual HOA components one can compensate not only for the finite loudspeaker radius but also for the sound sources appearing outside or inside the array. An exemplary set of filters required for accurate wide-area rendering of sound sources outside the array (array radius $R = 2m$, sound source distance $\rho = 4m$) and inside the array (array radius $R = 2m$, sound source distance $\rho = 1m$) are shown in Figures 3.36(a) and (b).

From the practical point of view, the necessity to correct for the near-field effect of the sound sources and pre-compensating for the loudspeaker distances should be carefully considered, especially for lower orders. First of all, the bass frequencies are affected at most, so the extent to which errors will be introduced in the reproduction will be strongly programme material



(a) (b)

Figure 3.36: Distance coding filters required for HOA components up to and including 8^{th} order to correct for the sound sources: (a) outside the reproduction array (array radius $R = 2m$, sound source distance $\rho = 4m$; (b) inside the reproduction array (array radius $R = 2m$, sound source distance $\rho = 1m$.

dependant. Also, as was shown, for the centred listening position there is no risk of azimuthal bias due to wrong wave front curvatures. One has to note however that in order for sound sources in an auditory scene to be encoded correctly, their distance information must be available at runtime. When the scene contains dynamic or interactive elements and when reverberation had been applied to the sources it necessitates the filters to be re-computed and changed in real-time (e.g. whenever the sound sources change their distances with respect to the listener - centre of the array).

Moreover, as pointed out by Adriaensen in [2], digital realisation of NFC filters may prove to be problematic if limited numerical precision is available. This may lead to unstable filters unless the high precision arithmetic is used. For example, at the time of the writing of this thesis, the NFC filters could not be implemented in a straightforward way in the real-time, visual programming environment of *Pure Data* [181] because of the single-precision ($16bit$) floating-point numbers it operates on. However, Adriaensen in the same paper [2] proposes a modification of the NFC filters algorithm to allow for the single-precision filter design process. The practical realisation of his filters is also proposed using C++ classes.

3.4.5 Higher Order Sound Field Synthesis

Despite the fact that HOA, as opposed to FOA, can offer very sharp acoustic images, the synthesis of HOA sound fields is usually done by the means of encoding of dry or anechoic mono sources. The reason for this is the lack of the widely commercially available HOA microphones. Most of the HOA microphones are at the prototyping/research stages [24] and those which could

possibly offer HOA recording capabilities [147] are very expensive and also do not record "true" Ambisonic output. This is because true higher order spherical harmonic patterns are infeasible to obtain directly due to the fact that physical placement of multiple microphone capsules at a single point in space is impossible. Therefore, the higher order components usually need to be approximated by uniform sampling of acoustic pressure across the surface of a sphere [24].

That is why, it is very useful to analyse the directionality of first order sound fields, that can be nowadays easily recorded, in terms of their directional content. Two applications of this process are of importance in this work: (1) directional analysis of the B-Format SRIRs so that the their directional parts can be re-synthesised using higher order Ambisonic components. This approach will be used e.g. in investigations of the impact of early reflections on the perception of auditory distance in Chapter 5; (2) directional analysis of the B-Format SRIRs so that their directional components can be extracted out and replaced with another method for more accurate computation of the directional parts (e.g. early reflections). This approach will be utilised in the creation of real-time virtual auditory environments in Chapter 4.

The method of directional analysis of the incoming sound field has been proposed by Marimaa and Pulkki in [144]. It is based on the time-frequency derivation of an instantaneous acoustic intensity vector which describes the current flow of acoustic energy in a particular direction:

$$\mathbf{I}(t) = p(t)\mathbf{u}(t) \tag{3.75}$$

where $\mathbf{I}(t)$ is sound intensity, $p(t)$ is acoustic pressure and $\mathbf{u}(t)$ is particle velocity. It is important to note that $\mathbf{I}(t)$ and $\mathbf{u}(t)$ are vector quantities with their components acting in x, y and z direction. Now, let us recall that the B-Format signals comprise of one omnidirectional components (W) that can be used to estimate acoustic pressure, and also three directional components (X, Y and Z) that can be used to approximate acoustic velocity in the required direction x, y and z:

$$p(t) = w(t) \tag{3.76}$$

and

$$\mathbf{u}(t) = \frac{1}{\sqrt{2}Z_0}\left(x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}\right) \tag{3.77}$$

where $\mathbf{i}$, $\mathbf{j}$, and $\mathbf{k}$ are cartesian unit vectors, $x(t)$, $y(t)$, $z(t)$ first order Ambisonics signals and Z_0 is the specific acoustic impedance of air.

Thus, the instantaneous acoustic intensity vector in the frequency domain, approximated with B-Format signals can be expressed as:

$$I(\omega) = \frac{\sqrt{2}}{Z_0}\mathrm{Re}\{W^*(\omega)\mathbf{U}(\omega)\} \tag{3.78}$$

where $W(\omega)$ and $\mathbf{U}(\omega)$ are the short-term Fourier Transform (STFT) of the $w(t)$ and $\mathbf{u}(t)$ time domain signals, and * denotes complex conjugate. The direction of the vector $\mathbf{I}(\omega)$ corresponds to the direction of the *flow* of acoustic energy. That is why the plane wave source can be assumed in the $-\mathbf{I}(\omega)$ direction. The horizontal direction of arrival ϕ_S can by then calculated as:

$$\phi(\omega) = \arctan\left(\frac{-\mathbf{I_y}(\omega)}{-\mathbf{I_x}(\omega)}\right) \tag{3.79}$$

and the vertical direction:

$$\theta(\omega) = \arctan\left(\frac{-\mathbf{I_z}(\omega)}{\sqrt{\mathbf{I_x^2}(\omega) + \mathbf{I_y^2}(\omega)}}\right) \tag{3.80}$$

where $\mathbf{I_x}(\omega)$, $\mathbf{I_y}(\omega)$ and $\mathbf{I_x}(\omega)$ are the $\mathbf{I}(\omega)$ vector components in the x, y and z direction.

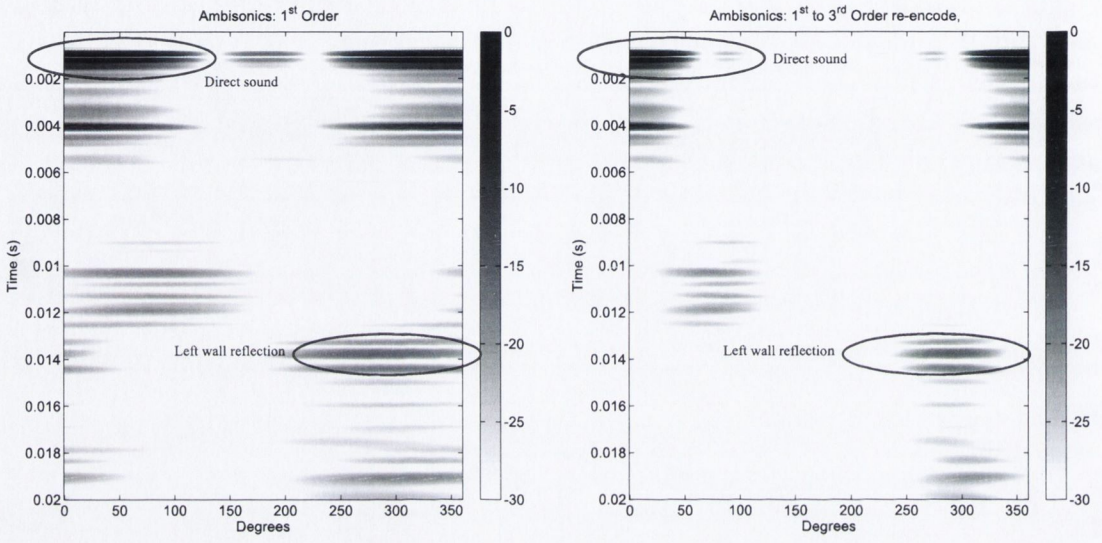
Now in order to be able to extract a directional portion from the B-Format SRIR, we have to estimate the *diffuseness* coefficient that is given by the magnitude of short-term averaged intensity referred to the overall energy density:

$$\psi(\omega) = 1 - \frac{\sqrt{2}||\mathrm{Re}\{W^*(\omega)\mathbf{U}(\omega)\}||}{|W(\omega)|^2 + |\mathbf{U}(\omega)|^2/2} \tag{3.81}$$

The output of the analysis can be subsequently subjected to spectral smoothing based on the Equivalent Rectangular Bands (ERB), as presented in [105]. The extraction of diffuse and non-diffuse parts of the SRIR is done by multiplying the B-format signals by $\psi(\omega)$ and $\sqrt{1-\psi(\omega)}$ respectively.

HOA re-synthesis is done by spherical harmonic encoding of the discrete sources into the directions obtained from the directional analysis and then by weighting i^{th} bands with the coefficient $\sqrt{1-\psi_i}$. On the other hand the diffuse field can be obtained simply by weighting each i^{th} band of the recorded B-Format SRIR with $\sqrt{\psi_i}$.

An exemplary re-encoding of the directional part of the 1^{st} order SRIR to the 3^{rd} order representation is shown in Figure 3.37 for a medium size lecture hall ($15.6m \times 5.85m \times 4.14m$). From this analysis, it is clear that due to re-encoding scheme, angular spread of acoustic energy diminishes and the direct and reflected sounds become more point-like.



(a) (b)

Figure 3.37: Ambisonic sound field resulting from 1^{st} order measurement with the Soundfield MKV system: (a) original 1^{st} order representation, (b) 3^{rd} order re-synthesis.

One limitation of this method is that for most A-Format microphones separation of capsules constitutes a problem and above some particular frequency spatial aliasing occurs [144]. For the *Soundfield* microphone the aliasing frequency lies around $10kHz$ [74]. Above this frequency, pressure and particle velocity components used for directional analysis are not estimated reliably any more.

The technique of directional analysis of 1^{st} and higher order re-synthesis will be utilised later in this thesis in the creation of the reverberation engine for real-time walk-through auralisations in (Chapter 4) and for studies of auditory distance perception in (Chapter 5) where more detail information on the implementation issues of the algorithm will be offered.

3.5 Binaural Reproduction

So far we have been considering different systems designed for loudspeaker-based reproduction of spatial sound. However, headphone reproduction offers several advantages over the previous mode, most notably increased privacy and immersion, and is also more tolerant to noise sources.

Moreover, it does not suffer from the sweet spot limitations of most loudspeaker reproduction techniques. That is why in this section the focus will be on binaural audio rendering presentations as a mean for delivering convincing spatial sound for virtual reality applications.

We have already mentioned in Chapter 2 that in order to auralise a monophonic source such that the sound field is reconstructed using headphones, one needs to convolve the source audio with the left and right Head Related Impulse Responses (HRIRs) pertaining to this particular position of the sound source. Thus, for multiple different source locations, many sets of HRIRs are required. Moreover, whenever the relative source-listener location changes, the sound field needs to be re-rendered for the new spatial configuration. However, switching of the directionally dependent HRIRs with source movement can lead to auditory artefacts caused by wave discontinuity in the convolved binaural signals [168]. Thus, another approach is required in order to dynamically change the sound fields and presented them over headphones.

3.5.1 Virtual Loudspeaker Approach

One possible solution to the dynamic binaural sound field rendering leads to the notion of the "virtual loudspeakers". This method was first introduced by McKeag and McGrath [140] and examples of its adoption can be found in the work of e.g. in Noisternig [165] and Dalenback [45]. In this approach, loudspeaker signals are generated according to a chosen playback methodology (e.g. VBAP, Ambisonics, WFS). However, these signals are not fed into the physical loudspeakers but instead, filtered with the left and right HRIRs corresponding to the spatial locations of these loudspeakers. Finally, the sums of left and right ear signals are fed into headphones. For example, to obtain the left ear headphone feed we have to perform:

$$L = \sum_{i=1}^{N} h_{Li} * q_i \tag{3.82}$$

where $*$ denotes convolution and h_{Li} is the left ear HRIR corresponding to the i^{th} loudspeaker location and q_i is its signal feed. The process is analogical for the right ear signal feed.

In the virtual loudspeaker approach, HRIRs are measured at the sweet spot, so the usual limitations of e.g. stereophonic systems are mitigated. The concept of forming the virtual loudspeakers from the regular octagonal array of loudspeakers is illustrated in Figure 3.38.

The block diagram in Figure 3.39 shows the signal flow in the virtual loudspeaker approach in the case of Ambisonics reproduction. The number of convolutions necessary for the binaural rendering is equal to twice the number of the loudspeakers used in the Ambisonic playback (two filters per virtual loudspeaker). Thus, in order to reproduce a sound source anywhere around the 360° azimuth, for the popular 3^{rd} order playback over the horizontal-only octagonal array, 16 HRIRs are employed, for 4^{th} order 20 HRIRs and so on. For 3-D reproduction, these numbers grow much faster and 3^{rd} order periphonic reproduction would already require 32 HRIR filters.

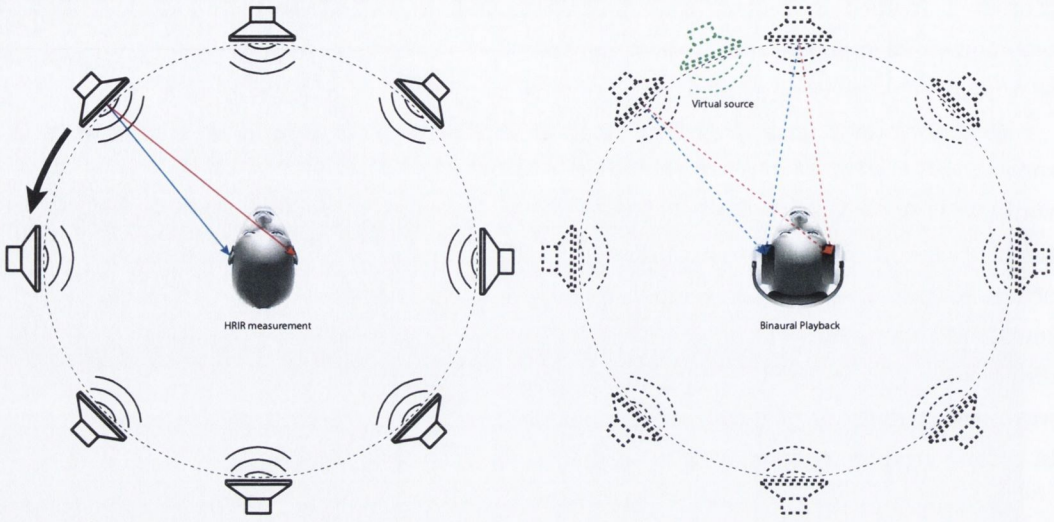


Figure 3.38: The virtual loudspeaker reproduction concept: (a) measurement of HRIRs corresponding to the spatial locations of all the loudspeakers in the setup; (b) playback of the loudspeaker signals convolved with HRIRs forming a 2-channel binaural steam

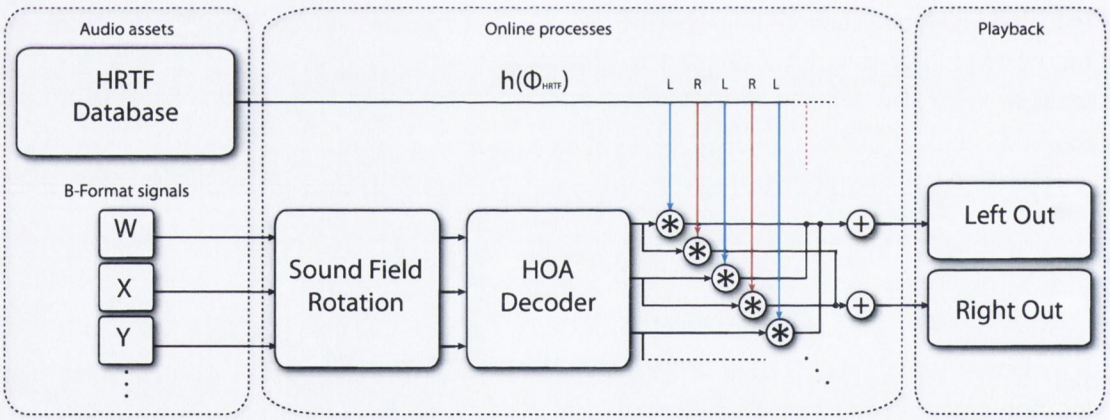


Figure 3.39: Binaural reproduction utilising "virtual loudspeakers" applied to Higher Order Ambisonics playback

The advantage of this approach is that still a limited dataset of HRIRs is usually necessary for all-around spatial sound reproduction. In theory, the directional resolution of the loudspeaker array and the spatialisation method used should be retained, especially that the simulated virtual array is formed around the centrally located listener (sweet spot). In practice, factors such as non-individualised HRIRs may deteriorate the listening experience, particularly for sources in elevation where the ILD and ITD cues are weak.

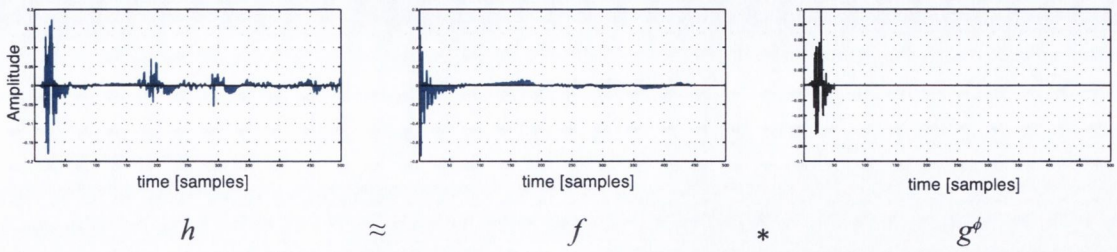


Figure 3.40: A concept of HRIR factorisation into a direction independent component (f) and a direction dependent residual g^{ϕ}

Another weakness of this method is that whilst the number of real-time filters necessary to represent all possible source locations is reduced, for large loudspeaker configuration it may still be significant. In most situations a block frequency-domain approach to this filtering process is utilised. However, given that the virtual loudspeaker feeds change interactively in real-time, a time-domain approach in a point wise manner avoids the inherent latencies introduced by block convolution in the frequency domain. The ideal solution to this problem would be to avoid the long filter kernels at run-time without the perceptual loss of localisation quality or timbre. A strategy for significant reduction of the filter lengths without artefacts will be presented next.

3.5.2 Optimised Virtual Loudspeaker Reproduction

In order to simplify and significantly reduce the computational burden associated with real-time virtual loudspeaker formation process, HRIRs (denoted $\mathbf{h}^{\phi}$) can be simplified by factoring each filter into the convolution of a direction independent subsystem (denoted $\mathbf{f}$) which is common to the whole set and a direction dependent residual (denoted $\mathbf{g}^{\phi}$). This idea is illustrated in Figure 3.40 where a 500-sample long HRIR filter is approximated as a convolutioin of a 451-sample long common (direction independent) filter and a 50-sample long direction dependent residual.

The long, directional independent component can be then pre-filtered with the audio assets off-line and only short direction dependent residuals used for forming of the virtual loudspeakers at run-time. This process is presented in Figure 3.41. Processing power savings gained in this way can be already substantial, particularly if the array of virtual loudspeakers is large (e.g. as in HOA or WFS reproduction) and if we consider that the run-time filter length reduction can be as much as 95% without perceptually noticeable artefacts [109].

The algorithm used in finding the common subsystem of a HRIR dataset is equivalent to finding the approximate greatest common divisor (AGCD) of the HRTF z-domain set. It can be formulated as a non-linear optimisation problem [128] where we attempt to find a global minimum of the reconstruction error function from N HRIRs, each m-point long:

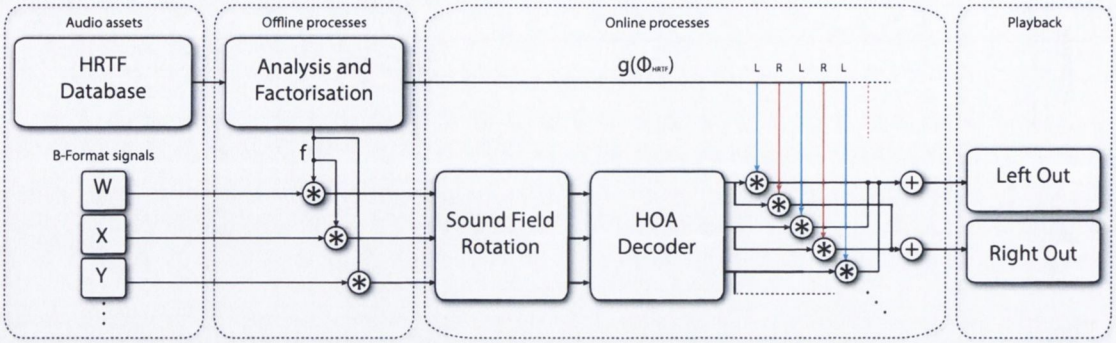


Figure 3.41: Optimised "virtual loudspeakers" reproduction utilising HRIR factorisation and off-line pre-processing applied to Higher Order Ambisonics playback

$$\min_{\mathbf{f},(\mathbf{g}^1,\cdots,\mathbf{g}^N)} \sum_{\phi=1}^{N} \|\mathbf{h}^\phi - (\mathbf{f} * \mathbf{g}^\phi)\|^2 \tag{3.83}$$

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

This problem can be solved using a variant of the well-known Gauss-Newton non-linear least squares algorithm with the exception that the usual step of linearisation around the current guess is already done, as the system is bilinear in $\mathbf{f}$ and $\mathbf{g}^\phi$.

Given an initial guess for $\mathbf{f}_0$, standard least squares can be used to find the residues, $\mathbf{g}^\phi$, which minimise the error between $\mathbf{f} * \mathbf{g}^\phi$ and $\mathbf{h}^\phi$. This $\mathbf{g}^\phi$ can then be used to generate a refined $\mathbf{f}$ again using least squares and hence a recursive process is defined.

Divisor-Quotient iteration

i=iteration count

1. Guess $\mathbf{f}_0$ $(i = 0)$

2. Solve for each residual, $\mathbf{g}^\phi$, as follows:

$$\mathbf{g}_{i+1}^\phi = F_i^\dagger \mathbf{h}^\phi \tag{3.84}$$

Where F_i is the convolution matrix formed from $\mathbf{f}_i$ and $\dagger$ denotes the Moore-Penrose pseudoinverse.

3. Solve for $\mathbf{f}_{i+1}$ using:

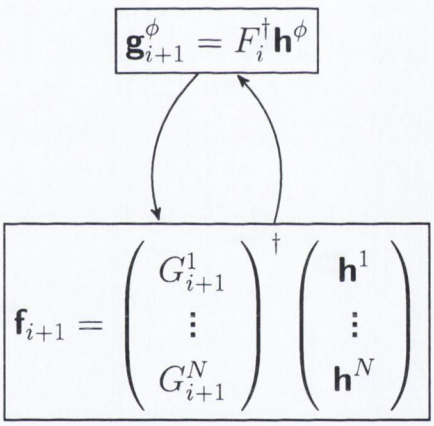
$$\mathbf{f}_{i+1} = \underline{G}_{i+1}^{\dagger}\underline{\mathbf{h}} \tag{3.85}$$

where $\underline{G}_{i+1} = \begin{pmatrix} G_{i+1}^{1} \\ \vdots \\ G_{i+1}^{N} \end{pmatrix}$ and $\underline{\mathbf{h}} = \begin{pmatrix} \mathbf{h}^{1} \\ \vdots \\ \mathbf{h}^{N} \end{pmatrix}$ is an Nm-point column vector

is an Nm-point column vector. G_{i+1}^{ϕ} is the convolution matrix formed from $\mathbf{g}_{i+1}^{\phi}$

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

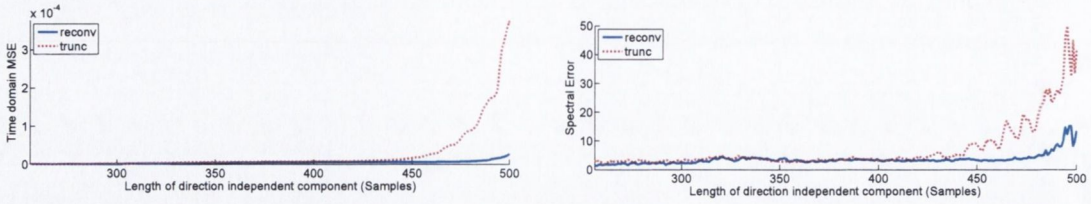
Visually, the above iterative algorithm can be represented as follows:



In order to evaluate the effectiveness of the algorithm objectively it is useful to investigate the differences between the original (h^{ϕ}) and reconstructed ($f * g^{\phi}$) HRIR sets. The mean square error has been used as the time domain error metrics and was computed for different lengths of direction independent components (f) as:

$$MSE = \frac{1}{NM}\sum_{n=1}^{N}\sum_{m=1}^{M}\left[h_(n,m) - h_r(n,m)\right]^2 \tag{3.86}$$

where h is the original HRIR set, h_r is the reconstructed HRIR set, N is the number of HRIRs in the dataset (in this case 72) and M is the length of an HRIR (in this case 512 samples). The frequency domain squared error in $[dB]$ between the magnitude spectra for different lengths of direction independent components (f) was calculated as:



(a) (b)

Figure 3.42: Comparative analysis of error metrics for HRIR factorisation (reconv) and truncation (trunc): (a) Time-domain mean square error; (b) Frequency-domain squared error $[dB]$

$$SE = \frac{1}{NK} \sum_{n=1}^{N} \sum_{k=1}^{K} [20log_{10}|H(n,k)| - 20log_{10}|H_r(n,m)|]^2 \tag{3.87}$$

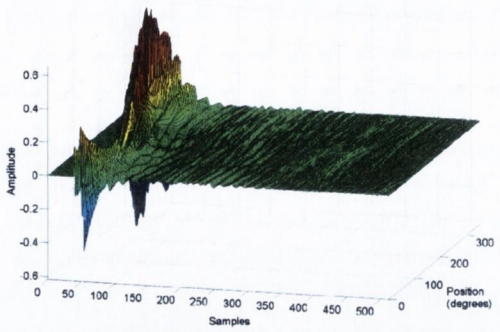
where H is the original HRTF set, H_r is the reconstructed HRTF set and K is the number of frequency bins. Figures 3.42 (a) & (b) show the error metrics obtained for the dataset of 72 HRIRs. These errors are compared to an alternative method of simple time-domain HRIR truncation of h^{ϕ} to the same length as the direction dependent component g^{ϕ}. The analysis show that truncation of more than approx. 60 samples from the original 512 point HRIRs results in increasing discrepancies between the factorised and truncated HRIRs. The max. spectral error for the method of factorisation is around $15dB$ for very long common filters f whereas for the method of truncation it reaches approx. $50dB$.

One of the problems with this method is that different initial guesses $\mathbf{f}_0$ may result in vastly different final results (i.e. $\mathbf{f}$ and $\mathbf{g}^{\phi}$ subsystems). The solution can be to use a regularisation technique that would allow to come up with the more psychoacoustically meaningful results whilst still maintaining low reconstruction error. This solution as well as more detailed objective analysis of the reconstruction errors is discussed in [134] and [136].

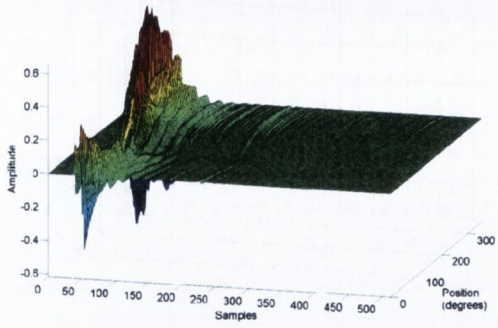
3.5.2.1 Perceptual Evaluation of Factorisation Algorithm

In order to subjectively evaluate the results from the factorisation algorithm, the performance of the virtual loudspeakers with reconstructed HRIRs was investigated through a subjective listening test. The test was conducted in the context of applying the HRIR factorisation to game audio [109]. The purpose of it was to assess whether the factorisation led to any perceptual differences when compared with full length virtual loudspeaker renders.

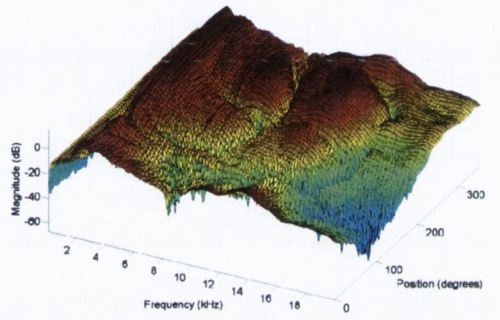
For this reason, three different types of HRIRs were incorporated into an Ambisonic based virtual loudspeaker array: 1) Full length HRIRs from the CIPIC database [6] (reference); 2) minimum phase HRIRs from the CIPIC database, with different lengths due to truncation; 3) minimum phase HRIRs from the CIPIC database, with different lengths due to the factorisation.



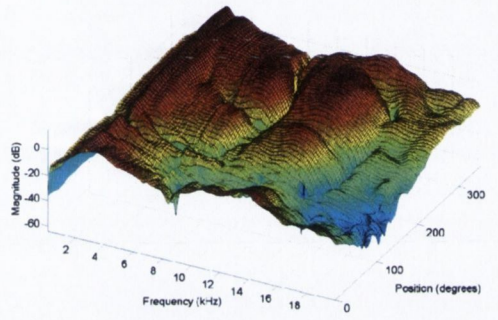
(a) Original left ear HRIR set



(c) Reconstructed HRIRs given a 450 sample direction independent component.



(b) Original left ear HRTF set



(d) Reconstructed HRTFs given a 490 sample direction independent component

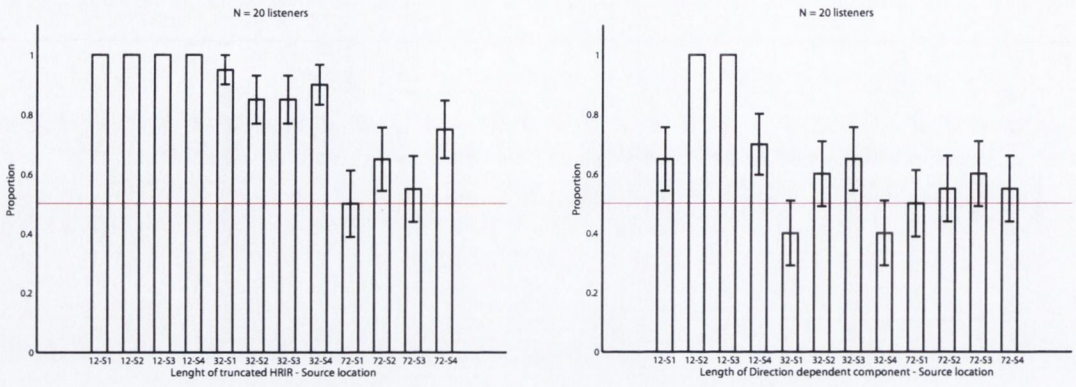
Figure 3.43: Original HRIR/HRTFs and their reconstruction from the convolution of two subsystems $\mathbf{f}$ and $\mathbf{g}^{\phi}$

Table 3.3 outlines the main parameters of the test. The source was pink noise bursts, from four different source positions: front, front lateral, rear and rear lateral (labeled S4, S1, S2, and S3, respectively).

HRIR Type	Full length (reference), Min. phase - truncated, Min. phase - factorised
Run-time HRIR Length	12 samples, 32 samples, 72 samples.
Angle	67.5° (S1), 157.5° (S2), 247.5° (S3), 337.5° (S4)

Table 3.3: Parameters for virtual loudspeaker renders in subjective assessment

The test was designed as an ABX comparison, where the listener had to determine which of



(a) (b)

Figure 3.44: Listening test results for virtual loudspeaker arrays from truncated HRIRs (a) and factorised HRIRs (b). A score close to 0.5 means that the simplified HRIRs cannot be distinguished from the original non-optimised HRIRs.

the A or B samples was the reference (full virtual loudspeaker render). At any time either A or B could be the reference file or the factorised virtual loudspeaker render.

Virtual loudspeaker renders using truncated HRIRs were also tested for comparison, as simple truncation is a commonly used technique for HRIR simplification. The 20 listeners employed in the tests were second year Music Technology students, each under 35 years of age and well experienced in music production.

In the comparison, the subjects were asked to concentrate on localisation deviation and timbre change. Performing the tests in ABX form also removed the need for individualised HRIRs, since subjects were asked only to identify changes between the reference and factorised virtual loudspeaker reproductions as opposed to absolute measures, such as angle of localisation.

AKG K-601 headphones where used as they are open back, have matched transducers and sit snugly on the subjects' ears. This satisfies the criteria described by Adams et al. [1] as being optimal for binaural synthesis i.e. minimising any distortion of localisation cues. The comparative nature of the listening test and the quality of headphones negated any need for headphone equalisation.

The results of the listening tests can be found in Figure 3.44(a) & (b). Here we plot the proportion of listeners who could identify the target reference during each test phase. A '1' means that all listeners were able to identify the reference. A '0' means that no listeners chose the reference source. A value of 0.5 means that the number of listeners selecting each source was approximately equal implying there was no perceptual difference between the signals processed with the full and factorised HRIRs. The proportions are plotted alongside the standard error for each test.

In the case of the minimum phase HRIR truncation, we clearly see that as the filter size is

reduced the rendering accuracy of the virtual loudspeaker array diminishes, and the proportion of listeners who can identify the target reference increases. At filter lengths of less than 32 taps, this is clearly the case, supported by subject feedback of localisation distortion at the lateral positions as well as significant low frequency timbre change. Such effects are not as prominent, however, using the factorised HRIR dataset. Here the proportion of listeners who identify the reference is statistically significantly lower for both the 72 and 32 tap filters and is close to 0.5 in all cases. Here the proportion of listeners who identify the reference is lower, by a statistically significant amount, for both the 72 and 32 tap filters and is close to 0.5 in all cases. For sources at positions S1 and S4 the method improves the localisation accuracy with 12-tap filters, but we noted that listeners perceived some differences from the reference in the cases of rear and rear-lateral source presentations from S2 and S3, as predicted from the mean squared error. The main difference reported for this HRIR length was timbral, as opposed to the localisation accuracy, although externalisation was also reported to diminish.

3.5.2.2 Further Optimisation

Further optimisation can be done to binaural Ambisonic reproduction when we realise that process of the virtual loudspeaker formation by HRIR filtering can be incorporated into the HOA decoding stage. As soon as the final reproduction loudspeaker setup is known, the HRIRs used for forming of the virtual loudspeakers can be encoded into the spherical harmonics domain. This step creates a set of short, decoding filters that are applied directly to the B-Format signals. However, for proper HRIR reconstruction, the B-Format signals need to be pre-filtered with the direction independent component off-line. For example, for any arbitrary N loudspeaker reproduction setup, we have to perform:

$$W_{HRIR,L} = \sqrt{2}\sum_{i=1}^{N} g_{i,L}^{\phi} \tag{3.88}$$

$$X_{HRIR,L} = \sum_{i=1}^{N} \cos(\phi_i)\cos(\theta_i) g_{i,L}^{\phi} \tag{3.89}$$

$$Y_{HRIR,L} = \sum_{i=1}^{N} \sin(\phi_i)\cos(\theta_i) g_{i,L}^{\phi} \tag{3.90}$$

$$Z_{HRIR,L} = \sum_{i=1}^{N} \sin(\theta_i) g_{i,L}^{\phi} \tag{3.91}$$

The output from the filters is summed together forming the left and right ear signals. For example, for the left ear signal it is done by:

$$Left = (W' * W_{HRIR,L}) + (X' * X_{HRIR,L}) + (Y' * Y_{HRIR,L}) + (Z' * Z_{HRIR,L}) \tag{3.92}$$

This approach is particularly useful whenever the number of reproducing loudspeakers significantly exceeds the number of B-Format channels. This is because the number of decoding filters is equal to the number of B-Format channels multiplied by 2 (to account for the left and right ear).

The final block diagram of the computationally optimised binaural rendering system is illustrated in Figure 3.45.

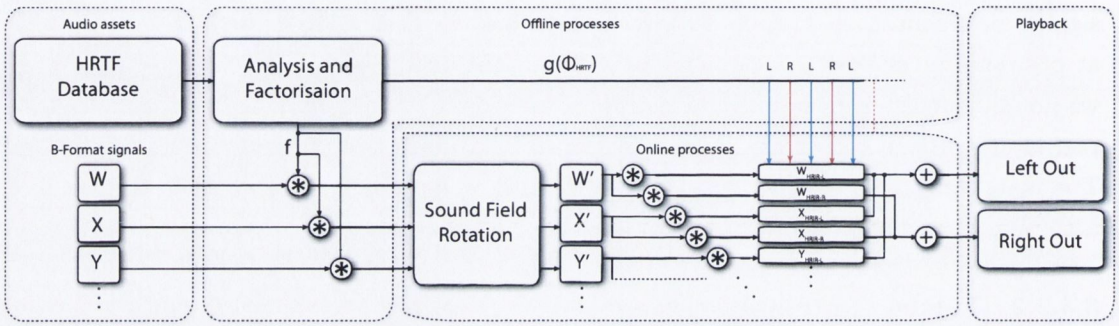


Figure 3.45: Optimised "virtual loudspeakers" reproduction utilising HRIR factorisation, off-line pre-processing and spherical harmonics encoding of the residuals, applied to Higher Order Ambisonics playback

3.6 Conclusions

In this chapter a review of state-of-the-art spatial audio reproduction methods was presented. This comparison was made mainly from the point of view of the adaptation of a certain system to the virtual reality presentations, i.e. where the robust real-time (and often interactive) rendering of the auditory information is required.

2-channel stereophony was first presented mainly as a base for further investigations into the multichannel stereo techniques. It was intended to show how virtual sound images are formed in between two loudspeakers and what psychoacoustic phenomena allow us to interpret multiple coherent sounds as a single source. The multichannel stereophony was shown to extend the foundations of the 2-channel system to other dimensions by the means of pair-wise panning concept. The notion here was that any arbitrary virtual source direction can be realised by the contribution of two loudspeakers. However, the creation of virtual acoustic images in this way is not equally effective in all directions. This is because one-sided loudspeakers are unable to correctly reproduce required ILDs and ITDs for lateral source directions.

VBAP was presented as a vector reformulation of the stereophonic panning law that makes it possible to form phantom sources in both horizontal and vertical directions. It also allows for arbitrary reproduction setups but the virtual sources are always created by selecting relevant pairs or triplets of loudspeakers. In this manner, the number of required reproducing loudspeak-

ers is always minimised leading to concentration of acoustic energy in a particular zone of the array. This, in turn, optimises the energy vectors that were shown to be good predictors of mid and high frequency localisation. However, changing the number of reproducing loudspeakers with e.g. dynamically panned sources causes audible artefacts (tonal colouration) that listeners are sensitive to [105].

The problem is less apparent in Wave Field Synthesis which makes use of all the available loudspeakers in the array in order to render the actual sound field as would have resulted from real sound sources in a real-world situation. These loudspeakers are used to re-synthesise acoustic wave fronts and that is why their amount and spacing are two crucial factors in the rendering process. For practical reasons, WFS arrays are usually truncated to horizontal-only arrays so elevation is either missing or has to be simulated using other methods, e.g. stereophony [81,237]. Also, finite loudspeaker spacing causes the wave field to be reconstructed correctly only up to a particular frequency, known as *spatial aliasing frequency.*

Nevertheless, it has been shown at several occasions (e.g. in [237]), that WFS indeed offers good localisation over a wide listening area so it seem to be perfectly suited for distributed audiences like in cinemas, concert halls or theatres. Initially, it was also hypothesised that due to correct reconstruction of acoustic wave curvatures, the WFS method could also lead to better discrimination of auditory distance. However, it was experimentally proven by Wittek [237] that this particular parameter in isolation is not able to produce solid sensations of distance for subjects.

Nonetheless, in spite of the great potential, support of multiple listener's within the listening area and movement, the usability of the WFS for domestic purposes was deemed rather minimal mostly due to the required number of audio channels and sophisticated and costly physical loudspeaker setup.

We identified that for a single user (e.g. a game player) Ambisonics can constitute a better alternative since in some aspects this method marries together the holophonic approach of WFS whilst keeping the number of reproduction channels as low as possible, as in stereophony. Ambisonics was shown to be scalable, which means that the required accuracy can be achieved by increasing the number of channels carrying the spatial description of the sound field. Ambisonics, or rather Higher Order Ambisonics can therefore offer localisation as good as e.g. VBAP but since all the loudspeakers are contributing to the sound field reproduction at all times, the colouration artefacts are significantly reduced [105]. It is well known in Ambisonic loudspeaker reproduction, that as the order of sound field representation gets higher, the localisation accuracy increases due to greater directional resolution. However, there are still many unanswered questions with regard to e.g. rendering of auditory distance at different orders. As we realised earlier in Chapter 2, early reflections may play a significant role in determination of the auditory source's distance. Thus, checking for the perceptual impact of the direct sound and early reflections spatial resolution is yet to be investigated. This topic will be discussed in Chapter 5.

Lastly, binaural rendering of spatial sound was identified as an effective way of creating

personalised sound fields. However, a standard methodology of convolving the monophonic sources with left and right Head Related Impuse Responses proves to be problematic as far as the multiple source locations are concerned. A "virtual loudspeaker" approach was shown to provide a more practical way of binaural sound rendering that minimises the size of the HRIR dataset but at the cost of increased processing power required for multiple, parallel HRIR filtering. However, two methods were proposed to optimise the virtual loudspeaker reproduction: 1) factorisation of the HRIRs into a long direction independent component and a set of short direction dependent residuals that are used at run-time; 2) Using the Ambisonic approach where the HRIR filters are encoded using the spherical harmonics basis functions and incorporated into the decoding process.

So far we have formulated a theoretical scaffold for recording and subsequent reproduction of sound fields. We will now use this knowledge in order to discuss the creation process of Virtual Auditory Environments (VAEs).

4

Real-Time Walk-Through Auralisations

In this chapter a practical framework is presented for creating interactive Virtual Auditory Environments with multiple sound sources located in a reverberant space. The discussion is centred around two examples of recording real musical performances and their subsequent interactive reproduction with the use of visual displays and multichannel loudspeaker arrays or headphones. The proposed methodology considers the parametrisation of real-world sound fields and their subsequent real-time auralisation using a hybrid Image Source Method/measurement-based convolution reverberation approach. First Order (FOA) and Higher Order (HOA) Ambisonics are utilised together in a single system to provide the optimal and psychoacoustically justified framework for interactive spatial sound reproduction.

Section 4.1 introduces the idea of headphone-based auralisation and reviews current approaches to auralisation in architectural and urban acoustics as well as in virtual reality applications and video games. It also proposes a novel approach to auralisation which is relevant to high fidelity, real-time and interactive applications.

Section 4.3 shows an application of the proposed method and describes the initial experimental work in which the performance of the traditional Irish trio in the reverberant medium-size lecture hall is recorded and subsequently recreated in the virtual settings. This section elaborates on the methods and techniques used for the audio-visual event reconstruction as well as the challenges that had to be addressed in the course of the work. The summary of the work is provided and possible future refinements are suggested.

Based on the initial study, the next section applies similar methodology to a larger scale project which is the live recording and subsequent virtual interactive reproduction of the per-

formance of the 19-singer choir of the Christ Church Cathedral in Dublin. In this initiative, a significant body of work had to be performed prior to the recording where both acoustical and geometrical measurements of the cathedral have been collected using the state-of-the-art measurement techniques. This work in itself significantly enriches the current state of historical preservation of acoustic landmarks in Dublin.

In Section 4.5 a formal evaluation of the proposed auralisation method by the mean of subjective listening tests is presented.

Finally, Section 4.6 summarises the work on real-time interactive auralisation, appraises it critically and gives perspectives for possible future improvements.

4.1 Introduction to Headphone Based Auralisation

The term "auralisation" refers to the process of presenting the listener with a recording of an acoustic event including the information about the acoustic environment in which the event takes place. Ideally, the process of auralisation would provide the level of faithfulness and immersion that the virtual event cannot be distinguished from the real event.

Ideally, in the process of auralisation a sound source should be represented as an anechoic pressure signal of an instrument, human voice or any other sounding object. However, for practical reasons, sufficiently dry microphone recordings (e.g. taken within the critical distance of a room) are also acceptable [112].

A dry sound source recording stored as a monophonic audio file contains information only about source radiation in a single direction. However, in real-world situations radiation patterns of different sound sources vary with recording angles. In the perfect scenario, new recording of a source should be used depending on the current listener position relative to the sound source. Several approaches to measuring and synthesising source directionality have been proposed in the literature, primarily with regard to Wave Field Synthesis reproduction [3]. The approximation can be done based on e.g. spherical harmonics decomposition of the 2-D radiation function [3, 141] or least-squares approximation [105]. The capture of source directivity is traditionally achieved using arrays of microphones surrounding a performer and databases of such measurements are available [170, 171, 175]. Directivity filters can be applied to single monophonic instrumental/voice recordings of a performance to simulate the change in frequency response with source/listener movement. A simple approach has been proposed by Giron [80], where the interference of several monopole virtual sources is used to synthesise 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 synthesised at reproduction using monopoles or virtual cardioid patterns [104]. This is not a practical solution to a real performance situation however.

Interactions between the sound source and room boundaries are encapsulated in the Room Impulse Responses. These responses are unique to different acoustic spaces but also they change with source-receiver spatial configurations. They can be measured *in-situ* and used in a so-called data-based auralisation. There are also methods of synthesising the Room Impulse Response based on room parameters.

In recent years, the formation of auditory scenes based on real-world spaces has benefited greatly from the use of convolution reverberation techniques and a significant body of work has been presented illustrating their possibilities and limitations [107,111,144]. Whilst such methods allow us to take an "acoustic snapshot", they do not accommodate a fully interactive listening experience, as in real-world listening. The acoustic changes that occur due to the movements of both the source and the listener within the space are an important aspect in the perception of music performance [105]. However, current methods in computational-based auralisation correctly consider this aspect through computing grids of acoustic response measurements over the entire computer model for different source/receiver configurations [45, 194]. Real-time auralisation is then implemented by choosing the appropriate acoustic response (or interpolating across responses [108]) for the intermediate source/listener positions. However, grid measurements are infeasible in the context of real-world acoustics as they are tedious and prone to error. Furthermore, one must not neglect the fact that performance within reverberant spaces such as halls or cathedrals will be markedly different than that in an anechoic chamber, due to the performer interaction with the acoustics [31].

Therefore, in general the convolution approach to the representation of room responses assumes that the source-listener-room interaction is one that is linearly time-invariant (LTI). In reality, RIR changes significantly with the relative spatial positions of the source and the listener. In this regard, it is a challenge to be able to capture the performances within the reverberant space such that this aspect is preserved whilst recording for an interactive auditory scene. However, if the geometric properties of the performance space, as well as the frequency dependent absorptive characteristics of the wall materials in the room are known, then the acoustic response can be computed instead. Highly accurate results can be achieved using wave-based methods, such as the Finite Element Method (FEM), Boundary Element Method (BEM) [176] or Finite-Difference Time-Domain (FDTD) method [200]. However, due to computational expense, such methods are generally limited to low-frequency RIR estimation. On the other hand, geometric-based solutions to calculating RIRs, such as the Image Source Method (ISM) [8] or ray-tracing are well suited to the mid- to high-frequency regions, although they do not consider phenomena such as diffraction or scattering. However, calculation of the propagation delays and magnitudes at low reflection orders using image sources is well suited to real-time auralisation, provided that room geometry is simple. Otherwise, further simplification to the modelled geometry might be deemed necessary as long as the real-time auralisation is concerned.

Finally, hybrid reverberation algorithms which combine computational and measured impulse responses have also been proposed, but have largely focused on the synthesis of the diffuse decay

as opposed to early reflections [33, 214]. In this work, we are concerned with the real-time rendering of the early reflections in conjunction with pre-rendered diffuse field recordings for walk-through auralisation.

At the reproduction stage, the direction of an auditory event can be controlled by using audio spatialisation techniques as outlined in Chapter 3. If the loudspeaker auralisation is considered, the process of spatialisation will provide the discrete loudspeaker signals which, when played back over a dedicated array of loudspeakers, will recreate the auditory scene. If the headphone reproduction is considered, the directional information is imposed on the sound source by the means of HRTF filtering.

All the above interactions between the directive sound source, reflective acoustic environment and the body of the listener are contained in Binaural Room Impulse Responses (BRIRs).

Because any auralisation system can be constructed using different building blocks, different results are to be expected in terms of the audible output and also required processing power. That is why some systems are well suited for real-time auralisation and some are designed to work off-line. Also, some systems are expected to provide a high fidelity (high faithfulness) output whereas in others it suffices to only approximate acoustic conditions to a certain degree. The next section reviews already existing solutions in two most common applications of auralisation - architectural/urban acoustics and virtual reality/gaming.

4.1.1 Auralisation in Architectural and Urban Acoustics

Auralisation can be an excellent tool in assessment of the quality/defects of acoustic environments such as concert concert halls or auditoriums [193]. That is why it should not come as a surprise that its applications in architectural/concert hall acoustics resulted in the development of several methods including full software packages offering comprehensive tools for auditioning of existing and conceived spaces. A short review of them is offered.

Obviously, the most straightforward way of quality assessment of the room/hall is real auditioning of sound sources in the physical acoustic space. However, this is usually not sufficient if a deeper analysis or a comparative study is needed. In this case, BRIRs can be obtained by the means of measurement using artificial heads or real subjects.

The above approaches however suffer from being time consuming and in some cases impractical. Also, this method cannot be applied to conceived spaces or, in a straightforward way, real-time and interactive auralisation. That is why in architectural and urban acoustics, a lot of research and development has been devoted to the problem of computational modelling of BRIRs.

From the information gathered by the author about existing software packages, besides auralisation they focus quite heavily on computing ISO acoustic parameters [98] of spaces as well as allowing the user to use databases of measurements including loudspeaker responses, wall material properties (absorbers, diffusers), etc.

From the point of view of RIR modelling, the most common are geometrical approaches that include Image Source Method and Ray Tracing. However, it is usually possible to use multiple methods and compare the results or even apply a hybrid combination of these methods. The most popular software packages used in architectural acoustics are ODEON[1], EASE[2], CATT-Acousitc[3], Ramsete[4], RayNoise[5] or Bose Auditioner[6]. The Bose Auditioner system is a software/hardware solution in which a specifically designed near-field speakers are used in order to provide the listener with the binaural signals in which the influence of the listening room acoustics is minimised.

4.1.2 Auralisation in Video Games and Virtual Reality Applications

On the other hand, video game audio engines is a fast moving branch of a large game development and middleware market dominated mainly by closed-source commercial software [50] in which the focus is on speed rather than the acoustic fidelity or faithfulness of auditory scenes. From the game technology surveys conducted by DeLoura [50] in 2009 and 2011, *FMOD*[7] and *Wwise*[8] are the two most widely used tools for video game audio. However, because of the commercial nature of these tools, there is a limited amount of information about the specific implementation of the audio DSP for the purpose of in-game auralisation.

There are other, open-source alternatives such as *OpenAL*[9] Application Programming Interface (API) which can be used for the purpose interactive 3-D positional audio. For example, *Rapture3D*[10] uses OpenAL and provides Ambisonics-based spatialisation (up to 4^{th} order) and binaural rendering of game audio using a choice of 5 HRTF sets to accommodate for different head shapes.

4.1.2.1 Reverberation in Video Games

In the vast majority of cases, video game auralisation uses procedural reverberation algorithms, amplitude based sound spatialisation and HRFT filtering for binaural reproduction. Procedural reverberation is usually obtained by using an algorithm which implements Feedback Delay Networks (FDN) in order to recreate dense but controlled repetitions of the input (direct) signal [69]. It is usually possible to specify basic reverberation parameters associated with the size, character and acoustics of the chosen environment such as reverberation time, early reflections strength,

[1] http://www.odeon.dk/
[2] http://ease.afmg.eu/
[3] http://www.catt.se/
[4] http://www.ramsete.com/
[5] http://www.lmsintl.com/RAYNOISE
[6] http://worldwide.bose.com/pro/en_us/web/auditioner_playback_system/page.html
[7] http://www.fmod.org/
[8] https://www.audiokinetic.com/
[9] http://openal.en.softonic.com/
[10] http://www.blueripplesound.com/product-listings/gaming

early reflection pre-delay etc. It is also possible to use presets (e.g. FMOD_PRESET_ROOM, FMOD_PRESET_AUDITORIUM, etc.).

Due to ongoing improvement of the available processing power in the modern video game computers and consoles, convolution based reverberation has also started to be present in games[11]. However, computational load associated with the convolution operation is directly related to the length of the RIR used in the game. That is why, it is recommended to use short or trimmed/gated responses, especially if the computational cost must be strictly kept low. Besides that, RIRs are only accepted in the 1- and 2-channel stereo formats which do not contain full spatial information about the complex spatial interactions between the sound source and the environment.

The other approaches to creating reverberation for in-game environments include statistical acoustics [160] in which randomly generated secondary sources imitate reflective surfaces of some in-game environment. This is a fast but not particularly accurate solution, especially as long as the early part of the RIR is concerned.

4.1.2.2 Binaural processing in Video Games

Binaural processing is already used in games and in most of the cases is done by the means of software extensions to popular game engines. For example, a well documented system called CLAM [167] has been developed on Linux platform to extend the functionality of *Blender Game Engine* (*BGE*). CLAM performs sound spatialisation using Ambisonics although the authors do not state what order is currently supported. The full RIR is generated at runtime using a ray-tracing algorithm and which is based on the room geometry from the game. Finally, the output can be presented using loudspeakers or headphones which is performed by decoding the sound field to the required number of channels. No details however are given about the B-Format to binaural decode.

In FMOD, binaural output can be achieved e.g. via BR2[12] which can simulate the direct sound and up to eight early reflections per source as well as diffuse reverberation. No scientific documentation is available however that will outline the exact implementation of the algorithms used.

It is also possible to achieve binaural output from a game using devices or algorithms that convert surround sound output (e.g. ITU 5.1 or 7.1) to two channels. This is analogous to creating virtual loudspeakers as outlined in Section 3.5.1 although in most of the cases, no scientific documentation is available for commercial products. One example of such a solution is Dolby Headphones[13] which delivers up to 7.1 channels of surround sound over headphones. A similar solution has been commercialised by DTS which is called Headphone:X[14] and supports

[11] https://www.audiokinetic.com/products/wwise-add-ons/convolution-reverb/

[12] http://mnsp.ch/products.html

[13] http://www.dolby.com/us/en/consumer/technology/home-theater/dolby-headphone.html

[14] http://www.dts.com/professionals/sound-technologies/headphonex.aspx

up to 11.1 channels of audio. None of the above systems however use head-tracking to stabilise the virtual loudspeaker sound field.

Commercial solutions which use head-tracking in the process of virtual loudspeaker creation are for example Smyth Research Realiser A8[15] or Beyerdynamic Headzone[16]. These are hardware solutions with external digital signal processors and head-tracking devices. No data has been found about their use in video games though.

4.1.3 State-of-the-art Auralisation Systems for Virtual Reality Applications

A general trend can be observed in research and application of real-time auralisation systems where deterministic parts of the BRIR are usually modelled using geometrical acoustics and the late reverberation uses some different approach, most notably an artificial reverberation. Below are some examples of such systems.

Sound Laboratory at NASA has been involved in the development of real-time auralisation solutions since the late eighties [230]. The current version of the system uses head-tracking and individualised HRTFs [11]. It provides a real-time binaural auralisation solution for aerospace display research, training, enhanced communications or improved situational awareness and is available as an Application Programming Interface (API) [18].

Some systems utilise external DSP processors for the purpose of real-time auralisation. First documented devices to do so were Convolvotron [66, 228] (1990) and HURON [192] (1994). Around the same time (1993-1996) a system named SCATIS [27] was developed in Ruhr-Universität Bochum with its primary purpose to be used in psychoacoustic research. However, in 1998, work begun on a software-based successor system which was named IKA-SIM[17] [203]. Generally, with the introduction of more powerful general purpose computers, the hardware approach seems to be slowly abandoned.

The auralisation system developed in the Acoustic Lab at South China University of Technology [241] utilises rotation and also translation information about the user's head using head-tracking. Also, the principal component analysis (PCA) technique is used for efficient binaural rendering of up to 280 virtual source at different distances.

A lot of research has been devoted to the development of real-time auralisation in the Helsinki University of Technology (currently Aalto University). A system called DIVA[18] (Digital Interactive Virtual Acoustics) has been developed since around 1994 [96,199]. The current version of the system renders static and dynamic virtual sources using their radiation patters, computes room reflections taking into account reflective and absorptive properties of the boundaries as well as air absorption.

[15] http://www.smyth-research.com/products.html

[16] http://europe.beyerdynamic.com/shop/headzone-headphone-surround-system.html

[17] http://www.ruhr-uni-bochum.de/ika/forschung/forschungsbereich_martin/speech_audio_processing/aud_virtual_environment/IKA-SIM.htm

[18] http://www.tml.tkk.fi/Research/DIVA/

Quite a different approach to VR auralisation has been proposed by Raghuvanshi et al. [188] which is based on physical synthesis of the source sound instead of using pre-recorded material. Their approach uses mechanical models with springs and masses for sound generation. These models are based on discretised geometries of 3-D objects. Whenever objects collide with each other, mechanical vibrations are translated into the domain of audio. However, this approach focuses so far on sound generation and does not take into account its propagation in a reverberant environment nor its interaction with the listener which are both key features in the process of auralisation.

Lastly, besides geometrical/artificial reverberation methods, some promising approaches have been made to create a real-time auralisation using wave-based acoustics (FDTD) and utilising massively parallel Graphics Processing Units (GPUs) [198]. However, due to enormous computational complexity of a full frequency range simulation, it has been limited to low- and mid-frequency ranges. Alternatively, pre-computing of SRIRs and using densely populated grid points was also implemented [210]. However, with the ongoing progress and increase of computing power, it is possible that further advancements in this field will be made. Until then, there is still an urge for improvement and optimisation of the other existing methods so that the auralisation can be performed in real-time including multiple sound sources, interactivity and 3-D visualisations.

4.2 Proposed Solution

Virtual acoustic recording refers to the capture of real-world acoustic performances in reverberant spaces and their subsequent plausible reproduction in a virtual version of the original performance space, otherwise known as a Virtual Auditory Environment (VAE). Such a walk-through auralisation still presents several challenges for production engineers, the most significant of which is the generation of the correct room acoustic response due to a given sound source-listener position. In particular, the correct direction of arrival of the direct sound and early reflections must be maintained since these signals contain the most vital cues for localisation of acoustic sources. From the point of view of the perceived distance, it is also crucial for the amplitude of the direct sound as well as the direct-to-reverberant energy ratio to follow the changes that would have naturally occurred under these particular acoustic conditions. This simplified model can be used as a guideline or a starting point in the process of the development of a real-time audio engine.

In the ideal scenario, a full-length RIR should be obtained for each of the source-listener combinations in the recreated space at runtime. Since the real-world impulse response measurements can easily develop very long filter kernels, their realisation by the means of time-domain FIR filters can be computationally prohibitively expensive. On the other hand, frequency-domain block-convolution in the frequency domain naturally induces latency that is proportional to the block size [157]. Some methods of improvement have been suggested that e.g. parallelise the

problem and divide the frequency domain blocks into smaller and non-uniform sub-blocks [158]. This approach can provide a trade-off between the processing burden and the system latency. However, for very long filter kernels like the RIRs of halls or cathedrals, we shall still seek a more efficient solution.

Pellegrini [172] proposed the hybrid model for auralisation which consists of: 1) A model-based early part, including direct sound and first- and second-order reflections obtained using geometrical approach; 2) A data-based second part, including all reflections of order three to nine, with a delay up to 80 ms again based on a geometric model; 3) The later part which is taken of a measured RIR.

In this thesis it is proposed to modify and expand the above approach in order to utilise computational acoustics aided with a single B-Format Spatial Room Impulse Response in the process of auralisation. For this reason, the real-time auralisation engine has been designed and created based on a hybrid model of the RIR. The idea was to decompose each RIR into two parts: the short, deterministic part (direct sound + early reflections) that can be synthesised at runtime, and the long, diffuse reverberation tail, that can be pre-processed with the source audio off-line. The deterministic part could be computed using one of the efficient geometrical approaches whereas the reverberation tail can be either obtained from the real SRIR measurements or also synthesised e.g. using more accurate acoustic methods (FEM, BEM, FDTD etc.).

The diffuse reverberation results from complex interactions between the emitted sound waves and the environment including physical phenomena such as scattering and diffraction. However, from the point of view of perception the diffuse reverberation lacks any prominent direction. It is also homogeneous across different points in the acoustic space.

The above knowledge can be utilised in order to parametrise the B-Format SRIR so that in each of the sound field components (W, X, Y and Z) the directional and diffuse components are identified. For example, the method of Directional Analysis by Merimaa and Pulkki [144] can be used to identify and subsequently attenuate the directional components from the early part of the diffuse-field SRIR. Diffuse field SRIRs are good candidates since the energy ratio between the directional and diffuse components is already low. The algorithm for the Directional Analysis has already been presented in Section 3.4.5 and more details can be found in literature [185]. Here the extraction is done by multiplying each time-frequency block of the B-Format signals (W, X, Y and Z) by the diffuseness estimate $\sqrt{\psi_i}$ where i signifies the i-th frequency band.

In the below example, a full SRIR has been processed in order to achieve a truly diffuse response. The SRIR used was measured in the large cathedral 32m from the sound source using a *Soundfield* microphone (see later Section 4.4.2.1 for more details).

As explained by Pulkki [185], different SRIRs may require different parameter values in the analysis in order to come up with optimal results. So far, no evaluation method of the effectiveness of the directional analysis has been proposed but it is suggested that the resultant SRIR can verified by the means of auditioning:

> So far all DirAC parameter values, such as the lengths of time windows for temporal averaging and the parameters for timefrequency analysis have been defined by informal listening during the development. The possibility of using some more advanced methods, such as formal listening tests or auditory modelling to find optimal parameter values, is reserved for future work.

Vilkamo in [223] gives an overview of directional analysis parameters, their influence on the analysis output as well as possible audible artefacts. He suggests the choice of parameters to best match the integration in human hearing e.g.

100Hz	200Hz	300Hz	400Hz	510Hz	630Hz	770Hz	920Hz	1080Hz	1270Hz
200ms	200ms	200ms	175ms	137.3ms	111.11ms	90.9ms	76.1ms	64.8ms	55.1ms

1480Hz	1720Hz	2000Hz	2320Hz	2700Hz	3150Hz	3700Hz	4400Hz	5300Hz	6400Hz
47.3ms	40.7ms	35ms	30.2ms	25.9ms	22.22ms	18.9ms	15.9ms	13.2ms	10.9ms

7700Hz	9500Hz	12kHz	15.5kHz	20kHz
9.1 ms	7.4 ms	5.83ms	4.52ms	3.5 ms

Table 4.1: Averaging window lengths used to compute the diffusion estimates at different frequency bands

The resultant full W component of a SRIR along with the frequency-averaged diffuseness estimate over time is presented in Figure 4.1. A good indication of the successful process of directional components extraction could be that the diffuseness estimate is low in the early part of the RIR and grows afterwards.

The early part of the SRIR is obtained at runtime using Image Source Method. Each direct and mirror-image source is subsequently spatialised using Higher Order Ambisonics. On the other hand, the diffuse B-Format SRIR is processed off-line and pre-convolved with the with the programme material. Then, at runtime it is mixed with corresponding W, X, Y and Z channels containing the directional sources from the simulation.

The advantage of the method proposed is that the complexity of late source-room interactions can be almost fully retained in the recreated virtual auditory event with the minimised computational effort. This is because the diffuse reverberation tail can be applied to the source material in the off-line convolution process and thus saving the run-time processing power. Precomputing the reverberation (using geometrical acoustics) has already been proposed e.g. by Tsingos [222].

It must be noted though that the assumption is made that a single B-Format SRIR can be obtained or approximated for the space being auralised. However, this assumption is reason-

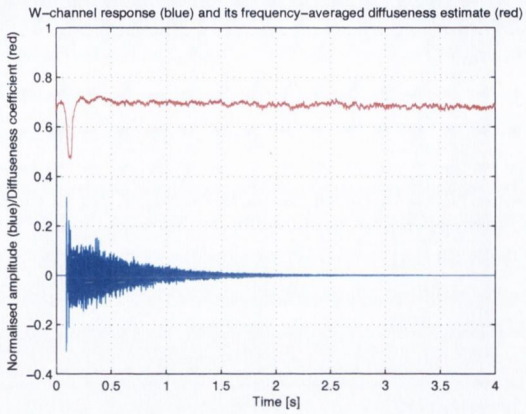


Figure 4.1: W component of the B-Format SRIR measurement together with its analysis of diffuseness - the frequency averaged diffuseness estimate.

able especially with the growing number of freely available B-Format SRIRs, like Openair[19], or simplified measurement techniques [58] which approximate B-Format SRIRs using multiple unidirectional microphone measurements.

Next two sections describe two applications which utilise the above approach and give more detail about its practical implementation.

4.3 Application Example I: Traditional Irish Trio Performance

4.3.1 Introduction

This section presents the preliminary approach to implement the proposed auralisation method to the virtualisation of a real-world musical performance.

4.3.2 Data Acquisition

In order for a real-time walk-through presentation to be created, a music trio (guitar, vocal and violin) performing a traditional Irish song have been recorded. The performance took place in the medium-size, reverberant Printing House Hall in the Department of Electronic and Electrical Engineering in the Trinity College Dublin. This hall is presented in Figure 4.2 and its main geometrical dimensions are gathered in Table 4.2.

This space can be regarded as a typical shoe-box type room. Spatially averaged reverberation time in the hall at $1kHz$ is around $1s$ and the critical distance (calculated using the Equation 2.10 is approximately $1.1m$. More detailed acoustical analysis of the hall, including the $ISO - 3382$ measurements [98] and geometrical simulations can be found in Appendix A of [105].

[19] `http://www.openairlib.net/`

Figure 4.2: Printing House Hall in Trinity College Dublin

Printing House Hall Dimensions:	
Main room dimensions ($W \times L \times H$):	$5.85m \times 15.56m \times 4.14m$
Overall surface area (S):	approx. $359m^2$
Overall volume (V):	approx. $376m^3$

Table 4.2: Principal dimensions of the Printing House Hall in the Trinity College Dublin

During the performance, each musician was recorded in the direct field with the use of a dedicated unidirectional (supercardioid) microphone. Microphones used for the direct sound capture were *AKG C-414B* (guitar and vocal) and *Rode NT2000* (violin). Reference stereophonic recordings (XY) have also been made from the distance of 1, 2 and 4m measured from the centre of the "stage". The performance was also filmed using a camcorder. The filming was done both in motion and using a tripod at a fixed position. In order for the further comparisons to be made, the *Soundfield* microphone was attached with the camcorder to the same boom arm or tripod at all times.

Lastly, the SRIR was taken using the *Soudfield MKV* system and the exponential swept tone technique. *Genelec 1029A* bi-amplified loudspeaker system was utilised for acoustic excitation. Source position was set up at the centre of the stage (vocal position) and the receiver was situated well within the diffuse field of the hall at the $8m$ distance.

4.3.3 Real-Time Visualisation

From the manually measured geometrical data, an *AutoCAD* model of the hall had been created and then imported into *Blender 3D* environment [28]. *Blender* is an open source, fully integrated 3-D modelling, animation and Computer Graphics rendering environment with a built-in game engine functionality (*Blender Game Engine - BGE*). This software also has a full support for the *Python* scripting language [187], which greatly extends its scope of possible applications

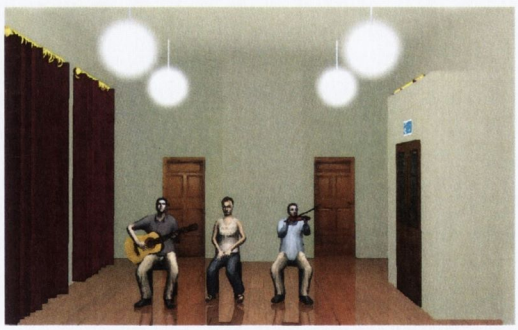
and allows, for example, for real-time communication with other software packages like *Pure Data* [181]. *Pure Data* (abbreviated later as *pd*) is visual programming environment designed mainly for real-time audio and multimedia signal processing.

The dimensions of the hall were faithfully transferred to the 3-D model and environmental units were set to match their magnitudes in the real-world metric units. In this way, virtual sound sources could be re-instated in the virtual space at locations matching the real-world co-ordinates. The process was very thorough and was compared with the real camera footage that was also taken during the performance.

In order to make the comparison feasible, virtual camera settings in *Blender* had to be adjusted to match the settings of the camera used for filming. In order to give the model a more visually pleasing look, simple texturing and lighting was applied. Also, static avatars of musicians were created and situated in the virtual world. Although animation would certainly bring another level of realism to the scene, it was decided that it would depart from the main objectives of this work and was left as a future activity. The original image, the 3-D world superimposed on the camera footage and the resultant 3-D environment can be seen in Figure 4.3.

(a)

(b)



(c)

Figure 4.3: (a) The original image captured during the performance; (b) 3-D world superimposed on the camera footage; (c) Final real-time render of the 3-D environment

4.3.3.1 Adding Interactivity and Integration with the *Pure Data*

Blender 3D was used not only for the visual representation of the virtual environment but also for its built in game engine functionality (BGE). Adding interactivity to the complete 3-D model comprised of placing a virtual camera in the environment and making it respond to the PC controllers: mouse and keyboard input. A simple *Python* script was written to assign mouse movement to the camera orientation and keyboard input to the camera translation, which are typical settings in many video games. The script allowed the user to move freely and explore the interior of the hall model.

As mentioned before, avatars of the musicians do not change their locations throughout the whole experience. Parameters that do change in real-time are: the location and orientation of the virtual camera in the virtual world (representing the listener) and horizontal angles between the camera and each of the virtual sound sources. From the auralisation point of view, whenever the camera's location/orientation shifts, the listening perspective alters as well and incurs the need to update the sound field in the audio engine. For this reason, one-way communication between *Blender* and *pd* has been established that allowed the communication of the user's location/orientation and source angle data at runtime. The communication was done by utilising the UDP network protocol and was governed by the custom-made *Python* script on the *Blender 3D* side and by the built-in object *netreceive* on the *pd* side. The data update rate matched the internal *Blender's* logic clock that by default is set to $60Hz$.

4.3.4 Real-Time Auralisation

Having created the interactive Virtual Visual Environment (VVE) of the Printing House Hall, we shall now consider the creation of the virtual acoustic model of the space, herein termed the Virtual Auditory Environment (VAE). The objective of this work was the presentation of aurally plausible reproduction of the traditional Irish performance in a virtual version of the hall. As we have already identified, an important aspect in the quest for realism in such auditory scene synthesis is user interaction. That is, how the movements of a person listening to the virtual auditory scene directly influence the presentation. In the former section, we have presented the methodology of converting a graphical 3-D model of an existing space into a virtual world that user can interact with. We shall now focus on explaining how user-driven data could be utilised in order to inform the audio engine to re-render the current sound field to reflect the current visual situation.

The audio engine used for auralisation was fully implemented in *pd*. On the *pd* side, the direct sound and early reflections were encoded into 3^{rd} order Ambisonics. Additionally, for each source a directivity filter was applied. A monophonic mixture of the direct field recordings, convolved with the first order diffuse field SRIR residual was added to the mix. The final decode was created for 8-loudspeaker horizontal only setup. The binaural rendering was implemented using the optimised virtual loudspeaker approach as outlined in Section 3.38. Figure 4.4 shows

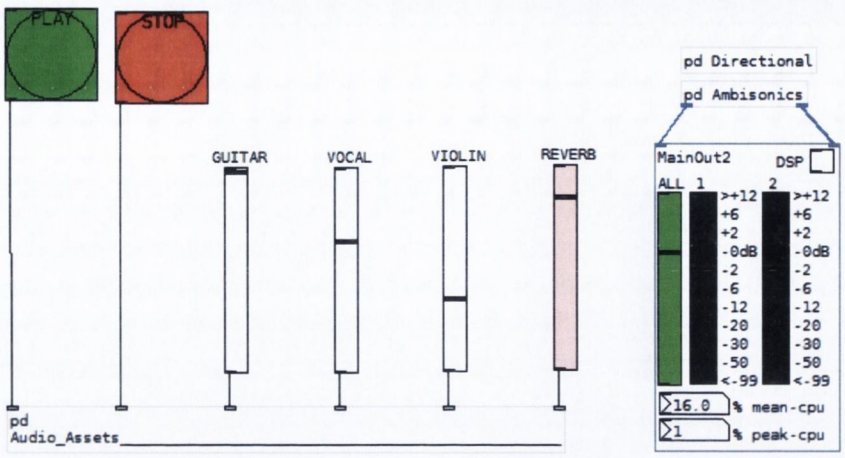


Figure 4.4: *Main_GUI* abstraction which is a top-level GUI for the implemented audio engine

the top-level *pd* Graphical User Interface (GUI) abstraction which was designed to hide the complexity of the above processes and allow for easy playback control.

The big green (Play) and red (Stop) button trigger and stop the playback of all the sources in the real-time auralisation. Four Sliders (Guitar, Vocal, Violin and Reverb) allow for the level control of individual sound sources and global reverberation. Following a typical Digital Audio Workstation convention, it was decided that it would be a good practice if the level of reverberation can be adjusted so that wet/dry signal ratio can be set to achieve the desired perceptual result. The output of the two buttons and four sliders is fed into the *Audio_Assets* abstraction which is responsible for loading all the audio assets (e.g. dry O-Format source signals (3^{rd} order) and dry, monophonic source signals pre-convolved with the 1^{st} order diffuse reverberation tail), decoding of the O-Format sources into single mono streams and summing together corresponding W, X and Y reverberation channels.

The decoded dry source signals are subsequently sent using individual delay lines to the *Directional* abstraction which is responsible for generating early reflections and spatialisation of the directional real and image sources. 3^{rd} order sound field representation of the directional components is then routed to the *Ambisonic* module which performs sound field rotation (based on current listener's orientation in *Blender* visual environment and also based on the information received from a head-tracking device), decode and binaural down mix. At the input stage of this abstraction the diffuse sound field components (W, X and Y) are also added to the sound field.

Finally, the *MainOut2* object has been created to help visualise and control master output levels of the binaural streams, mean and peak CPU usage and to switch the DSP on and off.

The following sections explain implementation of each of the above steps in a deeper level.

4.3.4.1 Direct sound and early reflections

In this work, the Image Source Method has been chosen as the efficient method to synthesise early reflections at runtime. For each sound source a map of image sources has been created using the approach proposed by McGovern [138]. The method allows for fast calculation of image sources that are stored in square matrices (3×3, 5×5 etc. for 2-D case or $3 \times 3 \times 3$, $5 \times 5 \times 5$ etc. for 3-D case) where the number of elements in the matrix indicates the total number of direct and image sources. If the virtual sources do not change their locations at runtime it is possible to pre-calculate the image sources and store them in lookup tables.

Here we are using a 2-D scenario in which at runtime each sound source source is represented as a direct sound and 8 reflections (3×3 matrix). This is realised using multi-tap variable delay networks which was first suggested by Schroeder [202] and implemented by Moorer [154]. In this way the computationally expensive FIR filter implementation via convolution is avoided, but more importantly it is possible to represent each reflection as a discrete sound source. Delay times are recalculated based on the actual distance from the receiver to the direct or mirrored sound source. Low-pass filtering (smoothing) is applied to delay curves in order to avoid the zipper noise. Intensity values of each of the sources obey the inverse square law in order to mimic the natural decay with the distance.

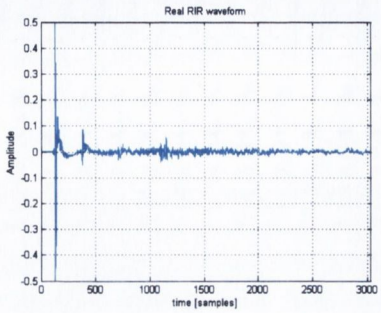
In order to simulate energy absorption by the boundaries of the hall, the magnitude of each reflection is scaled by the reflection coefficient β which expresses the average surface reflection properties of the modelled space. First, based on the hall dimensions presented in the Table 4.2 as well as the spatially averaged reverberation time at $1000Hz$, the average absorption coefficient α has been calculated using the empirically derived Sabine equation for the reverberation time in enclosures:

$$RT_{60} = \frac{0.161V}{S\alpha} \tag{4.1}$$

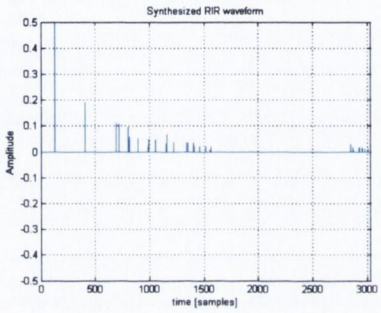
where RT_{60} is the spatially averaged reverberation time at $1000Hz$ in seconds, V is the approximated volume of the room in m^3, and S is the total surface area in m^2. Substituting $V \approx 376m^3$, $S \approx 359m^2$ and $RT_{60} \approx 1s$ and solving for α yields $\alpha \approx 0.17$. Then, the average surface reflection coefficient β can be calculated as $\beta \approx \sqrt{1-\alpha^2} = 0.985$.

Lehmann [120] has shown how phase inversion of image sources can lead to more realistic impulse responses, since a given RIR usually displays stochastic noise-like properties around a zero mean. Thus, it is proposed to use negative absorption coefficients as a first approximation to the directional portion of RIR. The improvement in RIR approximation over positive only impulse responses in comparison to real-world measurements in the case of the modelled hall is shown in Figure 4.5.

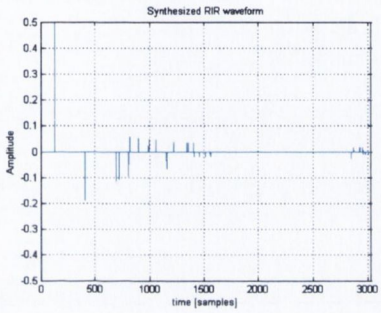
In *pd*, the *Directional* abstraction is responsible for performing all the above processes.



(a)



(b)



(c)

Figure 4.5: Room Impulse Responses: (a) Measured in the hall at $8m$ distance from the source; (b) Synthesised, using Image Source Method; (c) Synthesised, using Image Source Method, with negative reflection coefficient β

This abstraction is shown in Figure 4.6. *Directional* contains pairs of objects that represent directional sound sources used in the simulation. First of these objects - *EarlyReflectionsGen* - generates at runtime 2-D matrices which contain the information about directions, amplitudes and delays of the direct sound and early reflections. It accepts two arguments which are x and y co-ordinates of the current sound source. The output of the *EarlyReflectionsGen* is then sent to *EarlyReflectionsApp* which performs the 3^{rd} order Ambisonics spatialisation of the direct sound and early reflections. In the *EarlyReflectionsApp*, first of the arguments is the delay line name which contains the dry source sound. The second, optional argument is the *netreceive* UDP port number which is used to provide the information about the visibility of the sound source (not used in this example, however useful when more complex room geometries are used. Please, see Section 4.4.4).

Inside the *EarlyReflectionsGen* object (Figure 4.7) angle, delay time and gain coefficients are computed for the direct sound and early reflections. In this example, the location of a sound source is expected not to change. That is why, locations of mirror images can be pre-computed (based on the specified room dimensions) and stored in lookup tables. However, extending the current abstraction to allow for the source movement should be a relatively easy task and would

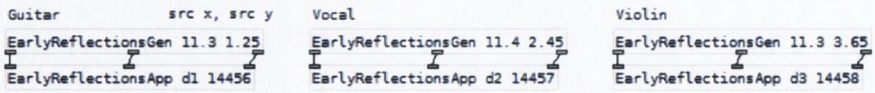


Figure 4.6: *Directional* object which is responsible for early reflection generation and Ambisonics spatialisation of all the directional sound sources in the auralisation.

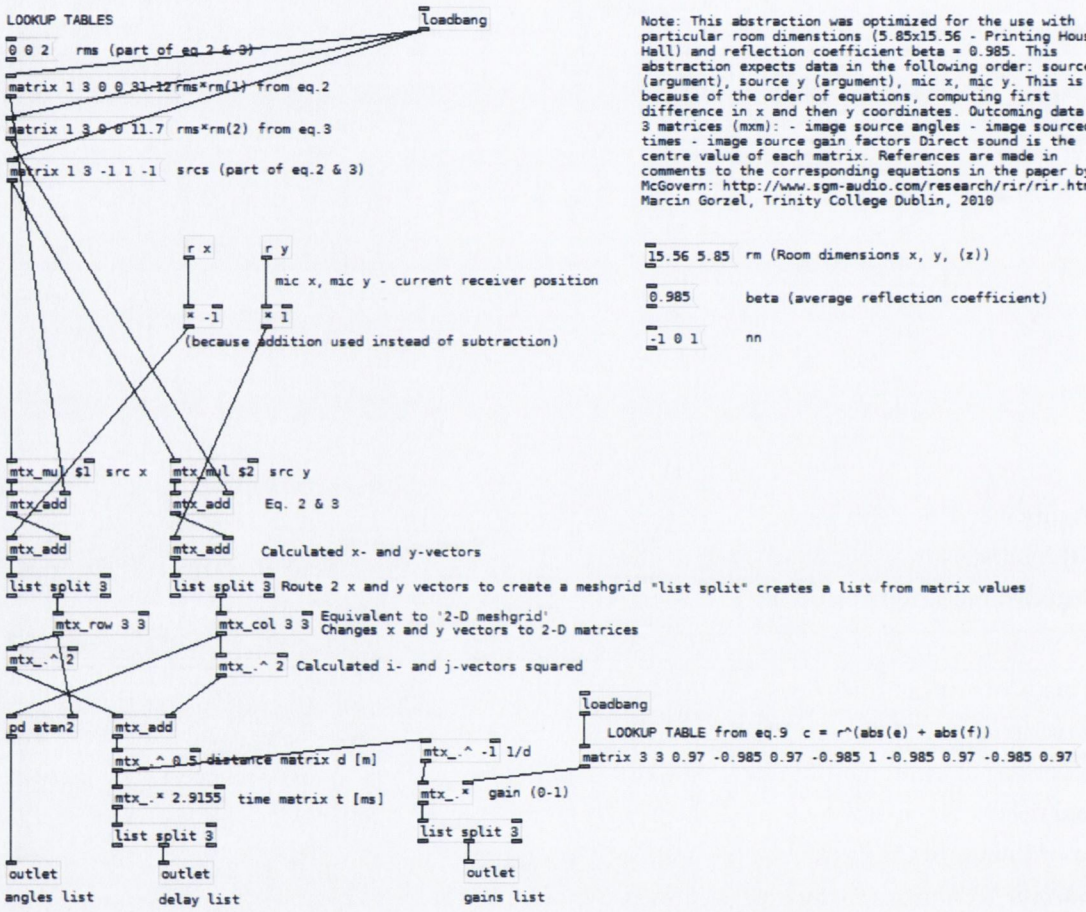


Figure 4.7: *EarlyReflectionsGen* object which calculates angle, delay and gain matrices for the direct sound and early reflections.

involve adding receivers of the source x and y co-ordinates to the *mtx_mul $1* and *mtx_mul $2* objects.

4.3.4.2 Sound Spatialisation

Reproduction of the auditory scene is highly dependent on the spatialisation method employed. VBAP is capable of producing very sharp acoustic images, especially if dense arrays of loud-

speakers are used for the reproduction and if the virtual sound source coincides with one of the reproducing loudspeakers. The localisation accuracy however is not uniform and drops at between loudspeaker locations. Wave Field Synthesis also necessitates a high count of reproducing channels and is rather better suited for large area faithful sound field reconstruction. On the other hand, HOA provides a scalable and modular method for sound field reproduction. It means that the sound field can be rendered with arbitrary accuracy and the encoding/transformation stages are usually independent on the final decoding stage (e.g. loudspeaker reproduction array used). Due to the fact that in general all the loudspeakers contribute to the sound field reproduction at all times, HOA also does not suffer from the colouration to the same extent as amplitude panned source [105]. Thus, this method is suitable for rendering dynamically changing auditory scenes as in the case of real-time auralisations.

When, dealing with the discrete representations of sound sources and their mirrored images, the process of encoding the auditory scene into its HOA sound field representation is rather straightforward. This is because each delayed version of the direct signal can be treated independently (i.e. as it was a *new* sound source). Then, all the respective spherical harmonic components for all the original and delayed signals can be summed up together accordingly.

Therefore, in the next order the audio engine encodes all the streams (from direct and image sound sources) into 3^{rd} order Ambisonics, taking into account the angle they make with the receiver - virtual camera (*angle* matrix is used for this task). The *EarlyReflectionsApp* (Figure 4.8) object accepts the output from the *EarlyReflectionsGen* object in order to perform spatialisation of the direct sound and early reflections. Here, the audio stream from a given delay line (e.g. *d1*) is demultiplexed into nine streams in order to represent the direct sound and early reflections. Each stream is individually delayed with the use of the *pd*'s variable delay object *vd~*. The input of delay time coefficients is however low-pass filtered in order to avoid the zipper noise (*lop~ 2* object). In the next order, each stream is scaled by the individual gain factor and (optionally) multiplied by 1 or 0 according to the visibility flag received from *Blender*.

3^{rd} order Ambisonics encoding is done using *pd*'s *ambi_encode* object from the *iem_ambi* library[20].

By combining the corresponding Ambisonic signals from all the sources and their mirror images the spherical harmonic representation of the whole auditory scene is created. This is achieved with the *throw~* objects and specifying the Ambisonic channel. At this stage the B-Format reverberation tail is also added to the sound field.

4.3.4.3 Sound Field Rotation and Decoding

The last abstraction *Ambisonics* (Figure 4.9) deals with sound field rotation and 2-stage decoding of the sound field signals (first into virtual loudspeaker feeds, and then into binaural streams).

The HOA B-Format signals are captured by the *pd*'s *catch~* objects which are located inside

[20] `http://puredata.info/downloads/iem_ambi`

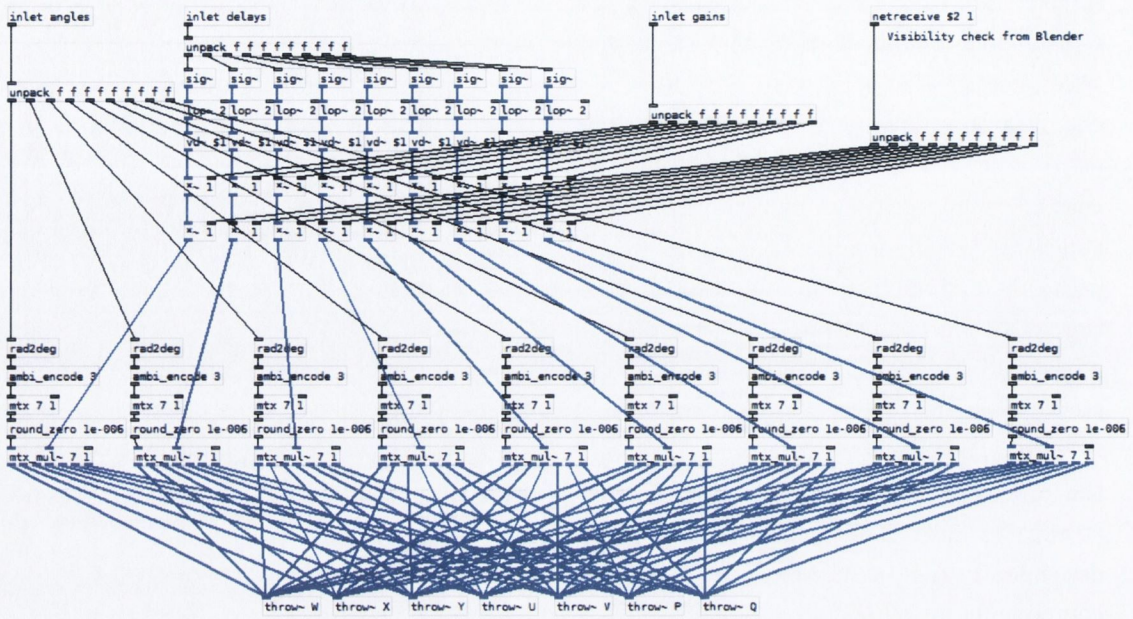


Figure 4.8: *EarlyReflectionsApp* generates audio streams for the direct sound and early reflections and performs 3^{rd} order Ambisonics spatialisation.

the *3rdOrderAMBRotationMatrix* object (Figure 4.10). The received HOA signals are fed into the rotation matrix that is constantly awaiting the data describing the listener's orientation. The orientation data can come from the mouse controller (if mouse is used in order to change the viewpoint) or the head-tracking device or both. In the last case, the transformation matrix accounts for the resultant listener's orientation being a sum of the virtual camera orientation and the head orientation. Using the Ambisonics approach here means that instead of recalculating the angle of arrival for every sound source and every reflection it is much easier to rotate the whole sound field according to the incoming orientation data. The mechanisms of HOA sound field rotations has already been introduced in this thesis in Chapter 3. In *pd*, *ambi_rot 3* object is used to perform the rotation against the z (vertical) axis (2-D scenario).

In the last stage, decoding coefficients are computed for the chosen loudspeaker configuration. In this example, the reproduction setup is by default the regular octagon (horizontal only). However, the systems is by no means restricted to this setup but using different layouts would require new 3-D encoding and decoding coefficients to be computed for correct reproduction. The decoding coefficients are computed for the chosen loudspeaker configuration using *pd*'s *ambi_decode3* object from the *iem_ambi* library. The coefficients are stored in a 8 by 7 matrix which is used to multiply the HOA sound field signals.

Binaural rendering has been also implemented as the "virtual loudspeakers" and as outlined in Section 3.5.1. In order to avoid the latencies resulting from multiple HRIR filtering the HRIR

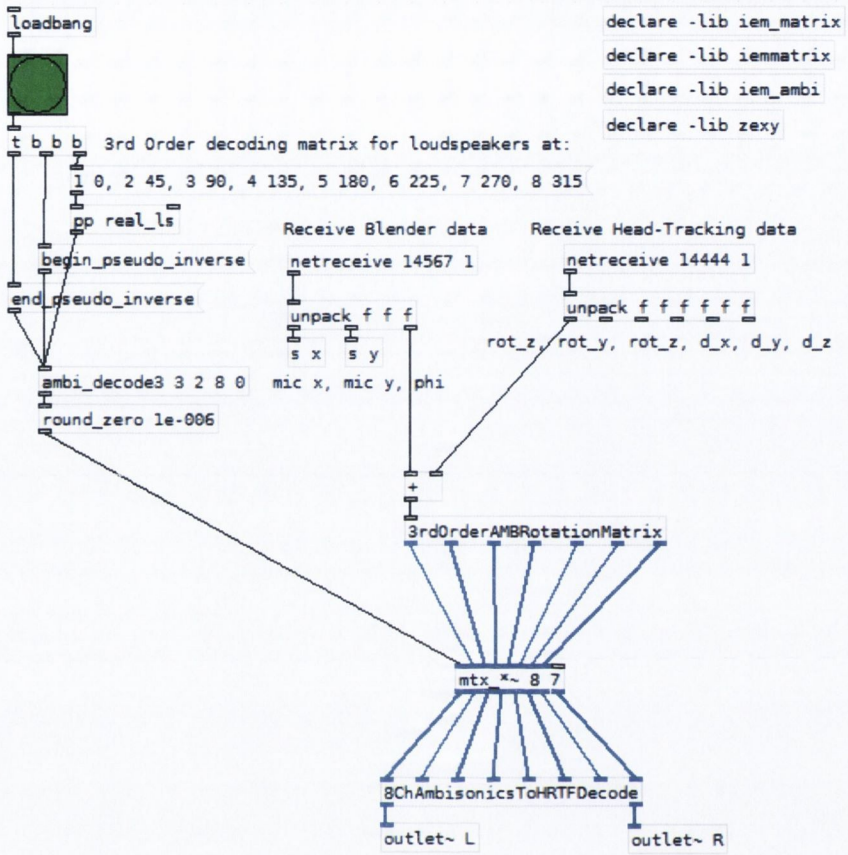


Figure 4.9: *Ambisonics* abstraction which deals with sound field rotation and binaural decoding of the HOA sound field components.

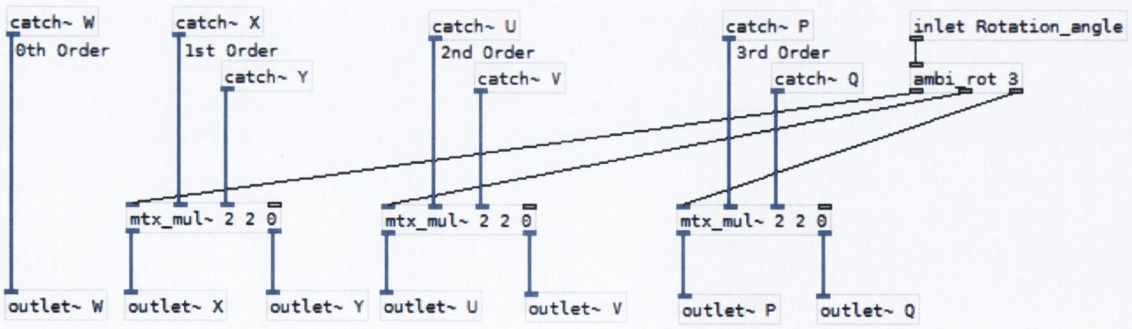


Figure 4.10: *3rdOrderAMBRotationMatrix* object which performs rotation of the HOA sound field components based on current listener's orientation in the virtual world.

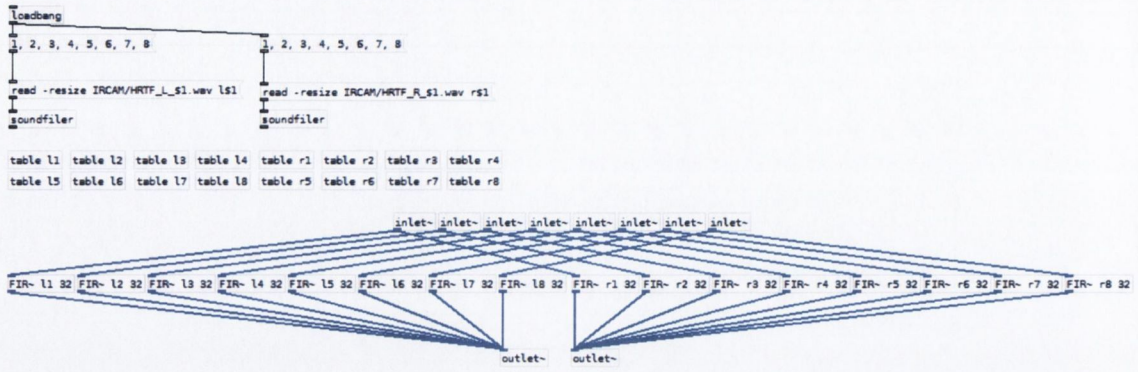


Figure 4.11: *8ChAmbisonicsToHRTFDecode* object which performs the binaural down mix of the virtual loudspeaker signals.

order reduction by approximate factorisation has been utilised as discussed in Chapter 3 Section 3.5.2. The optimisation allowed for significantly shorter HRIR run-time filters of 32-taps in the default configuration based on the HRIR filters pooled from the *LISTEN* database [97] (originally 512-taps long).

The *8ChAmbisonicsToHRTFDecode* (Figure 4.9) performs the binaural down mix of all 8 virtual loudspeaker signals. First, the HRIRs are loaded into memory (*tables*) with the use of *soundfiler* objects. Then, each loudspeaker signal is passed through 8 pairs of FIR filters which kernels are the left and right HRIR functions for the virtual loudspeakers at 8 horizontal locations.

4.3.4.4 Adding Sound Source Directivity

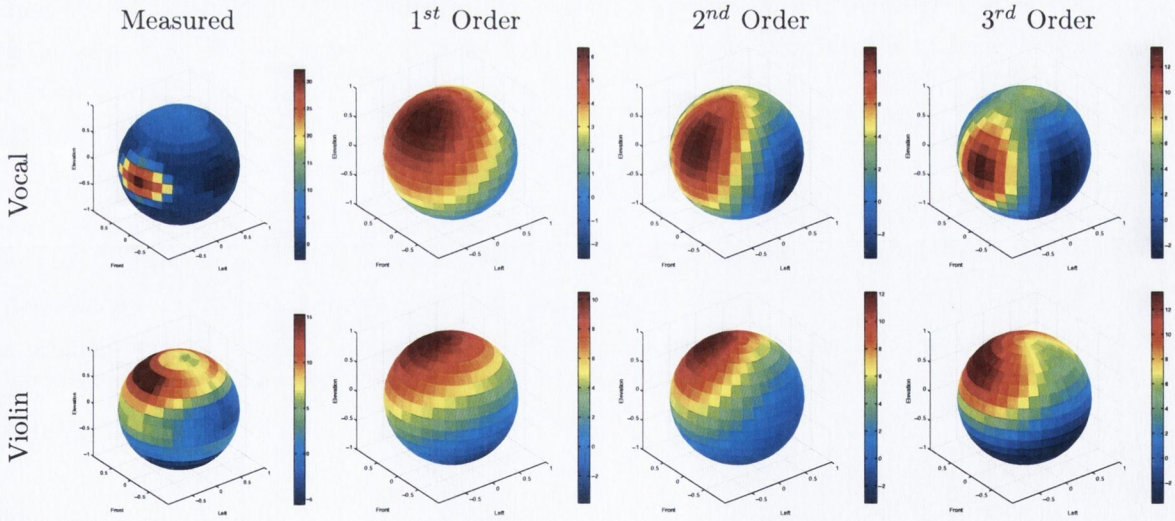
So far we have described the full signal processing path of the proposed auralisation system assuming that each sound source can be represented by a single dry or anechoic recording. Such approach creates a sound source whose directional output is not dependant on the listening angle in the virtual world. However, for any virtual acoustic recording to be convincing, the directional properties of the source audio and the subsequent effect on the early reflections must also be considered.

In this work the decomposition of the directional response into spherical harmonics is used, which has been proposed for computational based auralisation in numerous papers most notably by Menzies [142] and Ahrens [3]. Here, for a given musical instrument, an averaged frequency dependent source directivity measured in an anechoic environment is encoded into HOA, forming a HOA directional filter. Such sources are also commonly referred to as O-Format Ambisonics as long as the 1^{st} order decomposition is concerned [141]. The recorded direct sound can be pre-rendered before runtime as a directional HOA source. The frequency response of the source audio utilised for each early reflection is therefore dependent on the angle of incidence of that

reflection. An important aspect of filtering direct source audio with a HOA directional filter is de-emphasis of the directional responses. Calculation of the directional response functions must not assume that the magnitude spectrum of the recorded musical source audio is flat, and that application of directional filtering will yield the appropriate directional magnitude response. A reference source angle should in fact be taken (e.g. the angle of the direct field microphone to the instrument) and deviations from the recorded source magnitude response used to form the directional filters.

In this work the approximation of the sound source radiation patterns uses the method of approximating 2-D functions defined on the surface of the unity sphere, as described earlier in Chapter 3. For the vocal and violin sources the measured radiation responses in frequency bands were pulled from the Physikalisch Technische Bundesanstalt database [175]. Examples of approximations to directivity patterns of violin and human voice are presented in Table 4.3. Note that because of some energy concentration at the top of the sphere for the vocal response, the FOA approximation is slightly biased toward this direction. This situation is rectified as soon as HOA components are employed. The full set of frequency dependant radiation patterns used in this work to design directional filters can be found in Appendix C. In the 2-D implementation, the directional responses measured at the equators of the spheres were used. Due to lack of detailed measurements for the guitar, this instrument has been temporarily characterised with a frequency independent directionality of the first order cardioid.

Table 4.3: Vocal and violin radiation patterns at $1000Hz$ and their HOA approximations.



In this work, the decode of the O-Format source information from its HOA component channels is done using the *O-Format_Source* object (Figure 4.12). Inside this object, the monophonic source stream which is direction sensitive is created by the weighted sum of the B-Format channels. The weighting coefficients are the circular harmonics function coefficients. The argument

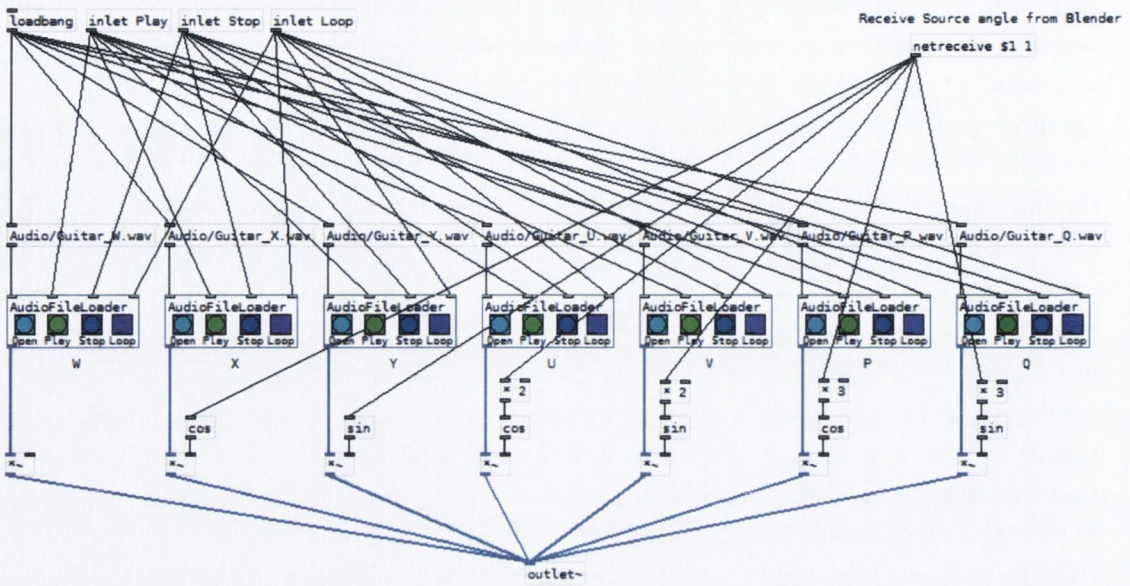


Figure 4.12: *O-Format_Source* object which performs the decode of the directional O-Format sound source to a single mono stream.

that is passed to these functions is the current horizontal angle between the source and the listener as obtained from the virtual world in (*Blender*).

One *O-Format_Source* object represents one sound source in the auralisation. These sources are located inside the *Audio_Assets* abstraction (Figure 4.13) along with the diffuse sound field components. This abstraction loads all the necessary sound files used in auralisation, performs the O-Format source decodes and routes the audio streams into relevant locations for further processing.

4.3.4.5 Note on Diffuse Field and Reverberation

As mentioned before, for the latter part of the RIR diffusion property of the measured RIR is used. Put another way, from the perceptual point of view the diffuse field is stochastic and this property holds at all locations in the modelled space. That is why, it should be possible to construct perceptually correct but also dynamically changed RIRs by re-computing only the early, directional part and complementing it with the diffuse residual.

A high-level diagram showing the complete signal processing chain of the implemented audio-visual rendering system can be viewed in Figure 4.14.

4.3.5 Summary

In this part we described the process of converting real-world acoustic performances into their virtualised versions with the use of signal processing and audio spatialisation techniques. The

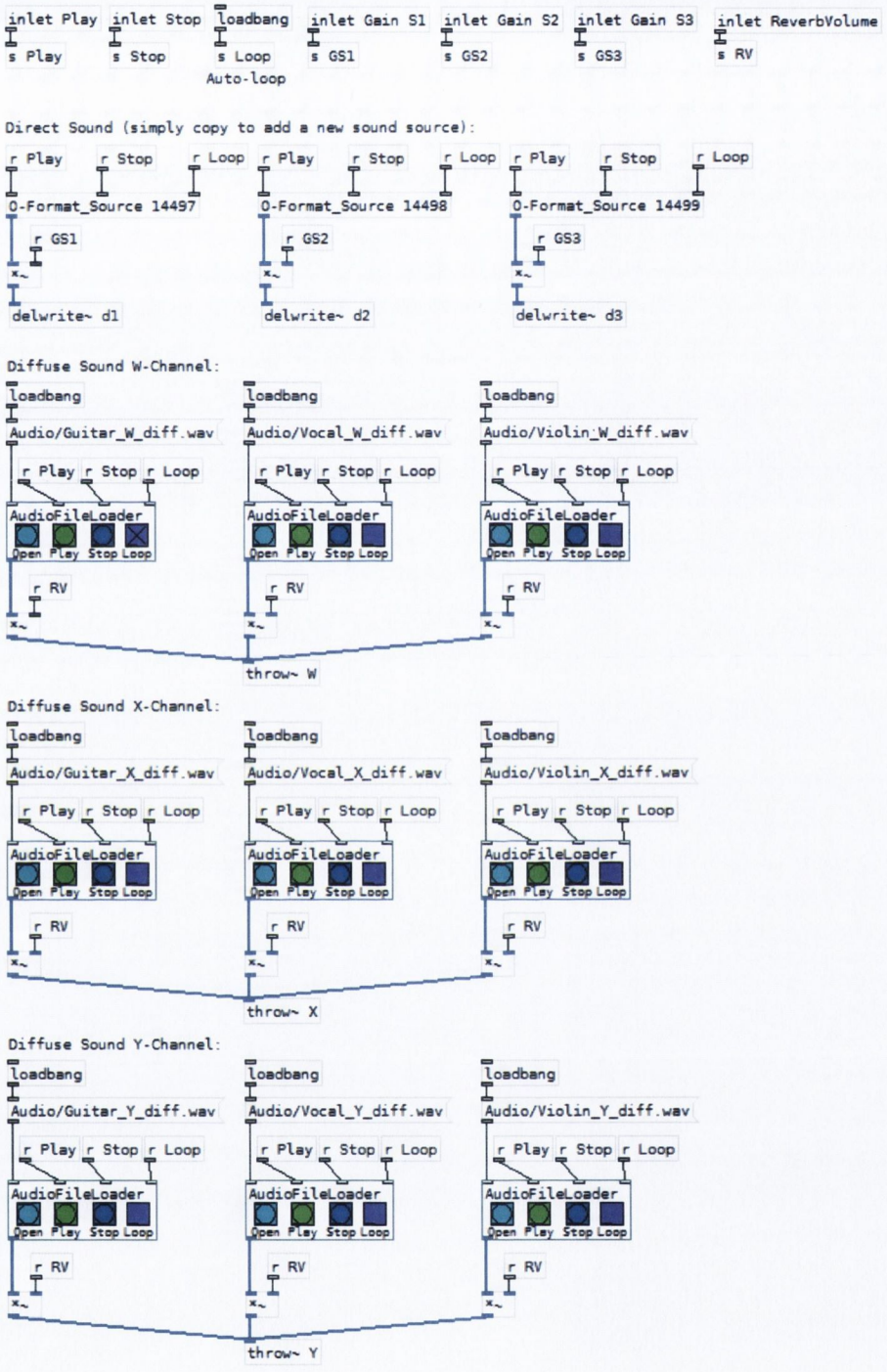


Figure 4.13: *Audio_Assets* abstraction which loads all the required sound files, performs the O-Format source decodes and routes the audio streams into relevant locations for further processing.

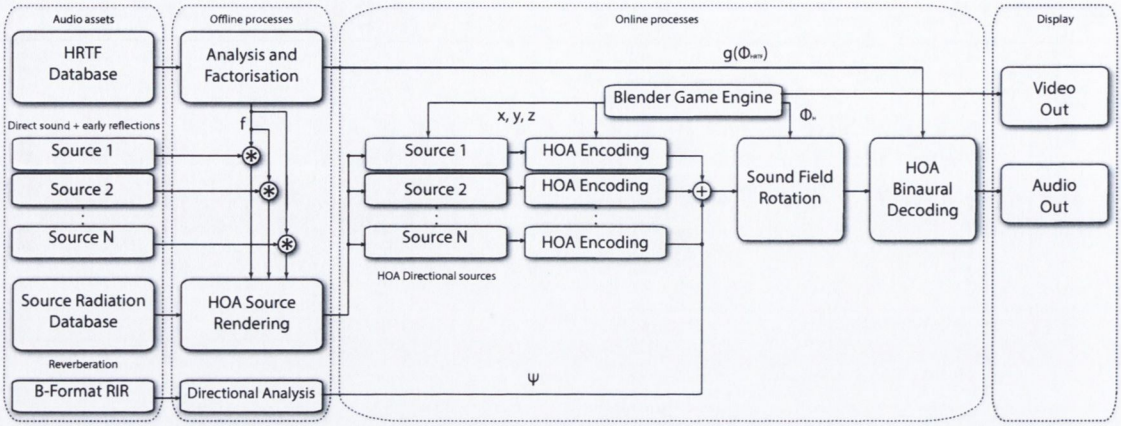


Figure 4.14: Complete flowchart of the implemented audio-visual rendering system. f represents the direction-independent or common component of the whole HRTF dataset. On contrary, $g(\Phi_{HRTF})$ represents direction-dependant residuals, specific to each loudspeaker location in the system. The direction-dependant components are used as run-time FIR filters, with kernels between 32 samples. Co-ordinates x, y and z denote the current location of the virtual camera in the modelled space. They are used in the HAO encoding stage to inform the encoder about the relative source-listener angle(s) ϕ_S and distance(s) ρ. Φ_H denotes the resultant virtual camera orientation with respect to the virtual-world axis corrected by the user's head orientation in the real-world co-ordinates. These variables are sent to the audio engine from *Blender Game Engine* in real time using the UDP protocol. Ψ denotes the 1^{st} order B-Format reverberation. Reverberation signals are obtained by offline convolution of the mono mix of all the real sources in the scene (image sources are disregarded) with the diffuse part of the RIR resulting from the Directional Analysis [144]. Finally, + denotes the process of summation of the corresponding B-Format channels.

model was based on real-world acoustic and geometric measurements of the real space. Hybrid reverberation is created, where early reflection synthesis is achieved using image source modelling, combined with the real-world diffuse field acoustic measurement. Before runtime however, the measured response has to be first analysed in terms of its directional and diffuse content, and only the diffuse components should be retained. The full response is then reconstructed at run-time by summing together the computed and measured parts. This algorithm is designed to give perceptually similar auditory impression as the full data-based approach, however without the need of using a dense grid of acoustic measurements. Also, because the long reverberation time is pre-convolved with the source audio before runtime, computational savings are expected that are proportional to the length of the RIR used.

However, although the algorithm optimises the data-based auralisation process, it is difficult to estimate the exact savings. This is because the standard convolution approach may use fast

convolution approach (e.g. FFT-based) to render binaural auditory scene using RIR and HRIR filters. In the proposed method, RIR computation is reduced only to computation of early reflections. However, interactive sound spatialisation taking into account both source's and listener's movements is a differentiating feature which cannot be implemented in a straightforward way in the standard data-based approach.

Informal listening tests with several subjects demonstrated a very good match between the real-world sound field recordings of the performance, and the interactive virtual walk-through model, although some colouration and tonal changes were unavoidable. This work was also presented to the wider public at several occasions, most notably at the First Irish Workshop on Music and Audio Signal Processing in National University of Ireland, Maynooth (January, 2010) where it was met with positive reception. The methodology and experience gained in this project also became a working bench for a more large-scale initiative that will be presented next.

4.4 Application Example II: Christ Church Cathedral Choir Performance

4.4.1 Introduction

Amongst Ireland's many acoustic treasures, Christ Church Cathedral, located in the heart of Dublin city, represents one of the most historically significant. The Cathedral has not only consistently played a prominent part in Dublin's civic as well as religious history but has also been noted throughout the ages for the calibre of its choir and music. Throughout the centuries, the landmarks such as the Christ Church Cathedral, tend to change and evolve in order to e.g. adapt to new roles and social needs or even because their constructions are tested by time, nature or political situation. In many cases it means that their former spirit associated with the unique acoustics is ultimately lost. It is quite remarkable, especially if we consider that early music, and particularly choral music, was often written with specific performance spaces in mind [36]. Christ Church Cathedral was no different in this regard. More information and historical background to the Christ Church Cathedral can be found e.g. in [106, 150, 213, 216].

A significant body of research has been undertaken in the measurement of the acoustics of performance spaces. These are mainly concerned with creating acoustical fingerprints of historically important spaces for posterity, most notably in Italy [60]. Current methodologies involve the acoustic measurement of performance spaces using *Soundfield* microphones and binaural mannequins [61]. Anechoic recordings of musical performances can then be filtered (convolved) with these measurements giving a plausible formation of an auditory scene in a virtual version of the acoustic environment as it exists today. This is referred to as "Data-based auralisation".

Despite the fact that much research has been undertaken into the architectural and musical history of this cathedral, prior to this work, its acoustics have not been studied, nor measured for posterity. This is in fact the case for many of Dublin's important acoustic splendours, including

St. Patrick's Cathedral, located only a short distance from Christ Church. The architectural mapping and measurement of the acoustics of such spaces is vital to Ireland's cultural heritage preservation.

In this work it is demonstrated how real-world acoustic and architectural measurements in combination with computational-based auralisation can be used to create a plausible, interactive and real-time walk-through experience, including the computer graphic visualisation, for the purpose of acoustical heritage preservation of Christ Church Cathedral. Section 4.4.2 outlines the approach to acoustic measurement and performance recording within the space for the purpose of the walk-through auralisation. Section 4.4.3 deals with visual aspects of the work e.g. how architectural scans of the cathedral can be utilised for interactive visualisation in a real-time gaming engine. Then, Section 4.4.4 discuses the practical and technical aspects of the auralisation and its implementation as the audio engine. This is done in terms of contrasts as compared to the methodology used in the previous auralisation work. Finally, Section 4.4.5 concludes the work performed and gives some details about the public demonstration of the model.

4.4.2 Data Acquisition

4.4.2.1 Acoustic Measurement

Hopefully at this stage we have shown that the historic significance of Christ Church fully justifies the requirement to measure its acoustics as they currently exist for cultural heritage preservation. To this end, acoustic measurements were taken for two source positions. The first source position was to the centre of the choir, the second was in front of the choir where large-scale choral events are also often staged. This was to provide a comprehensive assessment of both musical scenarios. The omni-directional dodecahedral loudspeaker *Brüel & Kjær Omnipower*, as specified by $ISO - 3382$ was used for acoustic excitation [98]. The receivers used consisted of a *Soundfield MKV* system and a *Neumann KU-100* binaural head as shown in Figure 4.15.

Measurements were taken along the length of the nave, and to either side of the pews within the nave at 1, 2, 4, 8, 16, 24 and $32m$ increments from each source (1 & 2). Each measurement position (with the exception of $32m$) was replicated for each source position. Both measurement arrangements for source positions 1 & 2 are shown in Figures 4.16 & 4.17 respectively. The additional height of staging was accounted for in determining the height of the source position. The excitation signal used was an exponential sine swept tone of duration 60 seconds, ensuring that any loudspeaker induced distortion in the resultant impulse responses was removed [59]. A detailed account of these measurements and results have been documented in [208][21].

In this work the particular attention has been paid to the two major acoustic parameters of Reverberation Time (RT) and Inter-Aural Cross Correlation (IACC). Definitions of both these

[21] The author was an active participant in the data acquisition process that was conducted by an inter-university group of academics and staff. This fact is officially acknowledged in the paper cited

Figure 4.15: Acoustic measurements taken using *Brüel & Kjær Omnipower* omni-directional source, *Soundfield MKV* microphone system and *Neumann KU-100* binaural microphone

parameters will be briefly recalled here: The reverberation time is a measure of the decay of an impulsive sound source in a reverberant room, defined as the time it takes for the impulse to fall by $60dB$ after the direct sound. IACC gives a measure of the similarity of the left and right ear signals. In general, the higher the value of the IACC, the narrower the perceived source width. If this parameter is measured within the first $80ms$ of the impulse response, the influence of the early reflections can be ascertained. This parameter is specifically relevant in the 500, 1000 and $2000Hz$ bands, where the wavelengths involved are comparable to the dimensions of the head, and an average IACC of these bands is known as the IACC_{E3} [166].

The reverberation time for source position 1 (centre choir position) is shown in Figure 4.18(a). The resultant spatially averaged reverberation time at $1kHz$ was measured as 3.2 seconds. It can be observed that from $8m$, the reverberation time does not change across all frequencies and the diffuse field properties are dominant. The changes in IACC for the same source-receiver positions are shown in Figure 4.18(b). In general, the IACC values beyond $1m$ are similar with noticeable changes around $500Hz$. The spatially averaged IACC_{E3} was measured as 0.5717.

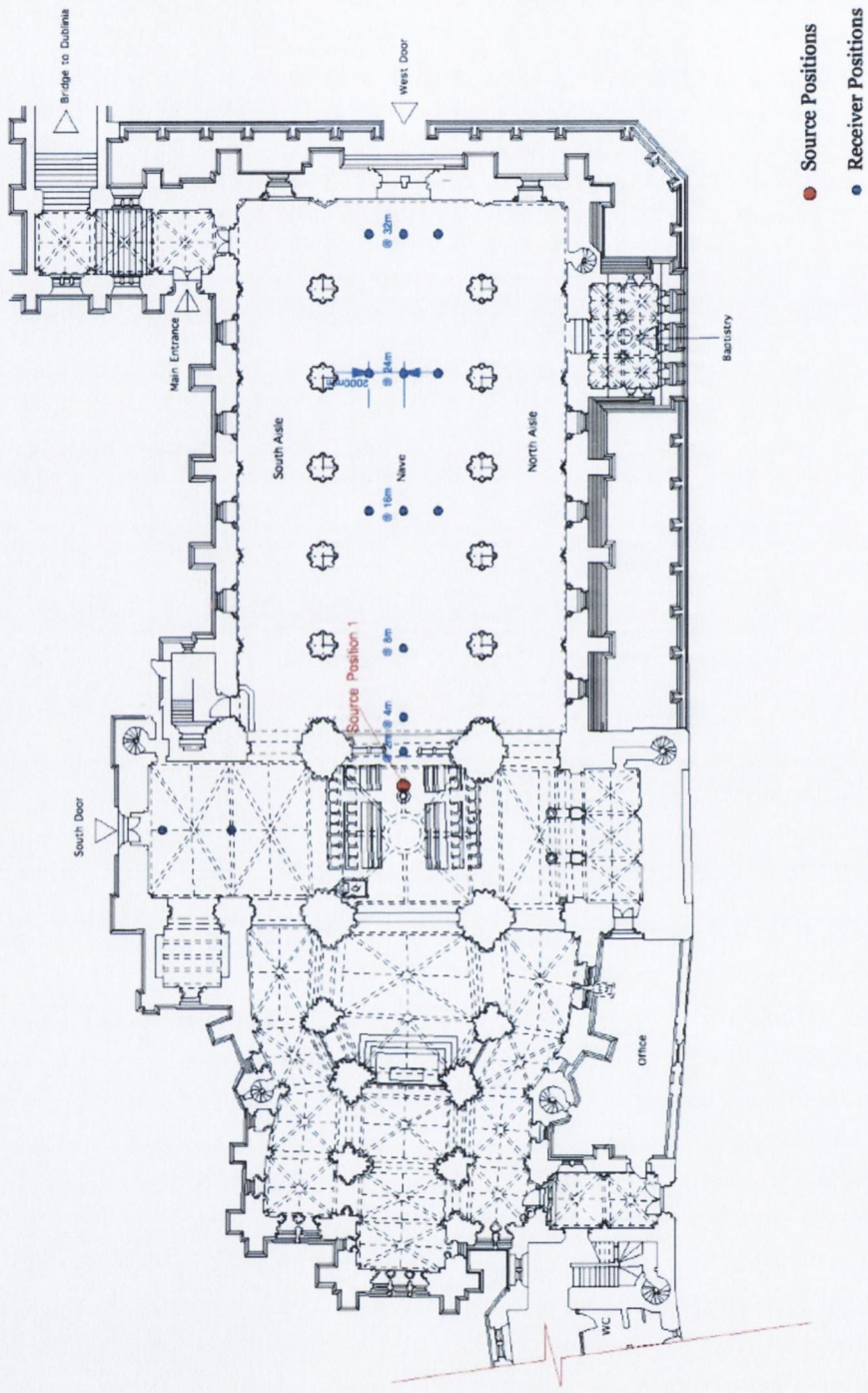


Figure 4.16: Acoustic measurement arrangements for source position 1

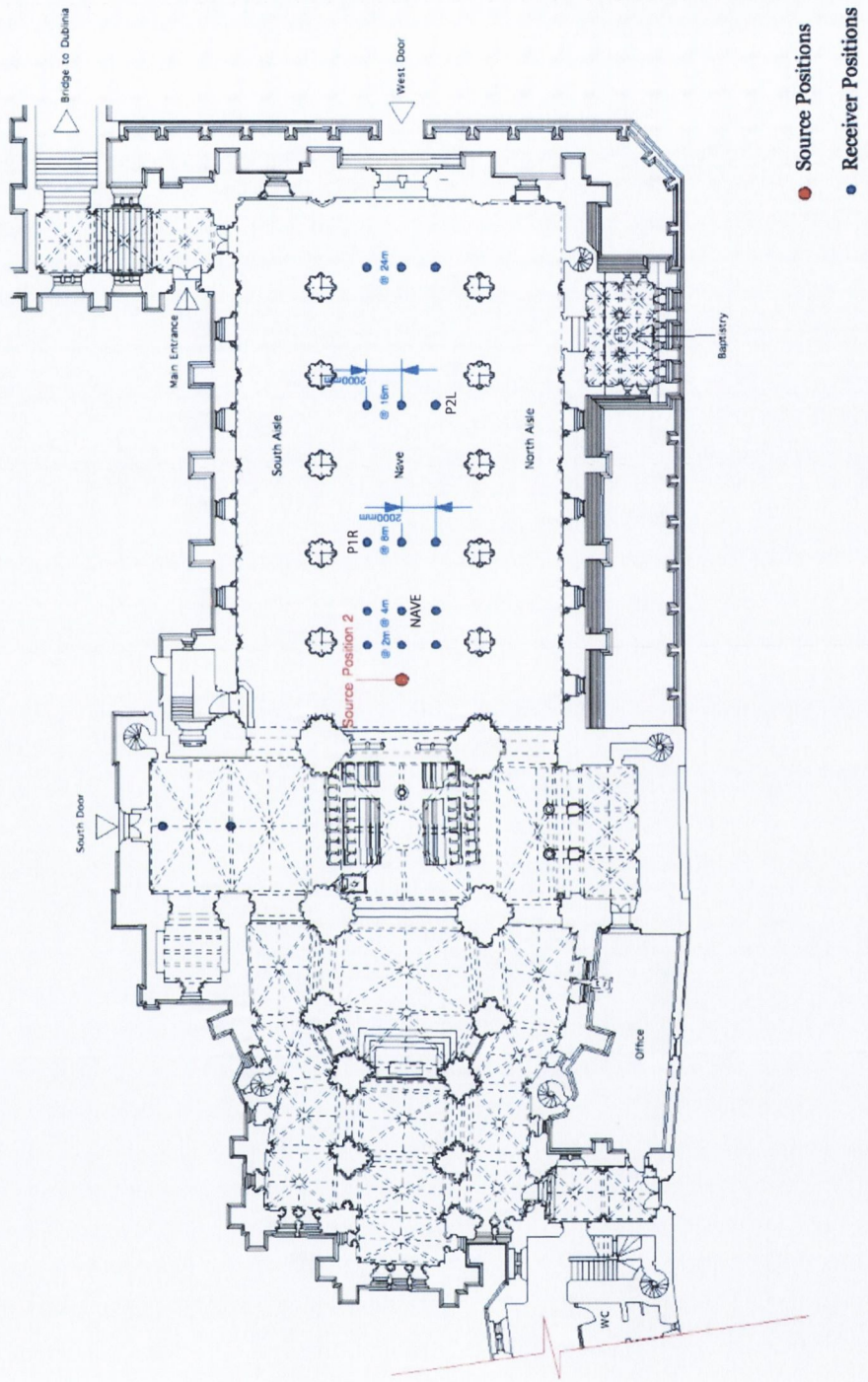
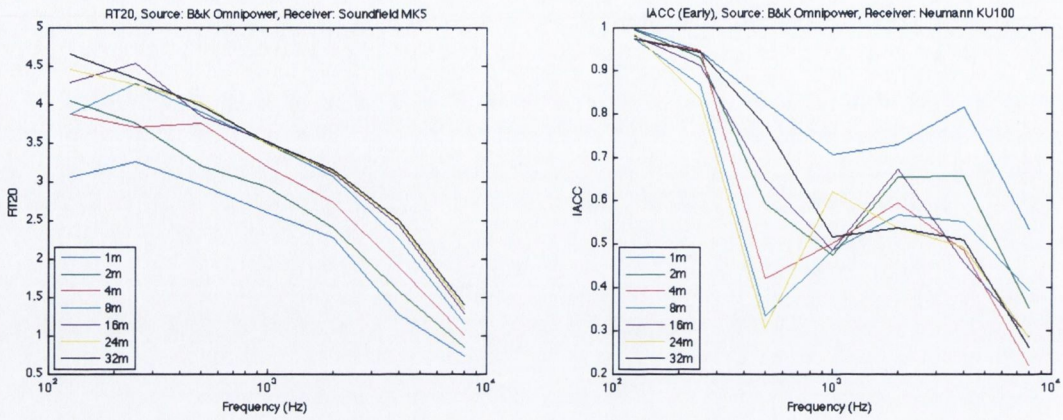


Figure 4.17: Acoustic measurement arrangements for source position 2



(a) (b)

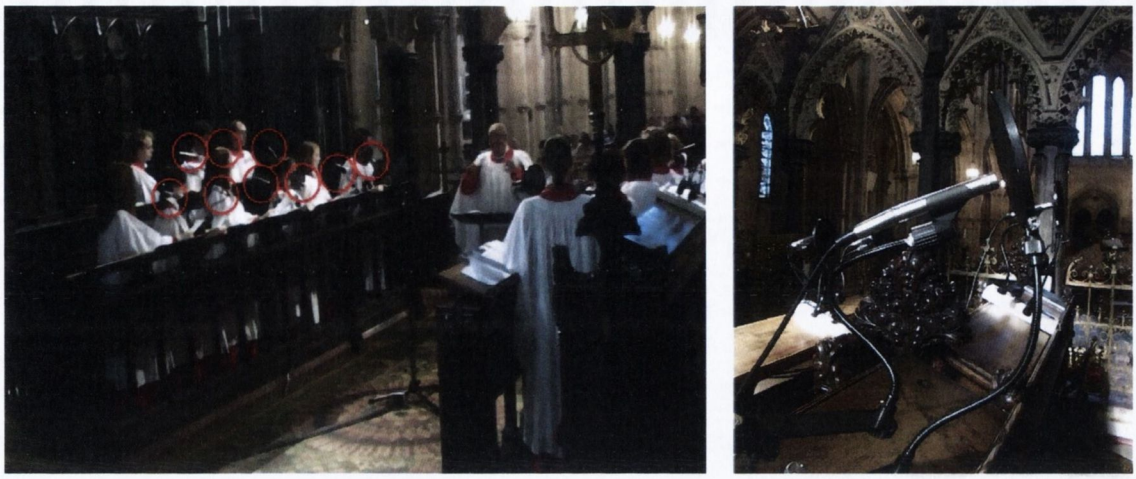
Figure 4.18: Reverberation time (a) and IACC (b) measurements, measured at Nave receiver positions from source position 1

4.4.2.2 Choral Recording

The strong musical tradition associated at Christ Church has been upheld to the present day and the calibre of the cathedral choir is outstanding. In order to create an effective virtual model of the cathedral, it was therefore also necessary to capture a real performance as sung by the choir. An anechoic recording is insufficient, as the resultant performance is not a true representation of the choir interaction with the acoustic, i.e. the performance characteristics and dynamics would be significantly different [31]. The choir was therefore recorded in the Cathedral during a Sunday service in June 2010.

Direct-field capture of each singer was achieved within their critical distance using a spot microphone. The positioning and directional characteristic of the microphone was important not only to the tonal balance, but also to minimise the amount of other sources (commonly referred to as cross-talk) in the recorded signal. Uni-directional microphones are frequently used in order to maximise rejection, but the cost of increased directional response can often lead to compromised frequency response in lower grade microphones as well as proximity effect. Such frequency response distortions must therefore be corrected in post processing. Thus, each of the 19 choir singers was provided with the individual cardioid microphone (*Rode NT5*) as shown in Figure 4.19(a). Each microphone had a custom mount and a pop-shield in order to reduce the amount of plosive sounds (Figure 4.19(b)). Although microphones were placed well within the critical distance of each singer, it was impossible to avoid cross-talk from other singers. However, the signal to noise ratio was deemed acceptable (greater than $10dB$ at each microphone) and did not cause any major tonal distortions during the process of auralisation.

Reference recordings were also made within the cathedral to later compare with the aurali-



(a) (b)

Figure 4.19: Choir recording: (a) Spot microphones (*Rode NT5*) on each performer circled in red; (b) (*Rode NT5*) in custom mount with a pop-shield

sation. These were made at the choir centre position using a tetrahedral microphone array [41], and at the crossing, using the *Neumann KU-100* binaural head and a set of *AKG C-414B* microphones configured as an ORTF pair.

4.4.2.3 Geometric and Spatial Data Capture

A comprehensive spatial and architectural survey of Christ Church Cathedral was carried out in order to build a detailed model with accurate surface, height, and volumetric information. To ensure both accuracy and speed in the data collection, the work was carried out using a terrestrial LiDAR (Light Detection and Ranging) laser scanner at a tightly set matrix to collect surface and height information and thus build up a three dimensional representation of the interior of the cathedral. The age and lengthy restoration history of the building have resulted in a number of volumetric and geometric idiosyncrasies. These constitute an essential part of the cathedrals character, but would be very difficult and time consuming to capture accurately using more traditional surveying techniques such as triangulation and macro-modelling. LiDAR captures information using laser pulses set to a user-specified matrix with distances calculated by measuring the time delay between emitted and reflected pulses, and is currently the swiftest way of capturing this type of information accurately [207].

The model used in surveying Christ Church Cathedral was a *Leica C10*. The *C10* captured data in individual scans at 360° along the horizontal axis and 270° along the vertical axis, thereby allowing for the capture of awkward geometries with relative speed and minimal overlap. Six-inch semi-spherical targets were used to aid the scanning process, and their locations were registered to the scanner for recognition and subsequent automatic merging of individual scans.

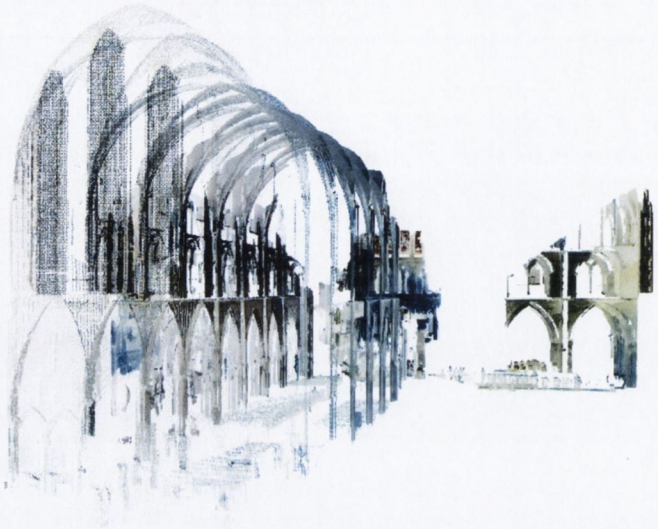


Figure 4.20: An exemplary point-cloud resulting from the LiDAR survey at one location

The entire process was completed in the space of 5 hours with a total of 16 individual scans taken from calculated scanner- and-target locations. Each of the 16 scans was carried out at a $10mm$ by $10mm$ resolution at a range of $10m$. The majority of the cathedral was captured in the first four scans. One of the point-clouds resulting from the LiDAR survey is illustrated in Figure 4.20. The geometry of the main body of the cathedral is quite open, but the interlocking spaces and built-up nature of the area around the transepts and the Lady Chapel required a greater density of scan stations and targets for complete data collection. A built-in, dual-axis level compensator to the scanner allowed for inconsistencies to floor level to be fully taken into account. The collected data from each scan station was saved directly to the hard drive of the scanner - an integrated data manager runs *Windows XP* - then downloaded to an external PC laptop for post-processing. The registered targets were then aligned and automatically merged using *Leica's* proprietary *Cyclone* software.

The laser-captured data sets are in point cloud format, which consists of millions of individual points. Data in this form is read by standard architectural drawing packages as a solid-block object. To this end, the survey data had to be converted into a workable spatial representation by translation into a series of planes and surfaces. This was carried out with the assistance of *Severn Partnership*, with the results exported to *AutoCAD*. From the resultant *AutoCAD* model it was then possible to read out the geometrical data, the most important of which are collected in Table 4.4.

4.4.3 Real-Time Visualisation

Due to audio-visual character of this work it was desired to faithfully model the interior space of the cathedral that could be subsequently used for interactive audio-visual walk-through pre-

Christ Church Cathedral Principal Dimensions:	
Internal Length	$61941mm$
Internal Width (Nave)	$19097mm$
Internal Width (North to South transepts - open area)	$25083mm$
Height	$28661.5mm$
Perimeter	$178440mm$
Internal surface area of cross - section	$607.716m^2$
Overall volume	approx. $38000m^3$

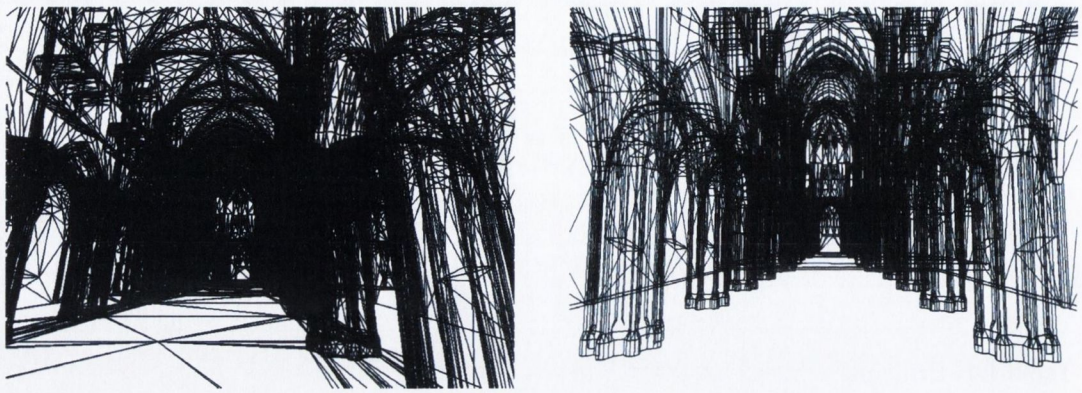
Table 4.4: Principal dimensions of the Christ Church Cathedral based on the readings from the *AutoCAD* model

sentation. For this purpose it was again chosen to visualise the model with the use of *Blender* software. From the previous experience with a much smaller environment, *Blender* passed the test as a convenient tool for creating interactive 3-D architectural reconstructions. What is not without importance, its capabilities to handle professional architectural models created in *AutoCAD* (e.g. *.dxf* files) have also been tested. It was then anticipated that the same pipeline could be successfully employed for modelling as long as the architectural model of the Christ Church Cathedral interior was created using the same file format.

4.4.3.1 Data Redundancy and Reduction

The laser scanning process (LiDAR survey described earlier) allowed the formation of a detailed and highly accurate (high-polygon) mesh reflecting the cathedral's interior. Initially, after importing into *Blender 3D*, the geometry consisted of 95580 polygons (triangles), 85674 vertices and was divided into 23 smaller, manageable sections. The initial wire-frame of the model is shown in Figure 4.21(a). Although such a detailed model would be interesting to work with from the point of view of off-line acoustical analysis, it was obvious that this level of complexity was definitely far too high from the point of view of real-time auralisation. The acoustical analysis and simulation employing the full detailed interior and using wave-based and/or geometrical methods was outside the scope of this work. Instead, methods were sought for adapting the model for efficient, real-time auralisation.

This level of detail would have been also demanding from the point of view of real-time video rendering, particularly on older systems. Although, modern video games can use visual environments with hundreds of thousands or even millions of polygons, it must be emphasised that for satisfactory results they usually employ dedicated hardware, e.g. Graphics Processing Units (GPUs) and are also optimised mainly from the point of view of displaying graphics. Initial tests with the high-level BGE showed that indeed the real-time rendering suffered from low rendered frame rates and it was deemed necessary to reduce the complexity of the model so



(a) (b)

Figure 4.21: Wire-frame model of the church: (a) before polygon reduction process; (b) after polygon reduction process

that it can be run on an off-the-shelf PC.

Thus, the optimisation process have been employed that allowed obtaining a simplified version of the mesh without perceptual loss of detail or quality. The process consisted of removing doubled or nearly adjoining vertices, *quadification* (converting several triangular faces into one single quadratic face, natively supported by the BGE) and polygon reduction. From the visual presentation point of view, the initial model proved to be highly redundant and after the optimisation process, we managed to reduce the polygon count to 35107 (from 95580) and the number of vertices to 30892 (from 85674), which was enough to run the demo on the middle class PC (*Dell Optiplex*, 4-core *Intel* CPU, $4GB$ of RAM, *ATI Radeon* GPU with $256MB$ of RAM). The wire-frame of the model after the reduction is shown in Figure 4.21(b).

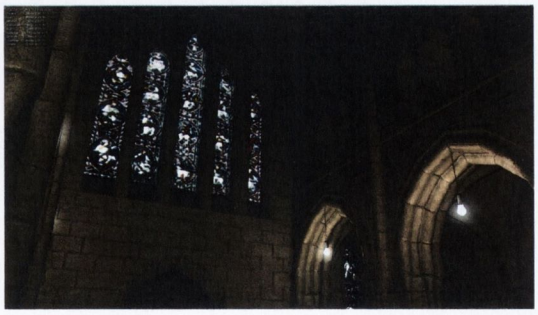
Finally, the 3-D model was also textured and lit in order to achieve a more natural appearance. Textures were created mostly from the photographs taken in the interior of the cathedral and applied to the model's faces using the technique known as UV mapping [156]. Some examples of textures used in this work are shown in Figure 4.22. It must be emphasised though that the ambition here was not to obtain a photo-realistic replica of the space but rather to improve the overall visual impression and recreate a feeling of spaciousness that was somehow distorted in a solid version of the cathedral model. The first person view real-time renders of the finished (textured and lit) model can be viewed in Figure 4.23.

Lastly, In order represent the sound sources in the 3-D space, silhouettes of choir singers were added as simple 2-D textures that were programmed to always face the virtual camera. In this implementation, 3-D co-ordinates of the "virtual" singers correspond to actual real-world positions in the cathedral, where the recordings were taken from. These co-ordinates do not change throughout the whole interactive experience.

Figure 4.22: Exemplary textures used in the model

(a)

(b)



(c)

Figure 4.23: First person point of view finished real-time renders of the Christ Church Cathedral interior as used in the walk-through presentation

4.4.4 Real-Time Auralisation

Having created the interactive Virtual Visual Environment (VVE) of the cathedral, we shall again consider the creation of the VAE. The objective of this work could be stated as the presentation of an aurally plausible reproduction of the captured choir performance in a virtual version of the Christ Church Cathedral.

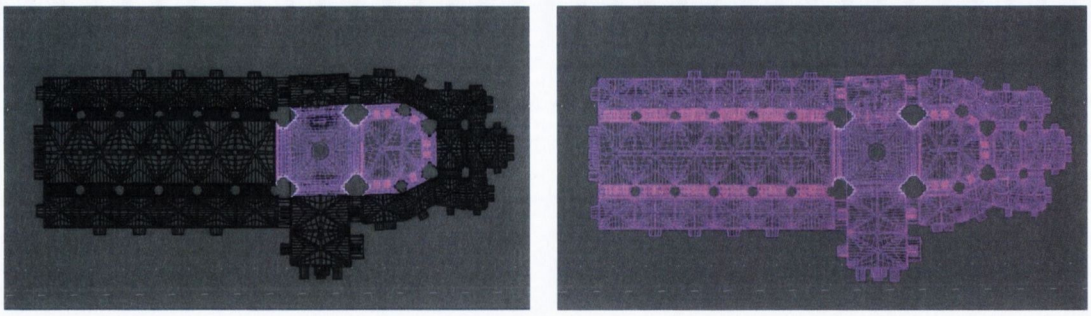
4.4.4.1 Implementation of the Audio Engine

The audio engine used in the auralisation was again fully implemented in *pd*. In fact only minor changes were needed as compared to the auralisation engine used in the Example I (Section 4.3). These changes are detailed below.

At runtime, for a given source-receiver position the magnitude and the time-delay of the direct sound and early reflections from 19 sources this time is computed using the Image Source Method. Then, each source with its image sources is encoded into 3^{rd} order horizontal-only Ambisonics. However, despite apparent simplicity, the ray tracing based methods in general become quickly computationally expensive when complex geometries are used. This is because the number of virtual sound sources tends to grow rapidly with added surfaces. Calculating image sources for a $> 35k$ polygon model (and also handling as many channels of audio) would not be feasible considering a current state-of-the-art technology so further simplifications were unavoidable. As a first approximation for calculating image sources in our model we use two bounding boxes of the interior depending on whether the user is inside (Figure 4.24(a)) or outside (Figure 4.24(b)) the choir area. The global bounding box has the 2-D dimensions of 60×24 meters whereas the inner bounding box is 20×11 meters. The audio engine has been set to logically switch between the two sets of mirror images whenever the user enters/leaves the choir area. For this reason, two sets of *EarlyReflectionGen* objects were used with different mirror image sources.

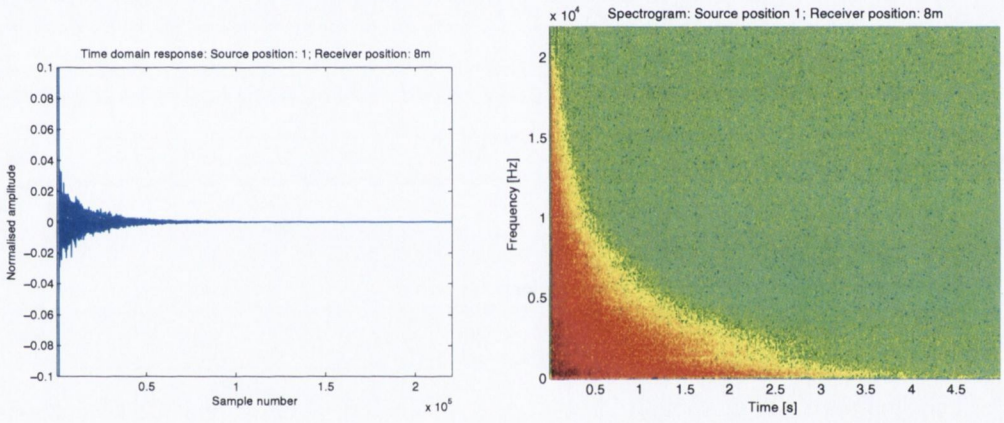
In this work we also represent each sound source as a direct sound and eight reflections (3×3 matrix). In this way, we have nine audio streams per source which gives 171 streams in total when we combine together all the sources in the scene. For direct source-receiver distances greater than $8m$, image sources are no longer used and the auralisation is only of the direct sound and diffuse field. This is due to the fact that at this distance there is little perceptible difference with or without the early reflections in the model, which is also corroborated by the largely consistent acoustic measurements of Section 4.4.2.1 beyond $8m$.

The average reflection coefficient has been chosen to give a good approximation of the general absorptive properties of the cathedral walls. It is based on the measured dimensions of the cathedral as presented in the Table 4.4 as well as the reverberation time at $1000Hz$ and is derived from the Sabine equation for reverberation time (Equation 4.1). Substituting $V \approx 38000m^3$, $S \approx 7200m^2$ and $RT_{60} \approx 3.3s$ and solving for α yields $\alpha \approx 0.25$. Thus, the average surface reflection coefficient β can be calculated as $\beta \approx \sqrt{1-\alpha^2} = 0.97$.



(a) (b)

Figure 4.24: Bounding boxes of the choir area (a) and the whole model (b) were used as a base for calculating image sources sources acting as discrete early reflections



(a) (b)

Figure 4.25: Time domain sequence (a) and the spectrogram (b) of the omni-directional impulse response taken at $8m$ distance from the source position 1. The sampling rate was $44.1kHz$.

In this simplified model, sound sources (choir singers) are represented as 2-D textured planes that are programmed to always face the virtual camera and each of the planes is constantly casting a ray in its normal direction. The ray-casting algorithm returns the logical value $TRUE$ or 1 whenever the virtual camera is visible to the sound source and $FALSE$ or 0 whenever there is an obstacle on the ray's path (like a wall or a pillar). In the case of baffles with semi-transparent textures (like fences) they hold a special property flag that makes them invisible to the algorithm. The visibility flag ($TRUE/FALSE$) is subsequently passed to the audio engine and determines whether to pass or mute the direct sound and reflections from the particular sound source. The ray-casting algorithm is one of the basic BGE built-in logic modules.

Also, because of the fact that sound sources can be accessed only from the limited arc in their front (due to physical arrangements of the choir area), it was decided not to implement the

directivity filters at this occasion. It also allowed for savings in terms of channel count. Otherwise, we have to remember that using the ambisonic approximation for source directivity, each individual source would have to been represented by as many channels of audio as the number of circular harmonics used for decomposition (e.g. 7 in the case of the 3^{rd} order approximation).

In this work the diffuse tail of the reverberation has been obtained from the real SRIR captured using the *Soundfield MKV* system (B-Format) in a similar way as explained in the previous example in this Chapter. The example time sequence of the W channel of the response at $8m$ is shown in Figure 4.25(a) and its spectrogram is shown in Figure 4.25(b). Now, however, the truly diffuse part of the of the impulse response has been extracted from the full SRIR measured at $32m$ as it assured lower initial dry-to-reverberant acoustic energy ratio.

4.4.4.2 Sound spatialisation

A method for sound spatialisation follows exactly the same concept as described previously in the context of the traditional Irish musical performance in the Printing House Hall. However, due to the simplification and lack of source directivity properties, on-line processing complexity has been reduced. On contrary, due to a significantly higher number of direct and mirror image sources, the overall acoustic scene was more complex. The high-level diagram showing the complete signal processing chain of the implemented audio-visual rendering system can be viewed in Figure 4.26.

4.4.5 Summary

This work presented a method for developing interactive audio-visual models of existing architectural spaces, based on the example of the Christ Church Cathedral in Dublin. We have shown how real-world recordings of the choir could be realised using direct-field pickup and subsequently auralised in a walk-through implementation. The project has yielded a significant amount of information about the architectural heritage and the current record of Christ Church Cathedral and has provided a solid framework for expanding the scope of such investigations to other spaces in Ireland and abroad. Both the data collected, and the methodologies utilised are unique to Irish heritage.

The current system or its parts have been also demonstrated to the public at several occasions, most notably: as part of National Heritage week at the Mansion House Dublin (August 2010), Second Irish Workshop on Music and Audio Signal Processing at the Trinity College Dublin (January 2010), Institute of Acoustics 8^{th} International Conference on Auditorium Acoustics (May 2011) [106] and Virtualisation and Heritage Symposium, University of York (February 2012). Whilst the current implementation requires further optimisation and validation, informal listening comparisons between the real and virtual environments resulted in a markedly positive public response. Moreover, the project succeeded in raising public awareness about the importance of acoustic measurement of Irish and foreign historical spaces for cultural

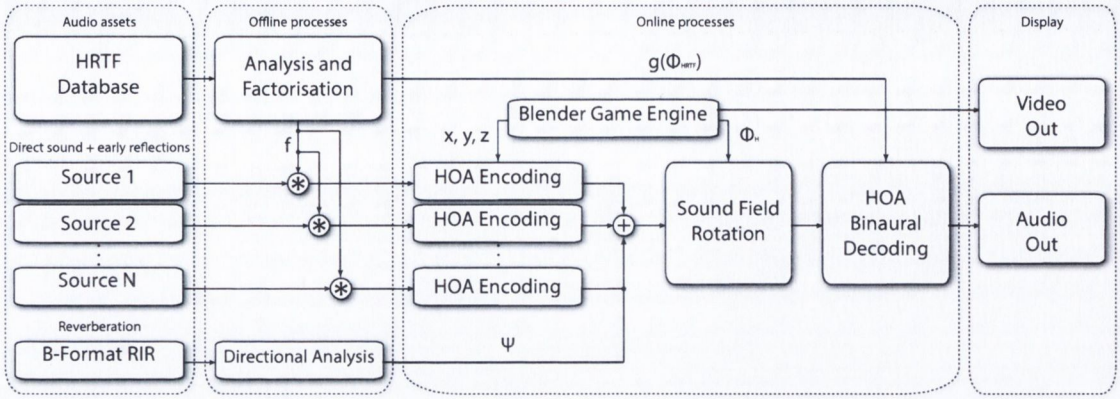


Figure 4.26: Complete flowchart of the implemented audio-visual rendering system. f represents the direction-independent or common component of the whole HRTF dataset. On contrary, $g(\Phi_{HRTF})$ represents direction-dependant residuals, specific to each loudspeaker location in the system. The direction-dependant components are used as run-time FIR filters with kernels of 32 samples. Co-ordinates x, y and z denote the current location of the virtual camera in the modelled space. They are used in the HAO encoding stage to inform the encoder about the relative source-listener angle(s) ϕ_S and distance(s) ρ. Φ_H denotes the resultant virtual camera orientation with respect to the virtual-world axis corrected by the user's head orientation in the real-world co-ordinates. These variables are sent to the audio engine from *Blender Game Engine* in real time using the UDP protocol. Ψ denotes the 1^{st} order B-Format reverberation. Reverberation signals are obtained by offline convolution of the mono mix of all the real sources in the scene (image sources are disregarded) with the diffuse part of the RIR resulting from the Directional Analysis [144]. Finally, + denotes the process of summation of the corresponding B-Format channels.

heritage preservation.

Further work should look into the optimisation of the acoustic model, in particular the early reflection synthesis and audio channel count reduction. A unified software solution that would help to automate the SRIR parametrisation and integration with game engines would also be a sought after effort.

Lastly, because the technique presented optimises the data-based auralisation approach and opens it for real-time and interactive applications, it would be useful to formally compare and contrast the audible result of the two approaches. Also, an objective verification of the model using e.g. $ISO-3382$ acoustic parameters would also help in the evaluation of the method used. This issues are talked about in the next section.

4.5 Evaluation of the Proposed Auralisation Method

In 2005 Lokki in [124] reviewed methodologies used for quality assessment of auralisartion algorithms. He concluded that the problem is still under-explored, especially when it comes to the real-time and interactive auralisations. A direct comparison of binaural recordings using a large number of source and receiver position combinations is recommended by Kleiner et al. [112] as a method for evaluation of the auralisation processes. They also recommend using multiple source signals including speech signals, solo instrument recordings and noise bursts.

Foteinou and Murphy [67] investigated the psychoacoustic effect when changing different acoustic parameters in the geometrical auralisation model. The study was also performed as a comparative listening test in which pairs of reference and test samples were rendered using a pre-defined grid of points within the virtual environment.

Pellegrini in [173] argues that different layers of perceived sound quality must be taken into consideration, namely physical elements (temporal resolution, frequency resolution and bandwidth, spatial resolution, dynamic behaviour), psychoacoustic features (perceived loudness, location accuracy, sound timbre, auditory spaciousness, source size, dynamic accuracy) but also psychological factors which is "Perceived quality of features, which are strongly related to cognition, action and the emotional state of the user" (source-dependent expectation, task-dependent expectation, expectation on interaction, active interactivity, passive interactivity or personal expectation). Some of the implications of the above may be that the physically correct auralisation may not be the sole requirement for plausible reconstruction of an acoustic event but other factors, like visual or tactile modalities may also play a significant role. That is why, experience of a sound in a given room may still differ from the experience of a physically correct binaural recording which is played to the subject in a different environment.

As already explained, the aim of the auralisation process is to "render audible" some physical or conceived space in such a way that the virtual experience is indistinguishable from the real-life experience of this space [112]. However, perceptual evaluation of the auralisation method is not an easy task due to the lack of clearly defined sound qualities the must be met in a "good auralisation" [123]. That is why, a comparative study is usually performed [26, 123, 125] in which the auralisation output is compared to the real listening experience in the auralised space or the binaural recording. Such a methodology has been used e.g. by Kearney in [105] in which he proposes the comparison of the following subjective sound attributes: reverberance, source-width, source-clarity, source-movement, natural-timbre.

The method of auralisation described in this thesis optimises the standard convolution-based method in which binaural recordings are created by the means of filtering of the dry source audio with the RIR and the HRIR functions. That is why it was decided to directly compare these two methods in a series of listening tests. A third method utilising a procedural reverberation was also used in the study. The third method was meant to represent a typical approach to auralisation implemented in game audio, as described earlier in Section 4.1.2.

4.5.1 Method

A comparative study using 3 different methods of auralisation has been performed in which subjects were asked to assess perceptual differences between pairs of audio samples. One of them was the reference virtual recording and the other one was generated based on either the proposed optimisation technique or a simple procedural reverberation:

1. **Full convolution method (reference):** virtual 1^{st} order recordings created using dry or anechoic source material and full, measured SRIRs
2. **Hybrid convolution method:** virtual 3^{rd} order recordings created using dry or anechoic source material, early reflections created using the Image Source Method and diffuse reverberation tail obtained from the full, measured SRIRs using the Directional Analysis
3. **Procedural reverberation method:** virtual 3^{rd} order recordings created using dry or anechoic source material where early reflections and diffuse reverberation tail are synthesised using a procedural approach

The full convolution (reference) examples were created by convolving the dry (female speech) or anechoic (orchestral music) source audio with 1^{st} Order Ambisonic SRIRs measured in the Christ Church Cathedral. The SRIR positions used are marked in Figure 4.17 as NAVE, P1R and P2L and were chosen to represent distances within, around and beyond the critical distance of the space as well as centred and lateral positions. The Ambisonic binaural decode was performed in *pd* and recorded to separate 2-CH WAV files. The sound field information was first decoded to octagonal horizontal-only loudspeaker array using *pd*'s *ambi_decode3* object and then the virtual loudspeakers were created by convolving the resultant loudspeaker feeds with the corresponding pairs of HRIRs using the *8ChAmbisonicsToHRFFDecode* object.

Because of the comparative nature of the study, the need of individually measured HRIRs was mitigated and the same, randomly chosen set of HRIRss (LISTEN database, subject 1021 [97]) was selected for all the subjects.

In the Hybrid convolution method, 2-CH binaural recordings were created using the optimisation techniques described earlier in this chapter. The direct sound and early reflections were spatialised using 3^{rd} Ambisonics (using *pd*'s *ambi_encode* object) and subsequently decoded for an octagonal, horizontal only loudspeaker array (using *pd*'s *ambi_decode3* object). The diffuse reverberation tail was obtained by performing the directional analysis and applying the diffuseness estimate coefficients to one of the measured SRIRs as explained earlier in Section 4.2. Due to the stochastic nature of the diffuse reverberation, the same tail was used in auralisation of all the hybrid samples. The binaural Ambisonic decode was performed in the identical manner as in the case of the reference.

Lastly, the procedural reverberation approach was implemented as follows. The direct sound was spatialised using 3^{rd} Ambisonics, decoded to an octagonal loudspeaker array and then to

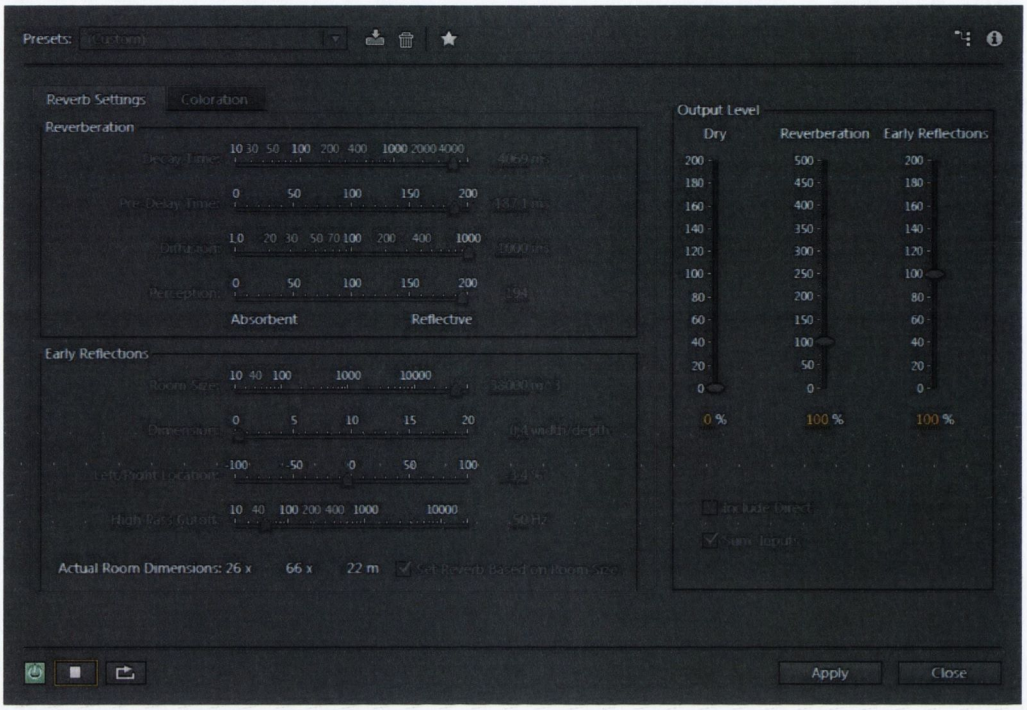


Figure 4.27: A screen shot showing settings of the *Full Reverb* effect in *Adobe Audition CS6*. These settings were used in synthesis of the procedural reverberation for a subset of test samples.

binaural as explained above. Then, the reverberation was procedurally synthesised in *Adobe Audition CS6*[22] using the built-in *Full Reverb* effect. With this approach it was intended to mimic the reverberation techniques that can be found in video game audio engines, e.g. FMOD. Settings of the *Full Reverb* were adjusted so that the reverberation was synthesised based on the cathedral's principal dimensions as listed in Table 4.4. These settings are illustrated in Figure 4.27. The synthesised reverberation was subsequently added to the binaural signals.

In order to avoid any bias due to loudness differences between sample and reference signals, the audio files with the same distance and stimulus types were levelled based on their dB LUFS levels, as per ITU-R BS.1770-2 recommendation [100]. Acoustic parameters - reverberation time (RT20) and IACC (Early) of the BRIRs resulting from three aforementioned methods of auralisation are presented in Figure 4.28.

The study was intended to investigate the perceptual differences between Reference-Sample test pairs. Sample pairs used in the test are listed in Table 4.5. These Reference-Sample pairs were played to the subjects in the pseudo-random order.

In each of these pairs, four subjective sound attributes were evaluated with the following descriptions given to the subjects [105]:

[22] `http://www.adobe.com/products/audition.edu.html`

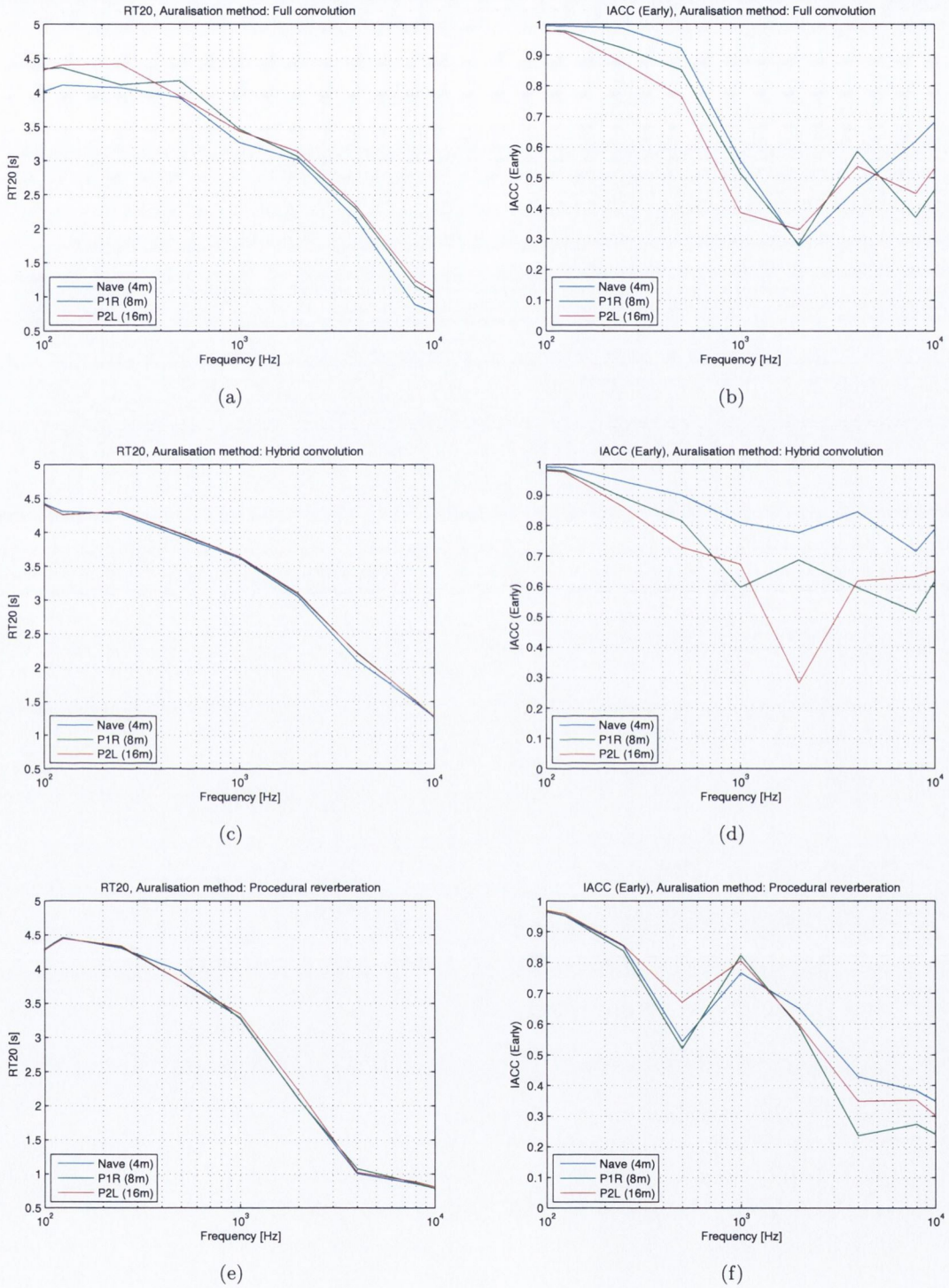


Figure 4.28: Acoustic parameters - reverberation time (RT20) and IACC (Early) of the BRIRs resulting from three different methods of auralisation: Full convolution: (a) & (b); Hybrid convolution (c) & (d); Procedural reverberation (e) & (f).

Test Item	Method	r_z	r_x	ϕ_S	Code	Stimulus
1	Hybrid	4m	0m	0°	O: Nave	speech
2	Hybrid	8m	+2m	−14.03°	O: P1R	speech
3	Hybrid	16m	-2m	+7.16°	O: P2L	speech
4	Procedural	4m	0m	0°	P: Nave	speech
5	Procedural	8m	+2m	−14.03°	P: P1	speech
6	Procedural	16m	-2m	+7.16°	P: P2L	speech
7	Hybrid	4m	0m	0°	O: Nave	music
8	Hybrid	8m	+2m	−14.03°	O: P1R	music
9	Hybrid	16m	-2m	+7.16°	O: P2L	music
10	Procedural	4m	0m	0°	P: Nave	music
11	Procedural	8m	+2m	−14.03°	P: P1	music
12	Procedural	16m	-2m	+7.16°	P: P2L	music

Table 4.5: Test samples used in the study. r_z and r_x signify the relative offset of the listening position with respect to the source position on the distance (z) and horizontal (x) axis respectively. ϕ_S is the sound source angle.

- **Reverberation:** *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 Sample?*** *Listen to the tails of notes, ends of phrases etc.*
- **Source width:** *The space a source occupies within the recording.* ***Is the source in Sample bigger or smaller than in the Reference?***
- **Clarity:** *How well-defined and intelligible is the source sound?* ***Is it more or less clear in the Sample than the Reference?*** *Does it sound more or less muffled?*
- **Timbre:** *Accurate and realistic tonal quality.* ***In overall tonality, is the Sample more or less realistic than the Reference?***

These subjective attributes were rated on a continuous hedonic categorical scale with seven points ranging from "Far Less" (-3) to "Far More" (+3). A custom test software has been built in *Matlab* which allowed for storing the data gathered and was used to randomise the order of test pairs for each subject. This was done to assure that any possible learning and ordering effects are minimised. The GUI of the test software is illustrated in Figure 4.29.

For each pair, subjects had to rate all four attributes before they were able to proceed. They were clearly instructed that they always rate the *Sample*, so for example by setting the *Reverberation* slider to "Far More" they indicated that there is far more reverberation in the *Sample* than in there is in the *Referenece*. There was no restriction in terms of how many times they could listen to both *Reference* and *Sample* files.

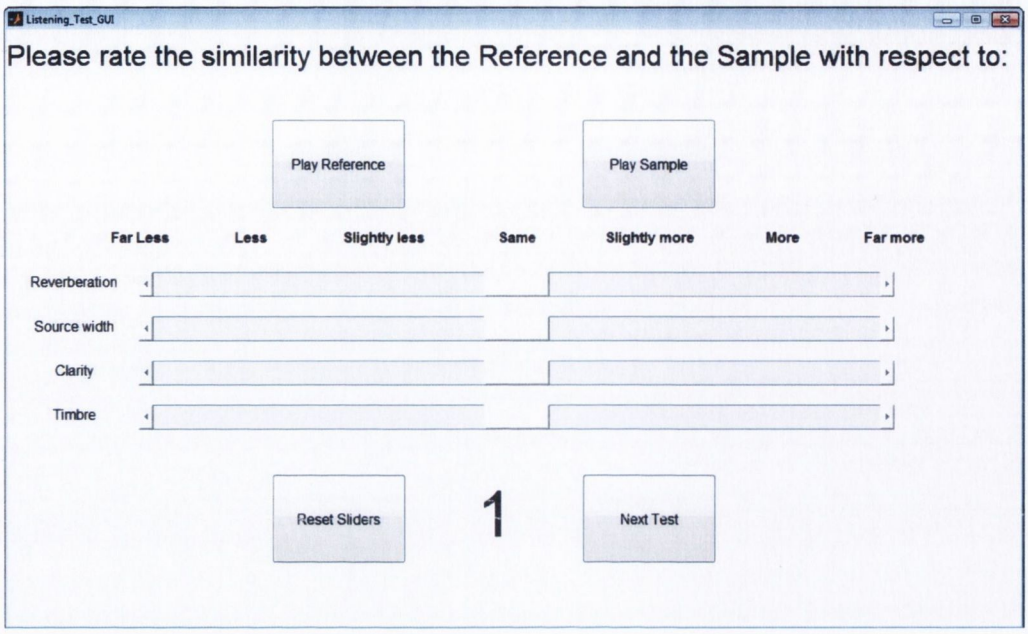


Figure 4.29: GUI of the *Matlab* test software which was used for subjective evaluation of auralisation methods.

4.5.2 Results

20 subjects participated in the study (14 male and 6 female). They were mainly postgraduate Music Technology students and staff members. This number of subjects is recommended as a minimum in the ITU-RBS.1284-1 recommendation for subjective assessment of audio quality [99] when non all the subjects can be considered expert listeners. The definition of the expert listener is quite open though and with enough exposure to the test material non-expert listeners can become sufficiently expert. The advantage of having the expert listeners though would be the reduced time and better accuracy in finding the likely results.

Prior to the test, all the participants were able to familiarise themselves with the test protocol as well as the test material used.

For all of the test pairs, the average scores from all the subjects with corresponding standard errors were computed. These results are presented in Figures 4.30 and 4.31. In order to identify the differences between different auralisation methods, a Two-sample t-test has been performed for each pair of samples. Additionally, in order to investigate the effect of stimulus, for each of the 3 source-receiver locations (Nave, P1R, P2L) a 2-way factorial ANOVA has been performed where factor A was the auralisation method (hybrid convolution vs. procedural reverberation) and factor B was the stimulus type (Speech vs. Music).

For the proposed hybrid convolution approach, in general the average rating of reverberation is not significantly different from the reference. The average ratings are always close to "Same"

score, with the exception of the P1R (Speech) sample, which was assessed on average as having "Slightly More" reverberation. The situation is different for the procedural approach in which more variation of user ratings can be found.

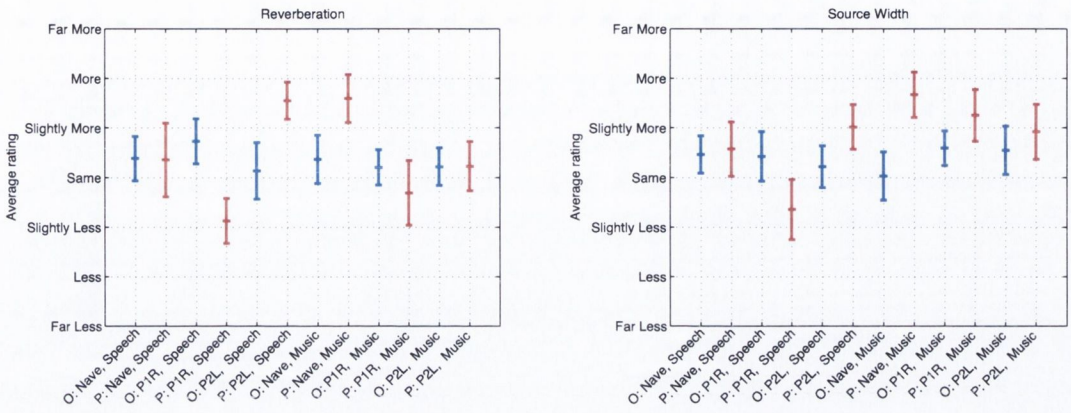
The speech sample using the procedural reverberation at the location P1R (P: P1R, Speech) is rated as having slightly less reverberation than the reference and also less than the (O: P1R, Speech) sample. This result is statistically significant ($p - value = 0.00002$). On contrary, at a further distance (P: P2L, Speech) the same speech sample is assessed as significantly more reverberant ($p-value = 0.00023$). More reverberant is also sample (P: Nave, Music) ($p-value = 0.0012$) so there is a statistically significant difference between the ratings for speech and music. The opposite is found for the position P2L where speech has been rated as more reverberant than music ($p - value = 0.00109$).

In terms of the source width parameter, for the hybrid convolution approach the average user ratings also oscillate around points 0 and 1 ("Same" and "Slightly More"). On contrary, the average scores for the procedural approach are in general higher in the range 0.5 to 2 ("Slightly More" and "More") with the exception of the sample (P: P1R, Speech) for which the source width was assessed between "Same" and "Slightly Less". The most prominent difference between two different auralisation methods can be observed for music sources and especially for the Nave position for which the source width in the procedural reverberation sample was rated as significantly wider than the one using hybrid convolution reverberation ($p-value = 0.00046$).

The hybrid convolution method, 3^{rd} order Ambisonics spatialisation was used for direct sound and early reflections. Although one should expect the source width to diminish if compared with the 1^{st} order reference, it was not the case and sound sources (using both stimuli) were on average rated as having the same size as the reference. These results corroborate the results obtained by Kearney in [105] which clearly show that using Image Source Method in virtual recordings of singing voice and violin may lead to sound source widening.

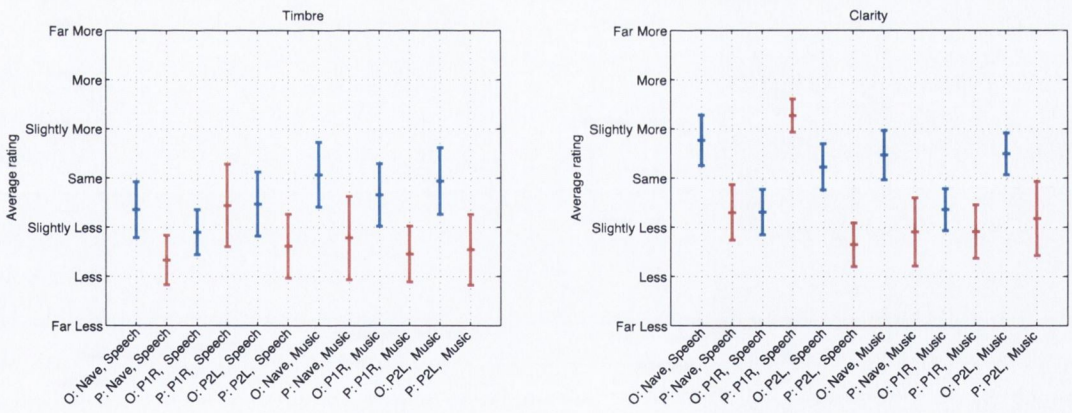
Regarding the timbre, the hybrid convolution approach is on average rated as "Same" and "Slightly Less" natural than the reference. Minimally worse results are obtained for speech but this is hardly surprising because humans are particularly accustomed to the timbre of human voice. The timbre of procedurally generated reverberation was in general rated as worse than the reference. It was also less natural than that of the hybrid convolution except for position P1R and speech source. This difference however is not statistically significant. The most noticeable and systematic differences are observed for music sources at all positions (Nave, P1R and P2L). All these differences are statistically significant ($p-values = 0.0243, 0.091$ and 0.0087 respectively).

In terms of clarity, except for the source-receiver position P2R, the hybrid convolution approach is rated on average between "Same" and "Slightly More". For P2R and for both speech and music samples, the rating is between "Same" and "Slightly Less". The procedural approach is evaluated as significantly inferior for all positions and source types except for the speech source at P2R. This sample was rated on average between "Slightly More" and "More" and the difference between the two auralisation methods is significant ($p - value = 5.1875 \times 10^{-8}$).



(a) (b)

Figure 4.30: Average user ratings for sound parameters Reverberation (a) and Source Width (b) within 95% confidence intervals. "O" indicates the hybrid convolution method whereas "P" signifies the procedural reverberation method.



(a) (b)

Figure 4.31: Average user ratings for sound parameters Timbre (a) and Clarity (b) within 95% confidence intervals. "O" indicates the hybrid convolution method whereas "P" signifies the procedural reverberation method.

4.5.3 Discussion

The results of the listening test show that subjective sound quality of the optimised auralisation method does not differ significantly from the sound quality of the reference recordings. From the results obtained as well as the informal comments by the subjects, the most prominent differences were observed in terms of timbre and clarity. For example, some people claimed that the optimised recordings sound slightly brighter than the reference and for some of them it made them sound slightly less natural. However, this timbral colouration in some cases translated into better clarity of both speech and music sources.

Also, some subjects reported that it is possible that a change in one aspect of the quality may have wider effects. A good example of this is the speech source at location P1R and using procedural reverberation. The clarity of this particular sample is rated quite high - between "Slightly More" and "More" than the reference. At the same time, we notice that the reverberation as well as the source width parameters for this same sample are rated on average around "Slightly Less". The lower amount of perceived reverberation and at the same time higher amount of direct signal could then explain the narrower perceived source width and better perceived clarity. However, this result is not replicated for the orchestral music source at the same location. The possible reason could be that the orchestral sample was rendered in the simulation as a point source which could create some difficulties in rating parameters such as source width or clarity.

In summary, the outcome of the above study is satisfactory. However, there is a strong indication that the focus on timbral aspects of the optimisation technique is required and more work on this aspect is needed in order to improve the realism of auralisations. A possible solution would be to equalise the hybrid SRIR against the reference measured SRIR to avoid any tonal distortion.

Lastly, the study was performed under static conditions (with no source and listener movements). However, the optimised auralisation method is designed for a fully interactive and real-time experience. The reference method (full convolution) on the other hand is designed for acoustic snapshots and does not work in dynamic scenarios. Although fast convolution is feasible due to fast partitioned block FFT-based algorithms the problem of sufficiently dense and large database of SRIRs still holds. Nevertheless, even if the latter problem has been solved and direct comparison of two auralisation methods under dynamic conditions was possible, then there remain issues regarding the design of such a comparative study. Of main concern would be how to compare the audible output of two dynamically rendered auditory scenes if we allow the user to interactively generate them? Because no such method has been devised yet the subjective comparison of auralisation methods is done in a static way using audio recordings. Future work is required that would develop auralisation evaluation methodologies that would take into account dynamic aspects of auralisations including listener/source movements and interactivity.

4.6 Conclusions

In this chapter a review of auralisation methods was presented. Also, a practical framework for creating real-time walk-through auralisations based on real-life acoustic events was proposed. This work can be directly applicable to virtualisation of musical performances but also to other virtual reality applications, including video games. Regarding the latter, the novel aspects and improvements of sound field reproduction that are still missing in state-the-art game audio techniques (see Section 4.1.2), like better reproduction of distance cues and stabilisation of acoustic images with head-tracking could be of advantage. Also, the Doppler effect to simulate fast source/receiver movement, comes virtually at no additional cost since it is naturally implemented with the variable delay networks for the direct sound and early reflections.

Another positive aspect of the system that should be brought forward here is undoubtedly its ease of integration with head-tracking devices. In this system the head-tracking can be seamlessly integrated with the user rotations in the virtual world. Using the Ambisonics approach rotation is applied to the whole sound field and does not depend on the number of sources.

However, it must be noted that the current system in its present form is by no means complete. Several challenges has been identified in the course of the work and by the formal listening tests that would certainly benefit from a deeper insight. First of all, implementation of source directivity currently imposes a high demand for audio channels since it is their weighted combination that results in the proper radiation pattern. Perceptual validation of the resolution of radiation for different instruments or sound sources in general would be definitely helpful since it could dictate the maximum approximation order necessary.

Secondly, for similar reasons, generation of early reflections is also still an open question, especially for more complex geometries like the one of the Christ Church Cathedral. Naturally, it is possible to achieve very accurate results (e.g. using wave-based simulations) at high computational cost. However, the question is rather what level of accuracy can actually guarantee the perceptually equivalent impressions as in the case of real-world listening.

Lastly, tonal colouration and its influence on the perceived quality of real-time auralisation also needs more attention. An equalisation mechanism that would de-colour the reconstructed hybrid SRIRs based on the reference, measured SRIR could be a solution here.

One of the informal comments by the users but also author's own observations suggest that the interactive experience in which SRIR are recomputed to account for the current listener/source location combined with head-tracking to stabilise the sound field lead to efficient externalisation of the sound source and give a convincing perception of the source distance. With the assumption that the diffuse field does not change in the simulation, user interaction means that the ratio between the direct sound and the late reverberation energy varies as the user explores the virtual environment. So, in this way some some very important distance cues should be preserved. However, one of the aspects of this hybrid system which is not clear so far is whether increasing the directional resolution of the early parts of the SRIRs only has any

effect on the perception of distance?

This question leads to the next Chapter which investigates the perception of distance in the context of Virtual Auditory Environments using First and Higher Order Ambisonics.

5

Perception of Auditory Distance under Real and Virtual Conditions

One aspect of a correctly auralised sound field that we have not considered so far is the locatedness of sound sources in the scene at their correct distance. As outlined in Chapter 2, human abilities and limitations in this regard are not fully understood yet, however there are at least some important cues that can affect our judgements.

As it was shown in Section 2.3, perception of auditory distance can be significantly different in a free field and in rooms due to the lack or presence of reflected acoustic energy. Thus, the influence of quality or accuracy of reconstructed room acoustics on the perception of distance in the rendered scene constitutes an important research question. Moreover, since the enhanced quality or accuracy of audio is usually connected with increased computational demand, this question seems to be valid from the practical point of view. For example, in video games, a monophonic sound source encoded into full 3-D 3^{rd} order Ambisonics sound field would require three more 4 times more channels of information than the sound source encoded using 1^{st} order Ambisonics. In practice, for a moving sound source it would mean the scaling each of the channel using 16 spherical harmonic coefficients instead of only 4 recalculated at each logic tick. If we account for the fact that usually more than 1 source is used at a time in a game, then these differences are even more prominent.

However, if not all the sound sources are equally important to the game play or some of them are distant enough, it may be the case that low resolution spatial rendering of these sources would still be sufficient. Although it is well understood [105] that the gradual reduction

of spatial resolution of sound fields presented to listeners makes it increasingly difficult to localise sound sources in both the vertical and horizontal planes, it is still an open question whether it also has an effect on the perception of distance. In particular, does the order of Ambisonics preproduction matter in the presence of strong monaural cues such as level changes and direct-to-reverberant ratio?

As pointed out earlier, trading-off the directional resolution of audio often means significant savings in terms of the bandwidth and computational overhead of the audio rendering systems. Therefore, testing for human perceptual limitations in this regard may lead to optimal methods of rendering auditory distance or even make it possible to adjust the quality of auditory information based on its importance to the rendered sources or provided processing power, for example.

We have already identified in Chapter 3 that amongst the audio spatialisation systems that are the most often quoted in the literature for their optimal means of rendering virtual auditory environments are Vector Based Amplitude Panning (VBAP) [182], Wave Field Synthesis (WFS) [22] and Ambisonics [72]. However, Ambisonics provides an attractive framework to deal with the research questions stated above. Since the spatial resolution of the sound field can be varied, it creates the opportunity for the direct comparison of sound fields of different orders with respect to provided sensation of distance.

To address research questions formulated above a series of subjective listening tests has been carried out with the extensive use of Ambisonic techniques. The pilot experiment was intended to preliminary examine the human ability to distinguish between sound sources at different distances in a reverberant space given clear visual anchors. Performance of a simplified headphone based rendering system utilising First and Higher Order Ambisonic-to-Binaural decoders with head-tracking was also evaluated in terms of its ability to provide externalised acoustic images. Lastly, estimations of distance using headphone based systems were compared to similar estimations of the real sound sources with a proper scientific rigour using formal statistical analysis.

Encouraged by and built upon the outcomes of the pilot study, two main experiments were carried out to examine more formally the differences (if existent) in the perception of distance for real and virtual sound sources. The first experiment focused on the virtual sources rendered using headphones whereas in the second experiment of main concern were sound fields rendered using loudspeakers.

The last experiment looked at the problem of modal dominance in the perception of distance in audio-visual presentations. Using the techniques and findings from the previous experiments, virtual sound sources were rendered at some predetermined distances. However, these distances were most of the time incongruent with the visual information displayed on the TV screen. The effect of conflicting audio-visual cues on the perception of distance was subjectively evaluated.

5.1 Pilot Experiment: Distance Estimation in Real and Virtual Auditory Environments using Visual Anchors

5.1.1 Introduction

The pilot experiment was intended to determine whether virtual sound sources rendered over headphones can be effectively perceived externally, outside the head. To find the answer a group of participants performed a task that was intended to show how well can they relate the sound they hear to the visual objects they see located at various distances. Then, the following question was how the estimates of the perceived distance of virtual sound sources would compare to the perceived distance of real sound sources and would the increased directional resolution of the virtual auditory scenes affect the localisation accuracy?

As pointed out earlier in this work, Virtual Auditory Environments VAEs can be realised using both headphone and loudspeaker based systems. The technique used should not make a difference as long as the correct signals (i.e. pressure changes at left and right eardrums) are extracted by the hearing system [152]. In practice headphone listening allows for greater control over the personalised sound field reproduction due to the fact that reproduced sound waves do not interact with boundaries of the listening environment and the cross-talk is minimised by the head. This was also the technique used in implementation of the auralisation engine in Chapter 4.

A common method in headphone auralisation is to incorporate HRTFs into the reproduced signals. It has already been discussed how HRTFs describe the influence of a listener's torso, head and pinnae on impinging source sound waves. However, as these features are highly individual it is quite easy to distort the true localisation of the virtual sound source if non-individualised filters are used. In order to mitigate this effect, head-tracking should be employed to control the spatialisation process [19].

However, the switching of the directionally dependent HRTFs with head movements can lead to auditory artifacts caused by wave discontinuity in the convolved binaural signals [168]. As described in Section 3.5.1 a more flexible solution is to form "virtual loudspeakers" from HRTFs, where the listener is placed at the centre of an imaginary loudspeaker array. Then, the loudspeaker feeds are changed relative to the head position and any technique for sound spatialisation over loudspeakers can effectively be used, including VBAP, Wave Field Synthesis or, as in this case, Ambisonics.

However, the spherical harmonic decomposition of the sound field used in Ambisonics, exhibits certain advantages over the other popular systems, especially when it comes to perceptual testing. Asymptotically holographic approach to spatialisation means that the when more spherical harmonic functions describing the sound field are used, then the more 'point-like' the sound source becomes. To an experimenter it means that the spatial resolution of the sound fields can be varied independently from other factors that might affect the perception of distance.

5.1.2 Methodology

In this experiment subjects were asked to identify the active source of audio from among 9 loudspeakers located directly in front of them. The loudspeakers were separated by 1m distance and slightly misaligned in order to ensure the "acoustical transparency". This is visible in Figure 5.1(c). Audio samples were presented first through the real loudspeakers and then through the headphones. The latter renderings were reconstructions of sound fields generated by the same set of loudspeakers but previously captured and processed in order to provide virtualised versions thereof. Processing of headphone signals consisted in convolving the dry source audio material with captured B-format Spatial Room Impulse Responses (SRIRs) for a given source-receiver location. From the first order responses, higher order responses were derived using Directional Analysis techniques outlined in Section 3.4.5. Such obtained Ambisonic signals were subsequently decoded for the purpose of binaural presentation using the "virtual loudspeaker" approach as described in Section 3.5.1. Additionally, head-tracking has been employed in order to compensate for any users' rotational head movements. Due to obvious differences in the apparatus requirements, this experiment had to be conveyed in two separate phases: first, for the real sources distance estimation and second, for the virtual sources distance estimation.

The remainder of this section explains in detail the technical aspects of the experiment as well as its exact process. The results are also presented and followed by the detailed analysis and discussion.

5.1.2.1 Stimuli

Two sets of stimuli were used in the experiment. In the first set there were presentations of anechoic music, extracted from the *Denon* anechoic orchestral database [51]. The second set consisted of phonetically balanced speech phrases selected from the TIMIT Acoustic-Phonetic Continuous Speech Corpus database [64] and recorded by a female reader. The sampling rate of the playback was 44.1kHz and the bit depth was set to 16.

5.1.2.2 Test conditions and apparatus

The experimental loudspeaker setup used for the test is shown in Figure 5.1. Nine loudspeakers (*Genelec 1029a*) were spaced 1m apart from each other starting from the 0m reference point (the listening position). In the design of this test, it was deemed necessary for the subject to clearly acoustically "see" each loudspeaker at each position. However, unconsidered placement of the loudspeakers could potentially lead to horizontal and vertical localisation cues that might bias the perception of source distance. Previous studies by the author have revealed that there is an azimuthal localisation blur of 0.25m and an elevation blur of 0.6m for the source located at 1m from the listener in this test environment [81]. Thus, the array was set up so as to not exceed these perceptual limitations, whilst giving the subject clear view of each loudspeaker. This is shown in Figure 5.1(b) & (c). The loudspeakers were calibrated to 80dBA at 1m from

their on-axis tweeter position.

In the second phase of the test presentation of samples was over headphones. The headphones used were open back *AKG-K601* which exhibit low levels of interaural magnitude and group delay distortion as recommended by Adams et al. in [1]. It is important to note, that non-individualised HRTFs were used in this preliminary test. However, the resultant sound externalisation was very effective, largely due to head-tracking, and no front-back confusion was reported. The HRTFs used were extracted from the IRCAM LISTEN database (subject 1021) and were diffuse field equalised [97].

Sound field rotation, tilt and tumble control was implemented via the *InterSense InertiaCube2+* head tracking system, resulting in stable virtual images when changing head orientation. *IntertiaCube2+* is a high accuracy (1° yaw, 0.4° pitch & roll), high rate (180Hz) and low latency (4ms) Inertial Measuring Unit (IMU) that can be mounted on top of most standard headphones in order to provide reliable estimates of subject's head orientation. However, in order to avoid drifts of data, which is a common problem with the IMUs, it requires frequent re-calibration using provided Compass Calibration Tools for static magnetic field compensation.

5.1.2.3 Participants

15 participants (5 female and 10 male) were used in total. The test subjects consisted primarily of music technology students under 35 years of age and of good hearing. Each participant was first subject to a training session prior to the test, where they were presented with stimuli from each of the loudspeakers and were asked to become accustomed to the source types and acoustics of the room, and most significantly the perception of depth. During the course of this training visual indicators in the form of LED lights attached on top of each loudspeaker confirmed the active loudspeaker.

5.1.2.4 Procedure

Because of the different apparatus requirements, the test was conducted in two separate phases. There was a short interval between the phases required for the participants to put on headphones and for the experimenters to calibrate the head tracking device. All 15 participants completed both phases of the test.

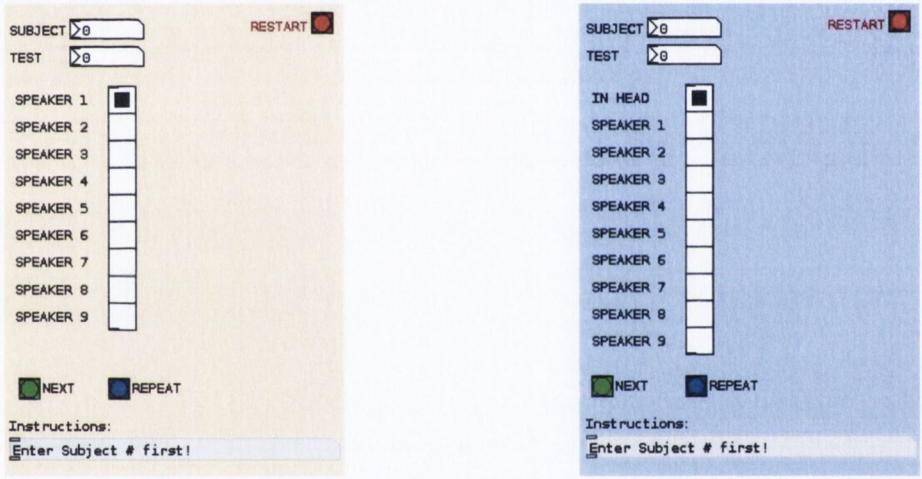
In the first phase of the test, each participant was presented with stimuli originated from randomly selected loudspeakers. The randomised method was used to negate any ordering effects during the test. After each presentation, they were asked to identify the location of the sound sources via the graphical user interface (GUI) shown in Figure 5.3. In the first phase of the test the number of choices for the sound source location was narrowed to 9 visible physical loudspeakers as shown in the Figure 5.3(a). In the second test phase, where headphone listening was involved, additional "IN HEAD" option was added for subjects to avail from whenever it was not possible to perceive the sound source externally (Figure 5.3(b)). A visual confirmation

(a) (b) (c)

Figure 5.1: Loudspeaker setup for the preliminary distance perception study: (a) Loudspeaker setup diagram; (b) Side perspective of staggered loudspeakers; (c) Subject perspective.

(a) (b)

Figure 5.2: Inertial Measurement Unit *Intersense InertiaCube2+* used for head tracking



(a) (b)

Figure 5.3: Graphical User Interfaces (GUIs) used in the: (a) first phase of the pilot experiment (reference - real sound sources) (b) second phase of the pilot experiment (virtual sound sources presented over headphones)

of the choice was given by the lit LED that was attached to the loudspeaker of choice. The GUI, audio engine as well as the loudspeaker LED lights control modules were all implemented in the *Pure Data* [181].

In the second phase of the test subjects were also asked to identify sound source distances using the same procedure. However, this time instead of physical loudspeakers they were assessing the distance of virtual sound sources in Ambisonic sound fields that were presented over headphones. The test stimuli again consisted of the same music and female speech samples, but convolved with B-Format Ambisonic impulse responses.

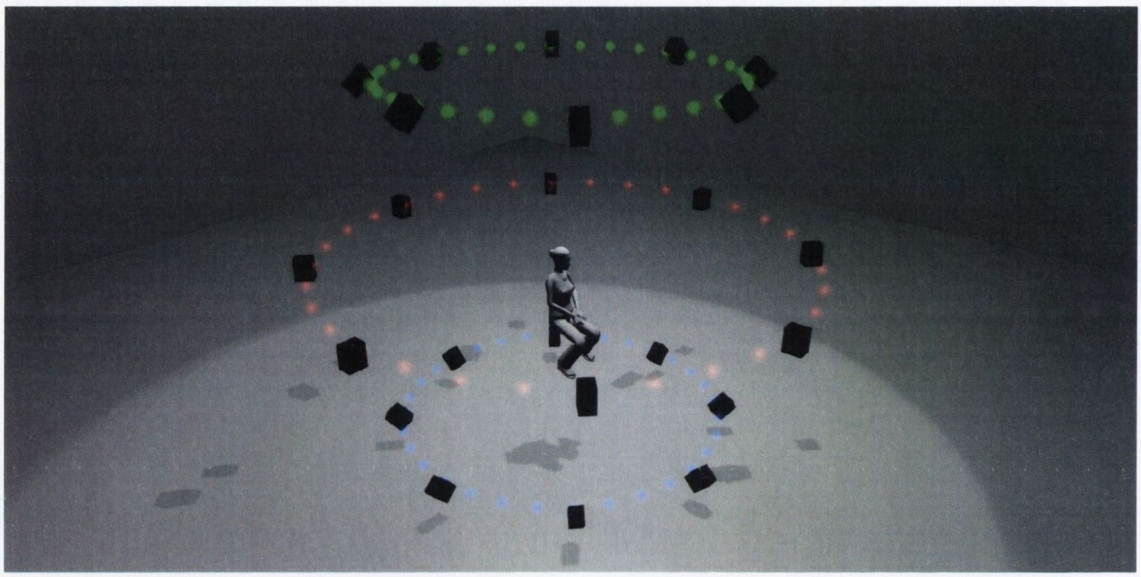


Figure 5.4: Virtual Loudspeaker setup used in the pilot experiment: 3 tiers of 8 loudspeakers each, arranged on an imaginary sphere of radius 1.95m. The angular distance between each tier was 45 °.

For this reason, prior to the test, first order Ambisonic impulse response measurements were taken from the listener position of each loudspeaker using the exponential sine swept tone method [59]. From these measurements, 2^{nd} and 3^{rd} order impulse response sets were extracted using the directional analysis approach outlined in Section 3.4.5.

To decode Ambiconic sound fields, 24 virtual loudspeakers (i.e. HRTF pairs) were implemented, arranged in three tiers of 8 on an imaginary sphere around the listener as depicted in Figure 5.4. The vertical spacing between the tiers was 45 ° with the middle tier located at the ears' hight of the listener. The radius of the sphere was 1.95m which was congruent with the distance of the HRTF's measurements.

The overall latency of the system was set at 20 ms as that was the lowest latency that did not cause audible artefacts in binaural audio stimuli presentation. Brugart et al. [34] investigated the influence of different head-tracked latencies on localisation accuracy and showed that "values of less than 70 ms are adequate to obtain acceptable levels of localization accuracy in virtual audio displays".

The only psychoacoustical optimisation applied to Ambisonics decodes was shelf filtering and was intended to satisfy Gerzon's localisation criteria for maximised velocity decode at low frequencies and energy decode at higher frequencies [79]. This involved changing the ratio of the pressure to velocity components at low and high frequencies as explained in the Section 3.3.3. For low frequencies, the ratio was set to unity which ensured the Velocity decoding scheme. For high frequencies and the Energy decode, correction gains k_n^m for higher order sound field

components were computed according to Table 3.2 and applied in the decoding process. The exact numerical values of the correction coefficients used are listed in Table 5.1.

k_n^m	$m = 0$	$\|m\| = 1$	$\|m\| = 2$	$\|m\| = 3$
$n = 0$	1			
$n = 1$	1	0.577		
$n = 2$	1	0.775	0.400	
$n = 3$	1	0.861	0.612	0.305

(5.1)

Table 5.1: Correction gains k_n^m for higher order sound field components used in the Energy decoder type at higher frequencies

Gerzon [78] recommends setting the crossover point for the high frequency boost in the pressure channel at around 700Hz for the first order systems. This is based on the Makita's theory which predicts the low-frequency (below 700Hz) localisation [131]. However, other authors [118] recommend lower values, e.g. 400Hz, for regular loudspeaker listening (e.g. where it is difficult to achieve a perfectly centred listening position). In this work the crossover point was restored to 700Hz, since the subject was always perfectly in the centre of the virtual loudspeaker array.

For higher orders, Daniel [49] suggests that for ideal listening conditions higher crossover frequency settings could be used for higher order sound field components (e.g. around 1200Hz for the second order system). The rationale for this is that at higher orders, the area of correct reconstruction grows and the sound field is reproduced correctly around the head also at higher frequencies using the basic or velocity decode (see Chapter 3 Section 3.4.3). However, since ideal listening conditions could not be ensured in this study (e.g. not perfectly regular array of loudspeakers was used), it was decided to keep the crossover frequency point at 700Hz also for the 2^{nd} and 3^{rd} order decoders.

Also, no NFC filters were employed in this study. Because the listener was always in the centre of the virtual array, there was no risk of angular bias due to near field effect. Regarding the possible low frequency boost/cut the requirement for distance compensation filtering due to near field effect for the array radius of 1.95m and different virtual sound source distances used to be prominent only below 100Hz as shown in Figure 5.5. Thus, for the female speech test stimuli, this would not have an effect since the first formant frequencies do not go down below 180Hz. On the other hand, it is acknowledged that for the orchestral music samples insufficient or excessive bass energy could be more prominent, especially in the case of 3^{rd} order sound fields. However, after auditioning the orchestral samples it was realised that no excessive colouration is present that would potentially affect distance judgements.

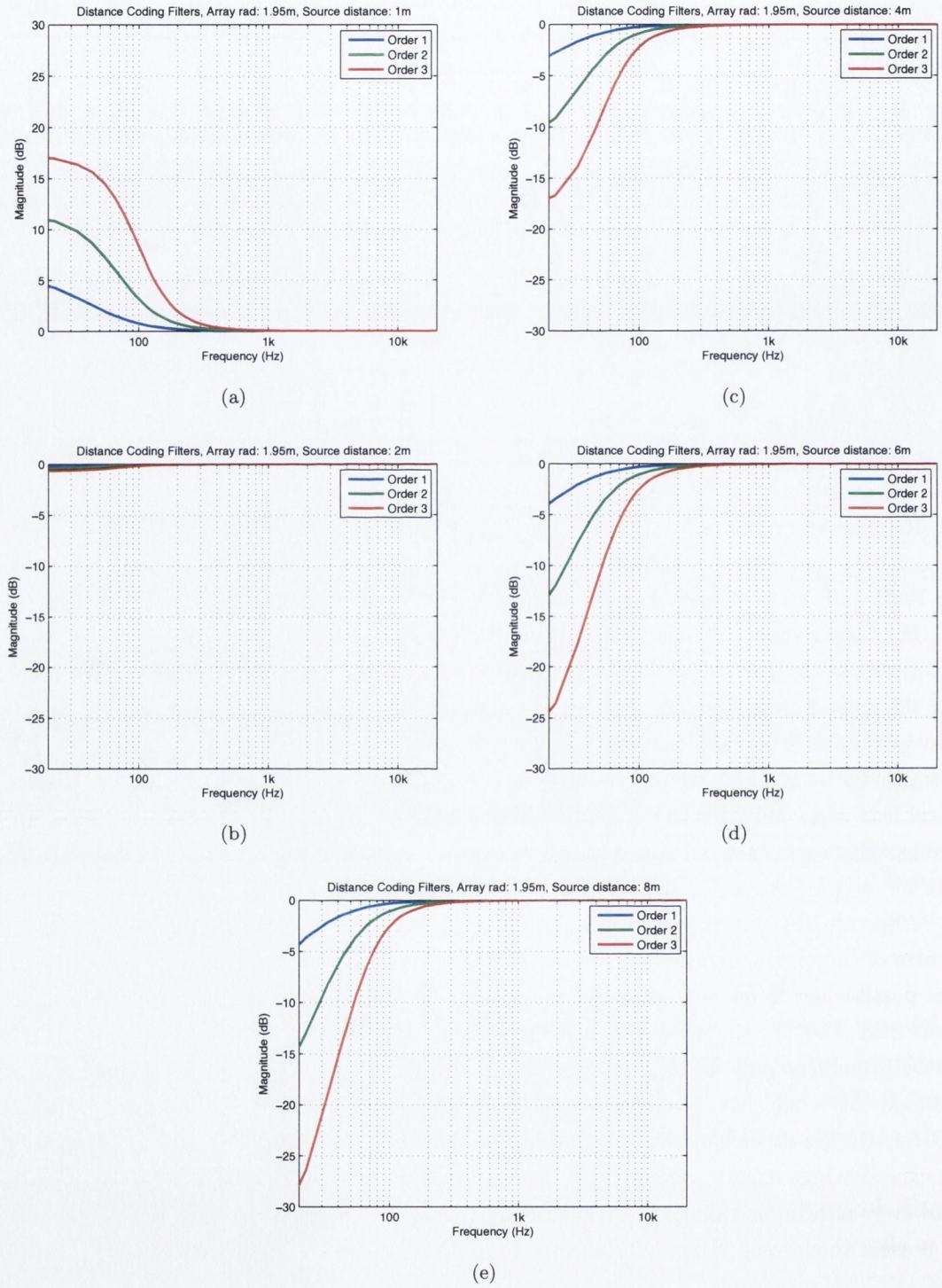
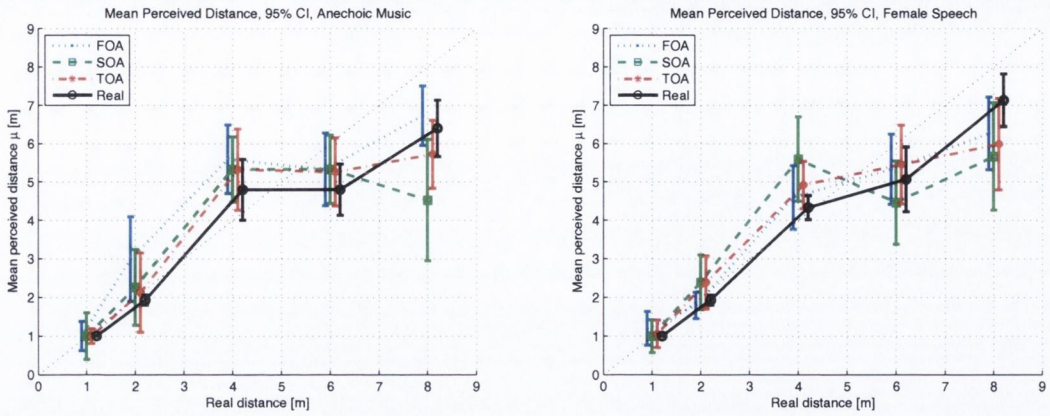


Figure 5.5: Distance coding filters required for direct sound compensation up to 3^{rd} Order Ambisonics in the Pilot Experiment. Compensation filters are obtained for loudspeaker array radius of 1.95m and sound source distance of: (a) 1m; (b) 2m; (c) 4m; (d) 6m; (e) 8m.



(a) (b)

Figure 5.6: Mean distance perception of real and virtual sound sources within 95% confidence intervals: (a) Anechoic music samples; (b) Female speech samples. 'FOA', 'SOA' and 'TOA' denote first, second and third order Ambisonics respectively whereas 'Real' indicates the real, physical sound sources.

5.1.3 Results

In this experiment, the perceived source distance from the 15 listeners for each stimuli was collected for 5 different positions (1m, 2m, 4m, 6m and 8m) for both the reference sources and for 1^{st}, 2^{nd} and 3^{rd} order Ambisonics . For each source type and for each source position, the average distance μ (over the 15 subject answers) and the corresponding standard error $se(\mu)$ have been computed. The results are presented separately for each stimulus type within 95% confidence intervals.

Since the study followed the within-subject factorial design with 2 (stimuli) * 4 (playback conditions), in order to investigate the effects of these two factors (referred later as factors A and B) as well as potential interaction effects, for each presentation distance a two-way ANOVA has been performed with the use of the MATLAB routine `anova2`. The null hypothesis tested here was that all the mean perceived distances for all the stimuli and playback methods are equal, i.e:

$$H_0\text{: } \mu_{FOA}=\mu_{SOA}=\mu_{TOA}=\mu_{Real}=\mu$$

$$H_1\text{: not all localization means } (\mu_i) \text{ are the same}$$

From the analysis, no statistically significant effect of stimuli (music versus speech) on the perception of distance has been found ($F_{1m}(3,112) = 0.26$, $p = 0.61$; $F_{2m}(3,112) = 0.56$, $p = 0.45$; $F_{4m}(3,112) = 1.74$, $p = 0.19$; $F_{6m}(3,112) = 0.09$, $p = 0.76$; $F_{8m}(3,112) = 0.48$, $p = 0.49$).

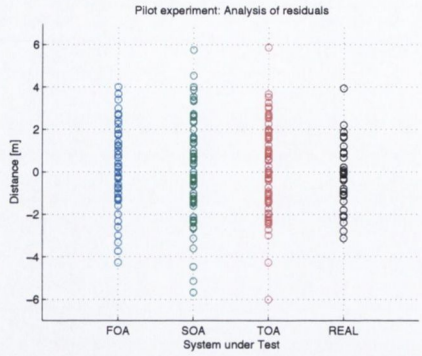


Figure 5.7: Pilot experiment: analysis of residuals. On the x-axis are four systems under test. On the y-axis there are differences between the mean perceived distance for all the test points and individual distance estimates from 15 subjects.

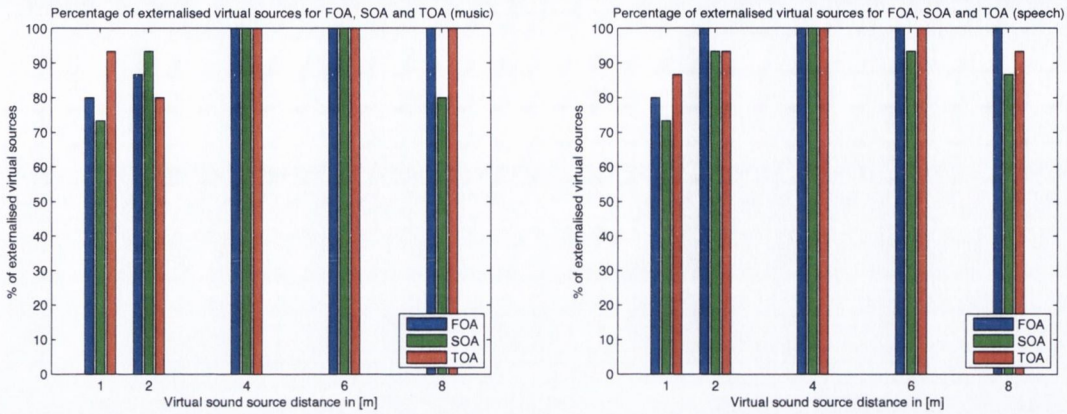
Also, performance of four tested playback methods did not affect distance judgements in any statistically significant way ($F_{1m}(3, 112) = 0.13$, $p = 0.94$; $F_{2m}(3, 112) = 0.61$, $p = 0.61$; $F_{4m}(3, 112) = 1.51$, $p = 0.22$; $F_{6m}(3, 112) = 0.58$, $p = 0.63$; $F_{8m}(3, 112) = 2.59$, $p = 0.06$). Lastly, no synergetic effect of factors A and B have been detected either.

5.1.4 Discussion

As expected, the perception of distance for both real and virtual sources was more accurate for near sources. The classic tendency (see Chapter 2 Section 2.3) to slightly overestimate the distance of sources up to 4m in some environments is also clearly visible. Beyond 4m, gauged distance was continuously underestimated which is also congruent with the previous studies outlined in Chapter 2 Section 2.3. Furthermore, the standard deviation of localisation increases as the source moves further into the diffuse field.

The mean localisation of the virtual sources follows the reference (denoted as "REAL") source localisation well and no statistically significant difference has been observed between the perception of music and speech sources. The answers for the virtual sources deviate from their means roughly in the same fashion as the answers for reference sources, as localisation becomes more difficult within the diffuse field. However, in general perception of real sound sources exhibits less between-subject variation which is better visible in the analysis of residuals presented in Figure 5.7. The most dramatic difference between real and virtual sound fields can be observed at the closest distances (1m and 2m) where virtually no or minimal deviation of answers occurs for the reference sources (Figure 5.6).

The effectiveness of virtual sound source externalisation using non-individualised HRTFs has also been assessed. In each trial of the test phase II subjects had the option of marking all the occurrences where in-head localisation was taking place. From this data, percentages of subjects able to perceive externalised sound images were computed for both stimulus types and all the



(a) (b)

Figure 5.8: Percentages of externalised virtual sound images: (a) Anechoic music samples; (b) Female speech samples

tested distances. These results are visualised in Figure 5.8. From this analysis, it can be inferred that the externalisation was more effective for further sources - beyond 4m. However, for close sources externalisation was still satisfactory and never dropped below 70%. Beyond that, no subject reported any front-back confusion problems. However, it is unclear whether it was the case because of the use of the head-tracking system or it was due to prior expectations due to the visual component of the test.

Other informal comments from the subjects after the tests indicated that whilst the overall spatial impression was more defined with the higher order samples, it did not affect their judgement of the source distance. Moreover, the level of the direct sound to the diffuse field, as well as the stable virtual imaging was reported to give strong distance cues.

5.1.5 Conclusions

In this pilot experiment the performance of virtual Ambisonic sound fields with respect to rendering of sound sources at their correct distance was assessed through subjective analysis. It was first established that headphone renderings with head-tracking can lead to externalised acoustic images even when non-individual HRTFs are used. However, some room for improvement has been identified, especially if the intention is to render close acoustic images. It is hypothesised then, that the use of individually measured HRTFs might benefit further studies in the subject.

Secondly, the influence of spatial resolution of the auditory scene on the perception of distance has been investigated. First order recordings were obtained by convolving dry source audio material with B-Format SRIRs. Then, higher order sound field synthesis was achieved using the directional analysis method of [144]. The assessment was carried out by the means of visual selection of available answer points. Interestingly, it was discovered that in Ambisonic

reproduction subjects' choices matched the choices for the real world source at each order. No significant statistical difference was exhibited by increasing the spatial resolution in this regard.

It must be emphasised though that this analysis applies to particular test conditions in which subjects had to choose from available visual anchors (a forced-choice test). It was expected then that the visual aspect could affect the auditory perception. However, the extent of this effect, if any, is yet to be investigated. Therefore, in the quest for more reliable perceptual results the following experiments should focus on assuring more controlled test conditions and eliminating possible visual distractors. In the light of this, first conducted experiment investigated further the perception of distance in the case of headphone rendering. The second experiment dealt with sound fields rendered over the multichannel loudspeaker array.

5.2 Experiment I(a): Distance Perception in Real and Virtual Environments using Headphones

5.2.1 Introduction

In the previous study it was shown that, having the visual anchors available, it is possible to distinguish between virtual sound sources located at different points in the distance. In fact, it has been shown that the localisation accuracy of sources presented over headphones with head-tracking and using virtual acoustic recordings is not worse than the localisation of the real, reference or original sound sources. However, there were signs that this ability deteriorates as the sound sources move further away into the diffuse field.

The previous test was based on multi-modal presentation of stimuli. Auditory information was always accompanied by a choice of visual anchors. Such a situation is not only commonplace in everyday life but also in a multitude of multimedia applications including video games or films. It also makes it easier for the examiner and the test participant to collect the user's feedback. However, in order to fully understand the effect of certain auditory cues on distance judgements it was intended that all the other sensory information is completely removed from the test protocol.

In the light this, the following study further examined to what extent the distance information can be inferred from the virtual audio recordings presented over headphones. However, this time all the possible visual anchors has been removed from the study. Also, making a choice on a continuous scale has been now made possible and the judgements of auditory distance have been collected in this way for real and virtual sound sources. Again, of special interest was whether the increased directional resolution of presented sound fields would directly translate to a sharper localisation of sound objects in the distance. For this reasons, sound fields were rendered using 1^{st}, 2^{nd} and 3^{rd} order Ambisonics utilising Ambisonic-to-Binaural decoding technique described in Section 3.5 of Chapter 3.

5.2.2 Methodology

Different protocols used for subjective assessment of distance perception can be found in literature, most notably direct verbal report [86, 239] direct or indirect blind walking [83, 127] or imagined timed walking [83]. All of these methods proved to provide reliable and comparable results for both auditory and visual stimuli with direct blind walking exhibiting the lowest between-subject variability [83, 127].

In the previous pilot experiment, subjects were indicating the perceived distance of real and virtual sound sources by pointing at physical loudspeakers lined up (and slightly off-set in order to provide "acoustic transparency") in front of them. However, in this study, in order to completely eliminate any possible anchors as well as visual cues, it was decided to utilise the method of direct blind walking.

In this method, participants listen to sound samples generated at some predetermined but randomly selected distances. The important aspect of the protocol is that vision is precluded (e.g. by using a blindfold). After several presentations of a test item, listeners are asked to indicate where they think the sound originated from by walking toward the sound source. At this stage, sound stimulus is no longer repeated. For safety reasons as well as to make the walking process easier and more comfortable, a guiding rope is usually stretched from the origin (or listening point) to the point several meters after the last intended sound source location. In this method, the distance walked by the subject is assumed to represent the distance at which the sound source has been perceived.

5.2.2.1 Participants

Seven participants aged 24-58 (5 male and 2 female) took part in the experiment. All of them were of good hearing and were either music technology students or practitioners actively involved in audio research or production. Before the experiment all the participants were subjected to a training session in which they familiarised themselves with the test protocol, space and the programme material used. That is why, according to the ITU-RBS.1284-1 [99] recommendation they could be classified as expert listeners.

Prior to the test, HRIR data for all the participants has been obtained in a sound-proof, large (18 x 15 x 10[m]) but quite damped (RT60 @ 1000Hz = 0.57s) multi-purpose room (*Black Box*) in the Department of Theatre, Film and Television in the University of York. Additional damping was assured by thick, heavy, curtains covering all four walls and a carpet on the floor. The measurement process consisted of a standard procedure in which miniature, omnidirectional microphones (*Knowles FG-23629-P16*) were placed at the entrance of a blocked ear canal in order to capture acoustic pressure generated by one loudspeaker at a time. The loudspeakers were located at a constant distance but various angular directions.

As shown in Figure 5.9, subjects were seated at elevation so that their ears were 2.2m above the ground level and their heads were in the centre of a custom-built spherical loudspeaker

Figure 5.9: Array of 16 loudspeakers used for HRIR measurements

array, arranged in diametrically opposed pairs. The array consisted of 16 full range loudspeakers *Genelec 8050A* since the intention was to reproduce virtual sound sources as ambisonic sound fields of up to and including 3rd order. The diagram with loudspeaker arrangement is also shown in Figure 5.10. This 3-D setup comprised a flat-front, horizontal octagon (marked in the picture with the red dots) and a cube arrangement (four loudspeakers on top, and four on the bottom of a cube marked with the blue dots). The radius of the loudspeaker array (and thus that of the virtual loudspeaker array) was 3.27m.

For FOA-to-binaural decodes, only virtual loudspeakers from the cube configuration were utilised since no directional resolution is gained by using a higher number of loudspeakers. Furthermore, despite careful alignment, oversampling of the sound field with higher number of loudspeakers has the potential to yield sound field distortions [23].

For higher orders, all 16 loudspeakers were used. Although the oversampled configuration was not optimal from the 2nd order reproduction point of view, it was not possible to easily and accurately rearrange the loudspeaker array in order to accommodate for a different layout. However, any potential misalignment was reduced by both careful positioning of the loudspeakers

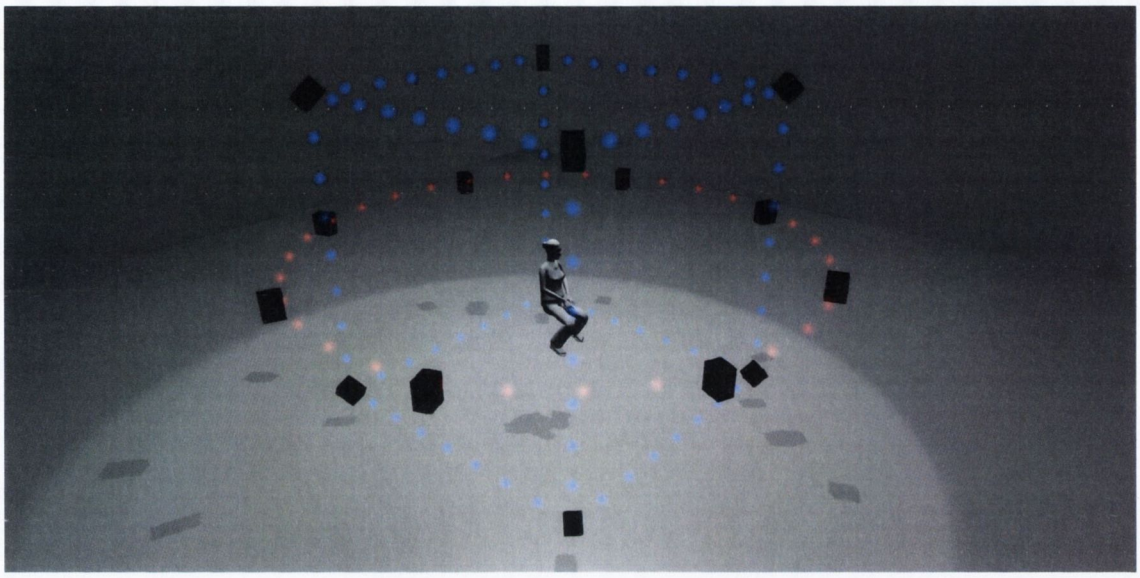
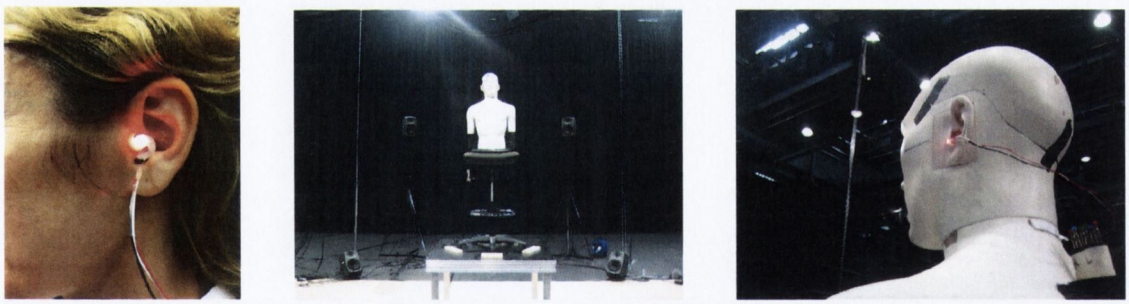


Figure 5.10: Virtual array of 16 loudspeakers used in the experiment: The setup consists of 8 loudspeakers placed at the corners of an imaginary cube (makred with the blue dots) and used for the First Order Ambisonic renderings. For Second and Third order renderings additional 8 loudspekaers were utilised that were arranged on an imaginary circle with radius 3.27m (marked with the red dots).



(a) (b) (c)

Figure 5.11: Measuring Head Related Impulse Responses with miniature in-ear microphones: (a) human test subject; (b) & (c) KEMAR binaural manequin

in the array and also by assuring that the subjects' heads are kept at exactly centre position using laser pointing (Figure 5.11).

For the 3rd order reproduction, the array used was also not perfectly regular with the undersampling of loudspeakers at the top and bottom of the array. That is why, sound field reconstructions could potentially suffer from errors, especially when the sound source was located above or below the listener. However, because of the technical (the number of available sound card channels and loudspeakers) and space limitations (it would be difficult to place a

loudspeaker e.g. below the listener without causing any shadowing effects), it was decided not to add additional channels. It must also be noted that all the virtual sources used in the test were rendered in front of the listener, at the hight of the horizontal ring of loudspeakers where the sound field was not undersampled which could minimise the potential rendering artefacts.

HRIRs were captured using exponential sine swept tone technique [59] at 44.1kHz sampling rate and 16bit resolution. Because of the fact, that the measurement environment was not fully anechoic further processing of HRIRs data was necessary. First of all, measured HRIRs were tapered before the arrival of the first reflection (from the floor) yielding filter kernels with 257 taps. The measured HRIRs were subsequently diffuse-field equalised in order to remove the colouration due to the individual microphone and loudspeaker frequency response characteristics in the signal processing chain.

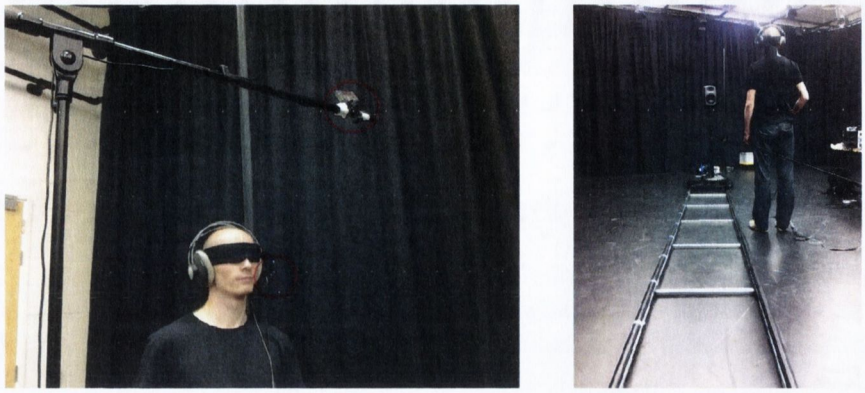
It must be noted, that the minimum number of expert listeners recommended for listening trials is 10 [99]. In this study it was not the case. Because of the complexity and multi-stage character of the experiment (HRIR aquisition session on one day and the main experiment stage on another day) it turned out that not all the subject could be available for both sessions. At the same time, due to technical and logistic reasons, the sophisticated setup for HRIR acquisition could not be recreated in order to obtain HRIR data for more subjects.

However, there are many reports in the related literature, where lower numbers of subjects than recommended are used. Nevertheless, meaningful and statistically significant results are obtained. For example, Zahorik in one of his journal publications [239] reports two subjective experiments in which acquisition of BRIR data for each of the subjects was necessary. He used 9 and and 7 subjects respectively. Similarly, in another journal publication Bronkhorst and Houtgast [32] report experiments on distance perception using 6 subjects.

It is worth noting though that the experimental design is equally crucial as the number of subjects used. For example, a factorial study design can lead to higher statistical power of the test than a simple pairwise comparison (i.e. two-sample t-test) even if the same sample size is used [159].

5.2.2.2 Stimuli

The stimuli used in the experiment were pink noise bursts and phonetically balanced phrases selected from the TIMIT Acoustic-Phonetic Continuous Speech Corpus database and recorded by a female reader [64]. It is important to note that it was decided to replace the music samples from the preliminary test for a couple of reasons. First of all, the results obtained using female speech and music samples did not differ significantly so one could potentially be dropped. Secondly, it was decided that pink noise bursts would better represent a source type that is unfamiliar to humans. Therefore, these two sample types were selected in order to represent both unfamiliar and familiar sound sources to test whether this could potentially influence the perception of distance. Sampling rate of 44.1kHz and 16bit resolution was used in



(a) (b)

Figure 5.12: Participant performing a trial during the experiment: (a) at the origin (head-tracking system has been marked with red circles); (b) on the walking path

both cases. They were presented to the subjects in a pseudo-randomised manner to avoid any ordering effects.

5.2.2.3 Test conditions and apparatus

A series of subjective listening tests has been conducted in the Large Rehearsal Room in the Department of Theatre, Film and Television, University of York. The room dimensions were 12m x 9m x 3.5m and the spatially averaged reverberation time at 1kHz was 0.26s. The low reverberation time was desired for this study so the walls were covered with thick, heavy curtains, as shown in Figure 5.12. Since the up-mix of the First Order Ambisonic sound fields concerned only the deterministic part of the measured Room Impulse Responses, it was assumed that no advantage would be gained from using a more reverberant space.

A professional camera dolly track has been set up roughly in the direction of the diagonal of the room. Doing this not only allowed for testing distances of the real loudspeaker up to 8m but also assured that early reflections did not arrive at subject ears at the same time. A single full-range loudspeaker (*Genelec 8050A*) was mounted on a camera dolly which enabled it to be noiselessly translated by the experiment assistant to different test locations. The guiding rope was hung along the dolly track with the intention to help and guide the participants when walking toward the sound source. Since it was not possible to walk exactly on the dolly track, it was decided that the walking path will be designated directly next to it, as shown in the Figure 5.12. The only weakness of this solution was that the sound source horizontal angle varied from 14.04° at the closest tested distance (2m) to 3.58° at the furthest tested distance (8m). However, as it turned out, it did not have any effect on the distance judgements for at least two reasons. First of all, the subjects were allowed (or even encouraged) to rotate their heads in order to fully utilise available ITD and ILD cues. Secondly, the initial head orientation

(a) (b)

Figure 5.13: *NaturalPoint TrackIR 5* head-tracking system [161] (a) Hi-speed infra-red USB camera; (b) *TrackClip PRO* - clip with three infra-red LEDs used for triangulation

was not fixed in any way. This, combined with the fact that there were no clear cues so to the subject's initial orientation in the room at the origin, made this small initial angular offset unimportant. Also, none of the participants reported any bias in their assessment based on the horizontal offset of the sound source.

For the trials with binaural presentations, again high quality open back headphones (*AKG-K601*) were used, which exhibit low levels of interaural magnitude and group delay distortion, as recommended by Adams in [1]. But this time, the sound field rotation, tilt and tumble control was implemented via the *TrackIR 5* infra-red head-tracking system [161], resulting in stable virtual sound images with head rotations. This solution, making use of the infra-red visual tracking technology, turned out to provide a good compromise between the precision of measurements, field of view, latency and additionally it did not required frequent re-calibration as opposed to the IMUs. It consisted of a high-speed infra-red camera connected via USB port to a laptop PC (Figure 5.13(a)) and a clip with three infra-red LEDs used for triangulation, attachable to a set of headphones (Figure 5.13(b)). The system responsible for playback of virtualised sound sources was completely built in the *Pure Data* and its latency was 20ms, similarly as in the Pilot Experiment.

5.2.2.4 Procedure

In the experiment, subjects entered the test environment blindfolded and without any prior expectation regarding such parameters as the room dimensions, the room acoustics or the test apparatus. They were guided by the experimenter to the test reference point (the "origin"). After a brief explanation of the experiment objectives a training session began with a short (3-

5min) walking-only trial until participants felt comfortable with walking blindfolded and using a guide rope. Next, they performed 4-6 training trials in which sounds were played by the loudspeaker at randomly chosen distances. No feedback was given and no results were recorded after the test trials. The end of the training session was clearly announced and after approx. 1min interval, the first phase of the test began.

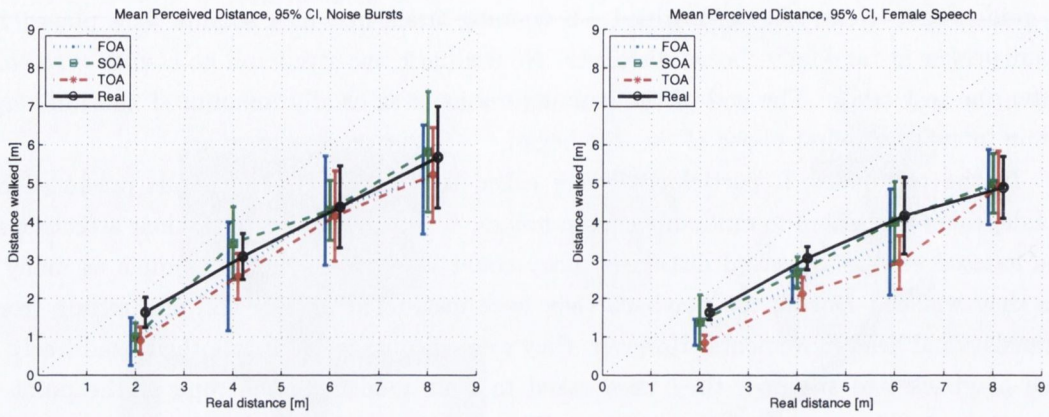
In the test phase I, participants were asked to listen to static sounds produced by the loudspeaker translated to randomly chosen points. It was emphasised that their attention should be focused on the perceived distance. They could listen to any audio sample as many times as they wished. During the playback they were instructed to stay still and refrain from any translational head movements. However, they were encouraged to rotate their head freely. After the playback had stopped, they were asked to walk guided by the rope to the point where they thought the sound originated from. The distance walked was subsequently recorded by the assistant using a laser measuring tool and participants walked back to the origin. In the meantime, the loudspeaker was noiselessly translated to its new position and the test proceeded. Similarly to the training session, no feedback was given at any stage.

During the first test phase participants had to indicate the perceived distance for sound sources randomly located at 2m, 4m, 6m and 8m. Taking into account that both speech and pink noise bursts samples were used (in a pseudo-random order), the number of test items in the first phase added up to 8. All the subjects performed all the trials only once.

Upon completion of the first phase of the test there was a short (approx. 2min) interval that was required to put on the headphones and calibrate the head-tracking system. In the phase II, subjects were also asked to identify the sound source distance but this time using Ambisonic sound fields presented over headphones. Beside the fact that the headphones and the head-tracking system was used, the test protocol remained the same as in the phase I. However, due to the fact that there were three playback configurations to be tested (1^{st}, 2^{nd} and 3^{rd} order Ambisonics), participants had to perform 24 trials instead of 8. Again, subjects performed all the trials only once and no feedback was given at any stage.

Similarly as in the pilot experiment and prior to this test phase, first order ambisonic impulse response measurements were taken from the listener position for each loudspeaker distance using the exponential sine swept tone method [59]. From these measurements, 2^{nd} and 3^{rd} order impulse response sets were extracted using the directional analysis approach outlined in Section 3.4.5. 0^{th} order Ambisonics, or the isolated "W" component of a sound field, does not provide any directional information. It means that the renderings would lack the cues that are investigated in the higher order renderings. Therefore, it was decided not to include them in this comparison.

Again, the only psychoacoustical optimisation applied to Ambisonics decodes was shelf filtering and was intended to satisfy Gerzon localisation criteria for maximized velocity decode at low frequencies and energy decode at higher frequencies [79]. The settings used for first an higher order components were exactly the same as in the Pilot Experiment (Section 5.1.2.4). Again, and as will be explained in more detail in Discussion section, no near-field compensation



(a) (b)

Figure 5.14: Mean distance perception of real and virtual sound sources within 95% confidence intervals: (a) Pink noise bursts; (b) Female speech

filters were employed.

5.2.3 Results

The perceived sound source distance (indicated by the distance walked) was collected from 7 subjects for 4 presentation points (2m, 4m, 6m and 8m), two stimuli (pink noise bursts and female speech) and four playback options (First Order Ambisonics (FOA), Second Order Ambisonics (SOA), Third Order Ambisonics (TOA) and Real Loudspeaker (Real)). With headphone trials, none of the participants reported in-head localisation, however there were 3 cases were the proximity of the sound source was very apparent so participants decided not to move at all. In some cases, the virtual sound source was initially localised behind the subjects but all participant were able to resolve the confusion by applying head rotation.

The mean values of walked distances μ for each test condition along with the corresponding standard errors $se(\mu)$ has been computed. The results are presented separately for each stimulus type within 95% Confidence Intervals.

As expected, the perception of distance for the real sources was more accurate for near sources. Beyond 4m, gauged distance was continuously underestimated which is congruent with the previous studies outlined in Chapter 2. The mean localisation of the virtual sources follows the reference (Real) source localisation well. The answers for the virtual sources deviate from their means roughly in the same fashion as the answers for reference sources, as localisation becomes more difficult within the diffuse field. Furthermore, the standard deviation of localisation increases as the source moves further into the diffuse field.

Again, the study followed similar experimental design as previously (2 (stimuli) * 4 (playback conditions)), so in order to investigate the effects of these two factors as well as potential

Table 5.2: Mean localisation [m] of virtual and real sound sources at 2m

	μ_{FOA}	μ_{SOA}	μ_{TOA}	μ_{Real}
Speech	1.119	1.389	0.841	1.638
Noise	0.877	1.001	0.902	1.641

interaction effects, for each presentation distance a two-way ANOVA has been performed. The null hypothesis being tested here is that all the mean perceived distances for all the stimuli and playback methods are equal:

$$H_0\text{: } \mu_{FOA}=\mu_{SOA}=\mu_{TOA}=\mu_{Real}=\mu$$

$$H_1\text{: not all localization means } (\mu_i) \text{ are the same}$$

No statistically significant effect of stimuli (familiar versus unfamiliar) on the perception of distance has been found ($F_{2m}(3,48) = 0.835$, $p = 0.365$; $F_{4m}(3,48) = 2.0462$, $p = 0.159$; $F_{6m}(3,48) = 2.575$, $p = 0.115$; $F_{8m}(3,48) = 2.0462$, $p = 0.159$). For distances of 4 m and more, playback option had also no statistically significant effect ($F_{4m}(3,48) = 2.192$, $p = 0.101$; $F_{6m}(3,48) = 0.665$, $p = 0.577$; $F_{8m}(3,48) = 0.202$, $p = 0.894$).

However, a statistically significant difference has been detected for the closest distance of 2m. At this point it should be mentioned that in larger study designs with multiple levels it is advisable to use the *Honestly Significant Difference* (HSD) approach since there is an increased risk of spuriously significant difference arisen purely by chance [159]. So, in order to investigate further if and where the difference occurs, the HSD has been computed which yielded $HSD = 1.423m$. If we now compile a table of mean perceived distances for the sound sources located at 2m (Table 5.2) we can see that all of the above values clearly lie within a single HSD to each other and cannot be distinguished. That is why the ANOVA false alarm (type I error) can safely be assumed and no statistically significant effect of playback method for the sources at the distance of 2m has been detected either.

Lastly, for all the distances no synergetic effects of factors A (stimuli) and B (playback conditions) have been detected.

Additionally, correlation coefficients ρ for pairs of distance estimations for real and virtual sound sources (either 1st, 2nd or 3rd order) and two stimuli have been calculated. In all cases, high correlation coefficients have been obtained, which confirms the findings that for these particular test conditions, the perception of distance of binaurally rendered Ambisonic sound fields of orders 1 to 3 cannot be distinguished from the perception of distance of the real sound sources.

Table 5.3: Correlation coefficients ρ and corresponding $p-values$ for pairs of distance estimations for real and virtual sound sources (Speech)

	ρ	$p-value$
Real vs FOA	0.9828	0.0172
Real vs SOA	0.9960	0.0040
Real vs TOA	0.9590	0.0410

Table 5.4: Correlation coefficients ρ and corresponding $p-values$ for pairs of distance estimations for real and virtual sound sources (Noise)

	ρ	$p-value$
Real vs FOA	0.9913	0.0087
Real vs SOA	0.9857	0.0143
Real vs TOA	0.9972	0.0028

5.2.4 Discussion

The results presented for real sources corroborate the classic underestimation of source distance, as reported in the literature (see Chapter 2 Section 2.3). However, they do not follow the same pattern as the results obtained in the pilot study. Specifically, greater underestimation of the source distances has been obtained here, especially for sound sources located at and beyond 4m. This can be explained by differences in the acoustics of both test environments and in particular their reverberation times. It has been shown by Bronkhorst et al. [32] that in damped environments sound sources appear to be closer than in reverberant spaces.

The results obtained for the real sound sources were used as a basis with which to measure the ability of Ambisonic sound fields of different orders to present sources at different distances. It was expected that a further underestimation of the source distance would ensue with the binaural rendering, as reported by Chan et al. in [37]. However, this was not the case, even for first order presentations, and the apparent distances of the virtual sources matched the real source distances well. One should note that the major difference between this study and that of Chan et al. [37] is the use of head-tracking, indicating the importance of head-movements in perceiving source distance, which develops the findings of Waller [226] and Ashmead et al. [13] on user interaction in a virtual space. Further work is required to quantify the effect of this.

Moreover, the presented study demonstrates that the enhanced directional accuracy gained by presenting sound sources in HOA through a head-tracked binaural rendering does not yield a significant improvement in the perception of the source distance. What is noteworthy is that

for each order, there is no significant difference in the perception of the source location when compared to real-world sources. It can be therefore concluded that sound field directionality for distance perception is sufficient with 1^{st} order playback.

The presence of the ANOVA false alarm at the 2m point is of interest. It is noteworthy that the 2m point represents a source inside the virtual array geometry. It is a known issue that virtual sound sources rendered inside the array of loudspeakers cannot be reproduced in a straightforward way without artifacts. Some of these artifacts include incorrect wave-front curvature and insufficient bass boost.

In the first case, there is ample evidence in the literature to suggest that the wave front curvature translates to a significant binaural cues for sound sources near the head [239, 240]. It was already shown in Section 2.3 that as the source moves closer to the head the levels of the monaural transfer function and the ILD both change significantly with the source angle. However, this effect is not strong beyond 1m. For sources further away, it has been shown by Wittek in [237] that it is very difficult to assess distance by binaural cues alone.

In the second case, the requirement for distance compensation filtering due to near field effects for the large loudspeaker array radius (3.27m) and the given source distances ($\geq 2m$) is only prominent below 100Hz as shown in Figure 5.15. Thus, for the female speech test stimuli, this will not have an effect since the first formant frequencies do not go down below 180Hz. Also, the current method employed for capturing HRIRs allowed for reliably obtaining filters with a frequency response reaching down to around 170Hz, thereby also band-limiting the delivery of the pink noise stimuli.

Finally, there was no significant difference in the results presented for different sources, although the greater variance in the results for pink noise suggests that the familiarity of the source does indeed play a role in the perception of source distance, as mentioned in Section 2.3. The analysis of residuals presented in Figure 5.16 indicates that the unfamiliar stimulus produces slightly higher variability in subjects' answers.

It has to be noted though that the order of test phases could potentially have some effect on the results, although it should not affect similarities in the reports for different orders of virtual sources. In [242] Ziemer et al. point out that when participants experienced the real environment first they tended to underestimate less the distance in the virtual environment than they normally would. They add however, that this difference is rather subtle.

Future studies should investigate the use of these monaural cues further, and will utilise 0^{th} order sound field rendering, since it will remove the influence of any directional information. Considering the aforementioned study of Bronkhorst et al. [32], where the accuracy of distance perception for binaural playback increases with the number of reflections, these findings demonstrate that the net effect of the monaural cues of direct-to reverberant-ratio, level difference and time of arrival of early reflections are of greater importance in distance perception for binaural rendering than Ambisonic directional accuracy beyond 1^{st} order.

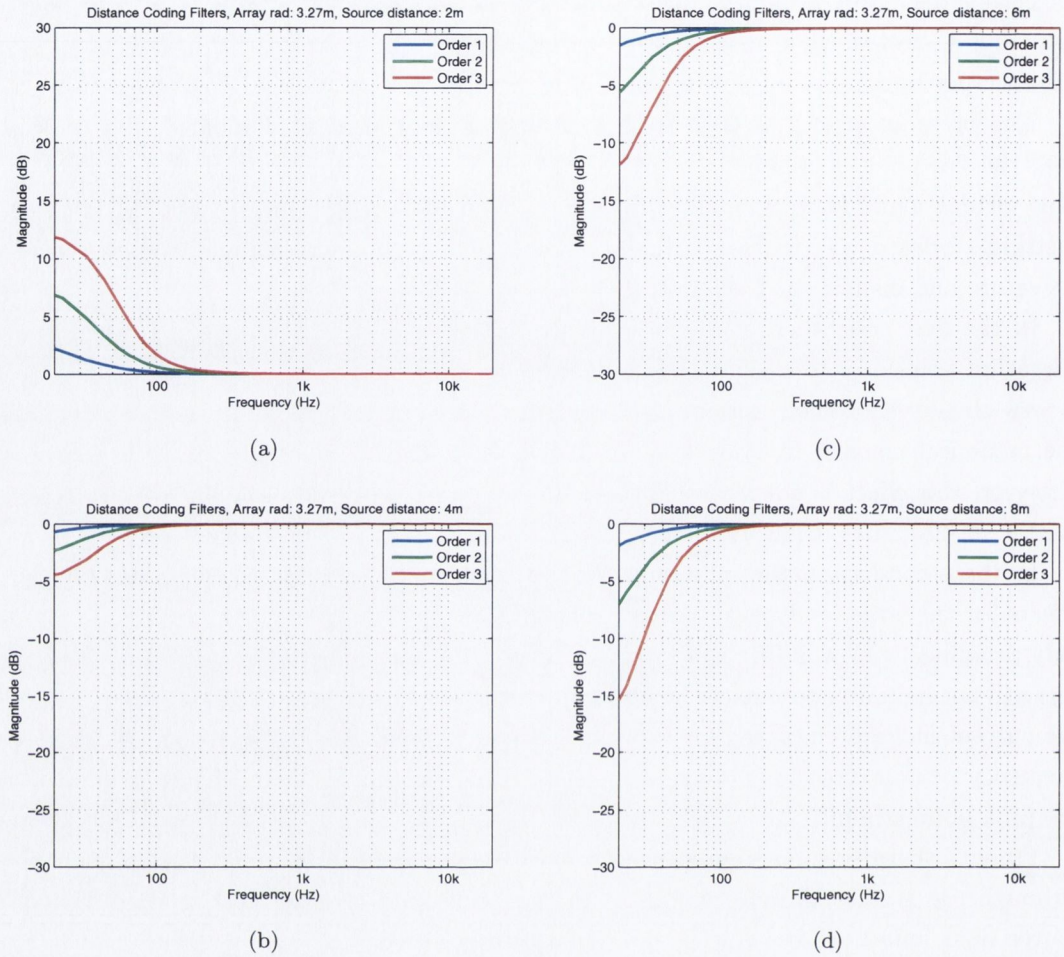
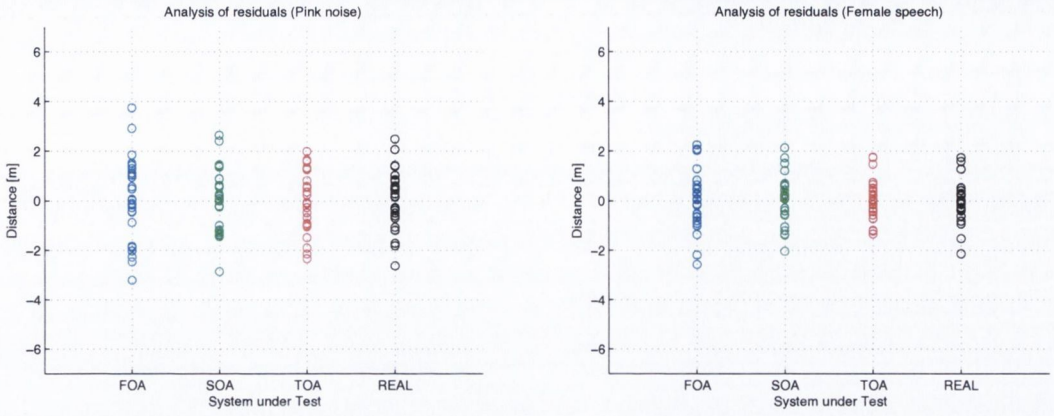


Figure 5.15: Distance coding filters required for direct sound compensation up to 3^{rd} Order Ambisonics in the Experiment I(a). Compensation filters are obtained for loudspeaker array radius of 3.27m and sound source distance of: (a) 2m; (b) 4m; (c) 6m; (d) 8m.

5.2.5 Conclusions

The perceived source distance in virtual Ambisonic sound fields as compared to the real world sources has been assessed through subjective analysis. The hypothesis tested was that enhanced directional accuracy of deterministic part of the sound field may lead to better reconstruction of environmental depth and thus improve the perception of sound source distance. However, it was shown that Ambisonic reproduction matches the perceived real-world source distances well even at lowest orders and no improvement in this regard was observed when increasing the order. These results are also compatible with the results obtained in the preliminary study as described in Section 5.1.

It must be emphasised though, that this analysis applies to Ambisonic-to-binaural decodes



(a) (b)

Figure 5.16: Analysis of residuals for: (a) Pink noise bursts stimulus; (b) Female speech stimulus. On the x-axis are four systems under test. On the y-axis there are differences between the mean perceived distance for all the test points and individual distance estimates from 7 subjects.

with higher order synthesis achieved using the directional analysis method of [144]. Therefore, further work should investigate the effectiveness of HOA synthesis in comparison to real-world HOA measurements should HOA microphones become widely commercially available. However, the fact that even 1^{st} order renderings cannot be distinguished from the real sources when it comes to distance judgements, allows us to predict that recordings of the Higher Order sound fields with a specialised microphones would be of no advantage in this regard as well.

The results obtained using the virtual loudspeaker approach should also be compared against the real, physical loudspeaker renderings. It would require additional considerations mainly with regards to the loudspeaker configurations and listening conditions. This topic will be discussed in next.

5.3 Experiment I(b): Distance Perception in Real and Virtual Environments using Loudspeaker Based System

5.3.1 Introduction

In the previous section it was shown that virtual recordings presented over headphones are capable of providing enough information from which sound source distance can be correctly inferred. It was also discussed that headphone-based renderings with virtual auditory scenes have the advantage of easier control over the personalised sound fields. It is due to the fact that the listener is already sitting in the center of (ideally) perfectly aligned array of virtual loudspeakers

and the signals generated by these virtual loudspeakers do not interact with the environment[1]. Also, individual modifications of acoustic waves due to the presence of the particular listener can also be performed if the process of obtaining the individualised HRIRs to create virtual loudspeakers has been performed.

In loudspeaker listening the situation is quite different. Acoustic waves emitted by each loudspeaker in the setup are now free to interact with the boundaries of the listening space and the resulting sound field is "contaminated" with the unwanted reflections. Although for most of the consumer uses such a situation is perfectly acceptable and in most cases some tonal colouration resulting from the comb-filtering will simply be added to recordings. The amount of colouration as well as the way it affects the spectrum of the signal often allows to classify the listening environments and help them assign different uses (like recording/mixing studios, cinemas, concert halls etc.). Larger environments, where it takes longer for the reflected acoustic energy to decay, will provide recordings with additional reverberation. If the listening space is really large though, there will be a risk that the reflected sound will arrive at the ears of the listener with a considerable delay. It will be possible to distinguish between individual reflections and discrete echoes will be heard. Such a situation is usually detrimental to the recordings, especially to the samples with recorded speech, since it can considerably degrade the intelligibility.

When it comes to perceptual testing, it is therefore desired that the acoustic conditions of the listening space are also strictly controlled. Whilst it is physically impossible to assure completely reflection-free listening environment, nonetheless there are ways of dampening and re-directing/diffusing the reflected acoustic energy so that it arrives at listener's ears considerably attenuated. Anechoic or semi-anechoic chambers are good examples of acoustically controlled listening/measuring environments.

The following experiment looks at the problem of distance perception in the loudspeaker listening. More precisely, headphone renderings of the second phase of the previous experiment test are replaced with analogous loudspeaker renderings. There are no major changes in the methodology of the experiment besides the fact that the radius of the real loudspeaker array as well as the listening environment had changed. Also, for technical reasons, it was decided to reduce the number of tested playback systems to 1^{st} and 3^{rd} Order Ambisonics only.

The remainder of this section explains in detail the rest of the technical aspects of the experiment as well as its exact process. The results are also presented and followed by the detailed analysis and discussion.

[1]It is true if the HRIR measurements have been performed under anechoic conditions or at least if the captured HRIRs have been further processed in order to remove any possible influence of the room

5.3.2 Methodology

In the previous part of this experiment, subjects were listening to samples of pink noise bursts and female speech either using the real, physical loudspeaker or its virtual auralisation using headphones. Then, they were asked to indicate where the sound originated from by walking toward the sound source. Carrying forward the results from the first phase of the previous test (when the physical source was used), the second phase was now replaced with the loudspeaker rendering of virtual sources.

Again, participants listened to sound samples generated at some predetermined but randomly selected distances. Then, their task was to walk toward the apparent location of the virtual source. The method of generating First and Higher Order Sound Fields was exactly the same as before, however, only First and Third Order Sound Fields were incorporated into the test.

5.3.2.1 Participants

10 participants undertook this experiment (3 female and 7 male). They were mainly postgraduate research students in the field of audio and acoustics, all of good hearing. Prior to the test, they were able to familiarise themselves with the test protocol as well as the test material used and according to ITU-RBS.1284-1 [99] recommendation they could be classified as expert listeners.

5.3.2.2 Stimuli

As in the Experiment I(a), the same pink noise bursts and female speech samples were used in order to represent both unfamiliar and familiar sound sources.

5.3.2.3 Test conditions and apparatus

For the trials, a purpose built, 16 channel loudspeaker array consisting of *Genelec 8050A* loudspeakers, arranged in a sphere was utilised. The array was housed in the Audio Lab at the Department of Electronics, University of York. The geometry of the array was an exact replica of the array used in the previous part of the test (Figure 5.10): 16 loudspeakers arranged in 3 tiers of 4 (bottom), 8 (middle) and 4 (top) loudspeakers respectively.

Again, the array used was not perfectly regular with the undersampling of loudspeakers at the top and bottom of the array. That is why, sound field reconstructions could potentially suffer from errors, especially when the sound source was located above or below the listener. However, because of the technical (the number of available sound card channels and loudspeakers) and space limitations (it would be difficult to place a loudspeaker e.g. below the listener), it was decided not to add additional channels. It must also be noted that all the virtual sources used in the test were rendered in front of the listener, at the hight of the horizontal ring of loudspeakers where the sound field was not undersampled which could minimise the potential

rendering artefacts.

Also, due to the space limitations of the current listening room, radius of the array was reduced to 1.9m. 1^{st} Order sound fields were reproduced over the inner cube of the array (bottom 4 and top 4 loudspeakers) and 3^{rd} Order sound fields were rendered over the full array. Because of the fact this setup could not be easily rearranged in order to accommodate the 2^{nd} order Ambisonic renderings as well as it was more difficult to ensure that listeners are fixed exactly at the sweet-spot, it was decided not to include the 2^{nd} order samples in this comparison.

As outlined in the introduction, acoustic conditions of the test were now different. Due to the technical difficulties with setting up such an expanded loudspeaker array safely in an anechoic chamber, acoustically treated and damped listening room was used for the tests instead. Acoustic absorbers and damping curtains were used to minimise the impact of the first early reflections in the listening space. Effectively, low reverberation time of 0.26s at 1kHz was achieved.

5.3.2.4 Procedure

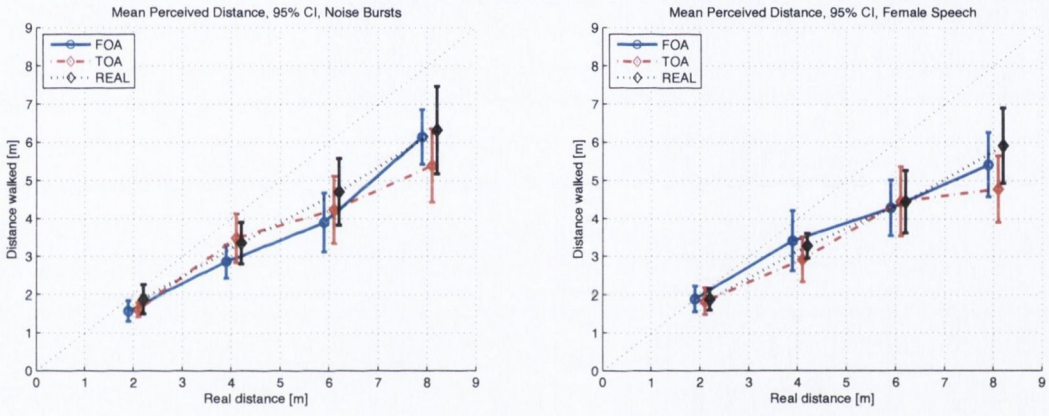
As before, subjects entered the test environment blindfolded and without any prior expectation regarding the room dimensions, its acoustic properties or the test apparatus. They were guided by the experimenter to the listening point which was at the centre of the loudspeaker array (sweet-spot) and which will be referred to later as the "origin". After a short explanation of the experiment objectives, a training session began with a short (3-5min) walking-only trial until participants felt comfortable with walking blindfolded and using a guide rope. Next, they performed 4-6 training trials in which the same test stimuli to be used in the experiment (speech and pink noise) were played by the loudspeakers at randomly chosen distances. No feedback was given and no results were recorded after each test trial. The end of the training session was clearly announced and after a 1 minute interval, the main phase of the test began.

In the proper part of the test, subjects were asked to indicate the perceived distance of the virtual sound sources randomly located at 2m, 4m, 6m or 8m. Virtual sources were rendered using Ambisonic sound fields and this time presented over the loudspeaker array. Other than the fact that the loudspeakers were used instead of headphones, and the reproduction environment was now changed to a dedicated listening room, the test protocol remained exactly the same as in the Experiment I(a). However, due to the fact that there were two playback configurations to be tested (1^{st} and 3^{rd} order Ambisonics), participants had to perform 16 trials instead of 24. All the samples were presented to the subjects randomly in order to negate any ordering effects. Again, subjects performed all the trials only once and no feedback was given at any stage.

5.3.3 Results

The perceived sound source distance (indicated by the distance walked) was collected for each subject for 4 presentation points (2m, 4m, 6m and 8m), two stimuli (female speech and pink noise bursts) and two playback options (First Order Ambisonics (FOA) and Third Order Ambisonics

Figure 5.17: Participant performing a trial during the experiment



(a) (b)

Figure 5.18: Mean distance perception of real and virtual sound sources within 95% confidence intervals: (a) Pink noise bursts; (b) Female speech

(TOA)). These results have been compared against the performance of the real sound sources from the part (a) of this experiment. The mean values of walked distances μ for each test condition along with the corresponding standard errors $se(\mu)$ have been computed. The results are presented in Figure 5.18 separately for each stimulus type within 95% Confidence Intervals.

Again, the perception of distance was more accurate for near sources. Beyond 4m, distance perception was continuously underestimated in all cases, which is congruent with the previous studies outlined in Chapter 2. Furthermore, the standard deviation of localisation generally increases as the source moves further into the diffuse field.

Since the study followed the within-subject factorial design with 2(stimuli) and 3(playback conditions), in order to investigate the effects of these two factors (referred later as factors A and B) as well as potential interaction effects, for each presentation distance a two-way ANOVA has

been performed. The null hypothesis being tested here is that all the mean perceived distances for all the stimuli and playback methods are equal:

$$H_0\text{: } \mu_{FOA}=\mu_{TOA}=\mu_{Real}=\mu$$

$$H_1\text{: not all localization means } (\mu_i) \text{ are the same}$$

No statistically significant effect of stimuli (noise versus speech) on the perception of distance has been found ($F_{2m}(2,54) = 1.8$, $p = 0.19$; $F_{4m}(2,54) = 0.01$, $p = 0.94$; $F_{6m}(2,54) = 0.13$, $p = 0.72$; $F_{8m}(2,54) = 2.28$, $p = 0.14$).

Also, performance of two tested playback methods did differ significantly from the perception of the real, reference sound sources ($F_{2m}(2,54) = 0.57$, $p = 0.57$; $F_{4m}(2,54) = 0.2$, $p = 0.82$; $F_{6m}(2,54) = 0.66$, $p = 0.52$; $F_{8m}(2,54) = 2.41$, $p = 0.1$).

5.3.4 Discussion

The results presented for real sources corroborate the classic underestimation of source distance, as reported in the literature. These results were used as a basis with which to measure the ability of Ambisonic sound fields of different orders to present sources at different distances. It was expected that a further underestimation of the source distance would ensue with the virtual source rendering, as reported in [37]. However, this was not the case, even for first order presentations, and the apparent distances of the virtual sources matched the real source distances well.

One should also note that distance compensation filtering (due to near field effects) was not implemented in this study. This is because the combination of the array radius (1.9m) and the source distances ($\geq 2m$) leads to near field effects only prominent below 100Hz. Appropriate distance coding filters are shown in Figure 5.19 up to 3rd Order. For the female speech test stimuli, such filtering was unnecessary, since the first formant frequencies do not go down below 180Hz. Furthermore, such filtering would only be relevant for the direct sound component, whose distance is known and not the early reflections. For these reasons pink noise delivery was also band-limited to 100Hz.

Moreover the presented study demonstrates that the enhanced directional accuracy gained by presenting sound sources in HOA over loudspeakers does not yield a significant improvement in the perception of the source distance. What is noteworthy is that for low and high orders, there is no significant difference in the perception of the source location when compared to real-world sources. It can therefore be concluded that sound field directionality for distance perception is sufficient with 1^{st} order playback.

Finally, there was no significant difference in the results presented for different sources, Interestingly, the difference in variability of judgements for noise and speech stimuli was less apparent this time, as visible in Figure 5.20). The variance was also in general slightly smaller that in the case of headphone rendered sound fields.

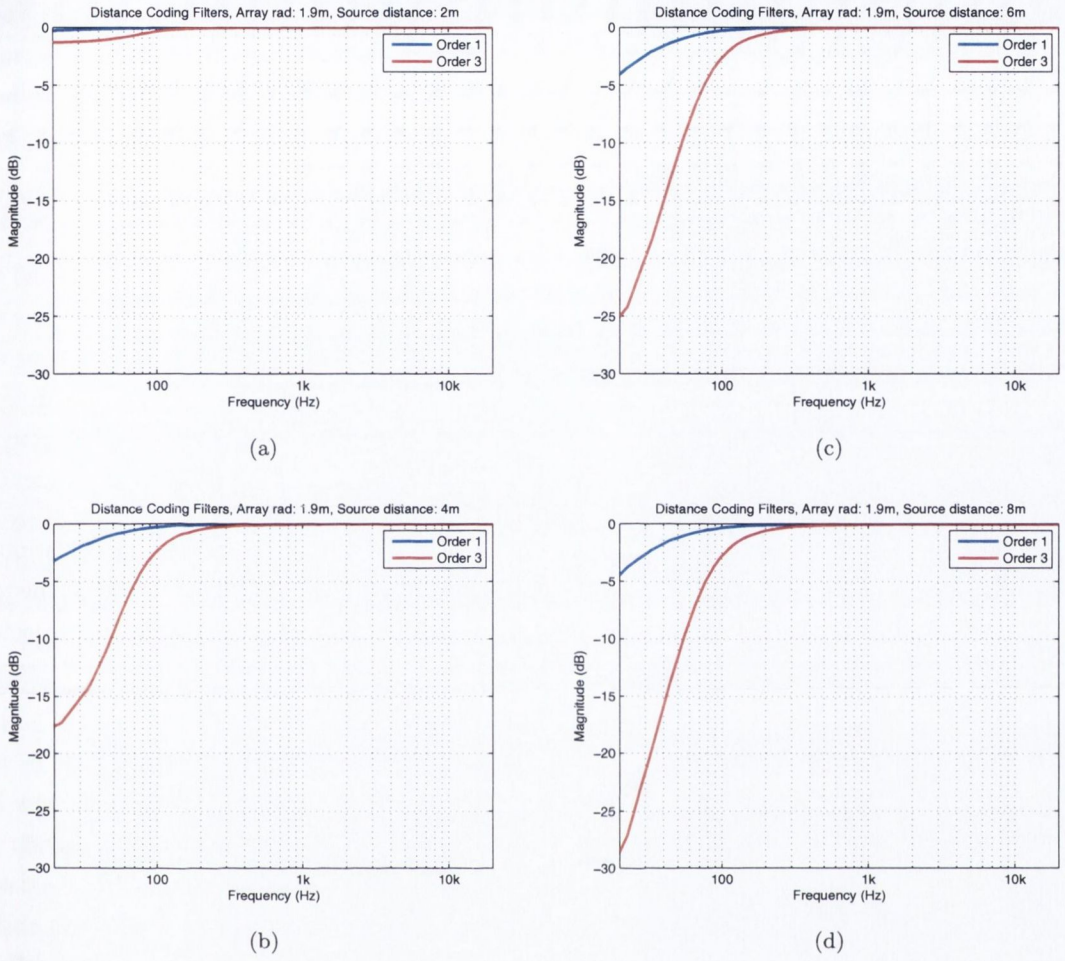
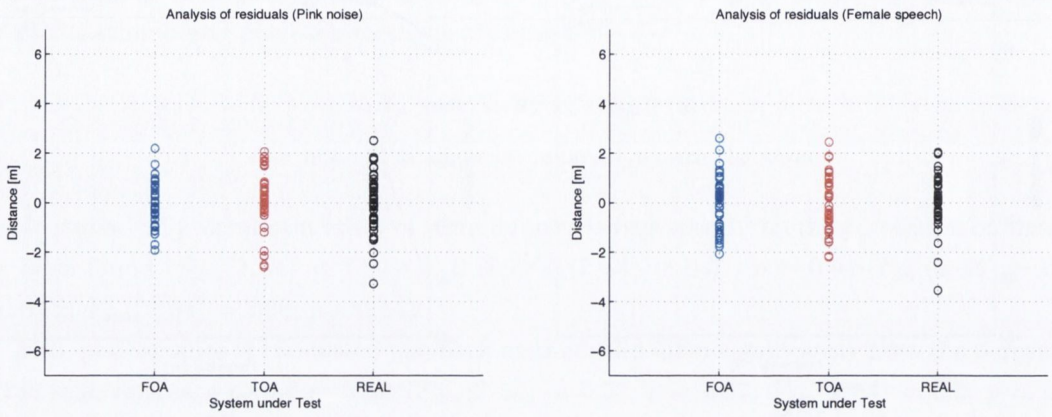


Figure 5.19: Distance coding filters required for direct sound compensation up to 3^{rd} Order Ambisonics in the Experiment I(b). Compensation filters are obtained for loudspeaker array radius of 1.9m and sound source distance of: (a) 2m; (b) 4m; (c) 6m; (d) 8m.

Considering the aforementioned study of Bronkhorst et al. [32], where the accuracy of distance perception increases with the number of reflections, these findings demonstrate that the net effect of the monaural cues of direct to reverberant ratio, level difference and time of arrival of early reflections are of greater importance in distance perception for loudspeaker rendering than Ambisonic directional accuracy beyond 1^{st} order. Future studies will investigate the use of these monaural cues further, and will utilise 0^{th} order sound field rendering, since it will remove the influence of any directional information.



(a) (b)

Figure 5.20: Analysis of residuals for: (a) Pink noise bursts stimulus; (b) Female speech stimulus. On the x-axis are three systems under test. On the y-axis there are differences between the mean perceived distance for all the test points and individual distance estimates from 10 subjects.

5.4 Conclusions

Investigations into perception of distance using Ambisonics binaural technology have now been complemented with loudspeaker reproduction. In all cases it was experimentally proven that recreating distance cues in virtualised sound fields is possible. More importantly there exists no significant difference in the mean perceived distance between the Ambisonic renderings and the real sources. This is true for both first and higher order Ambisonics, as well as across different source types (familiar and unfamiliar).

It must be noted though that in all studied cases, different apparatus requirements forced the tests to be conducted in two separate stages: the first stage where the real sound sources were used; and the second stage where the virtual sound sources were used. Such a situation is not ideal from the statistical point of view because possible learning effects cannot be fully eliminated. In other words, using real sound sources may have potentially resulted in training the subject that subsequently reflected on the perception of the virtual sources. Further work, in which the binaural testing is carried out first, would help to resolve whether the ordering effect on the perception of real and virtual sources was significant.

5.5 Experiment II: Distance Perception in Virtual Audio-Visual Environments

Spatial localisation of sounding objects is affected not only by auditory cues but also by other modalities such as vision. This is true particularly in the context of distance perception where

the number of auditory cues is limited in comparison to e.g. localisation in horizontal and vertical planes. This experiment was designed to investigate the influence of vision on the perception of audio. In particular, the effect of incongruent audio-visual cues was explored in the context of the perception of auditory distance in photo-realistic Virtual Reality Environments (VREs).

5.5.1 Introduction

So far in this work of the primary interest was how humans perceive distance when presented with an isolated auditory stimulus. In experiments I(a) and I(b) it was even desired that all the other sensory information is completely removed from the test protocol. However, multimodal sensory presentations are now commonplace in many aspects of everyday life. Cinematic or gaming experiences are probably two of the most popular examples. In some of these applications, the intention is to immerse the user so that the perceived virtual reality experience is equivalent to some real life experience. In order to achieve the best effect, the sensory stimuli must obey certain rules and conditions in order be physically correct or at least perceived as equivalent to the real world experience.

For instance, in audio presentations the goal is to recreate pressure variations at ear canals (using loudspeakers or headphones) that will be equivalent to those that would have occurred in a real acoustic field. This can be achieved by e.g. filtering audio material with an individualised set of filters (i.e. HRTFs) that fully describe the acoustic path from a source to the listener's ears [236], as discussed earlier in Chapter 2, Section 2.1.3.

From the visual point of view, a key challenge is to set up the stereoscopic images so that the perceived size and relative positions of all the visible objects match the recreated, authentic scene. However, vision has been proven in multiple studies to dominate auditory cues [37, 217]. A famous example is also that of the so-called "ventriloquist effect" [225]. It is hypothesised then, that if the visual dominance affects the perception of auditory distance then there must exist some margins of audio-visual incongruence within which the scene is still perceived as consistent.

In this study the above hypothesis is tested by the means of a subjective listening test where the visual stimulus is accompanied by audio whose distance is either the same or deviates from the visual distance. For each presentation, the proportion of the population able to detect the incongruence is estimated.

5.5.2 Visual Distance Perception

The perception of depth when viewing stereo images is related to the relative horizontal positions of a point between the left and right images, known as the stereo disparity. However, it is not sufficient to present the images as captured by a stereo camera to the viewer as the different modalities of image capture and display mean that the perception of depth is altered. This can be seen in Figure 5.21 and Figure 5.22 [243]. As the left and right stereo images are overlaid on

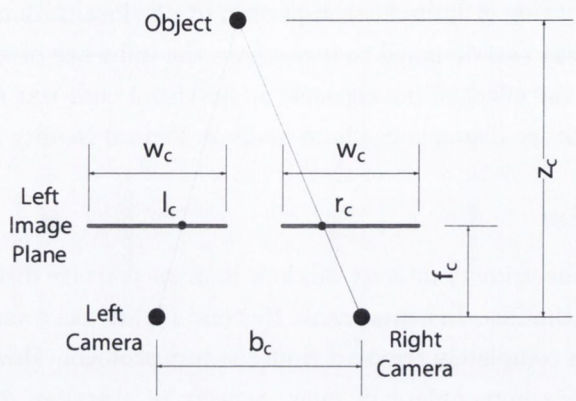


Figure 5.21: The model for Standard Rectified Geometry [218] stereoscopic images capture. The cameras are in parallel with the line through their centres being perpendicular to their orientation. In this model, an object at a distance z_c from the cameras is projected onto the left and right image planes at points l_c and r_c respectively

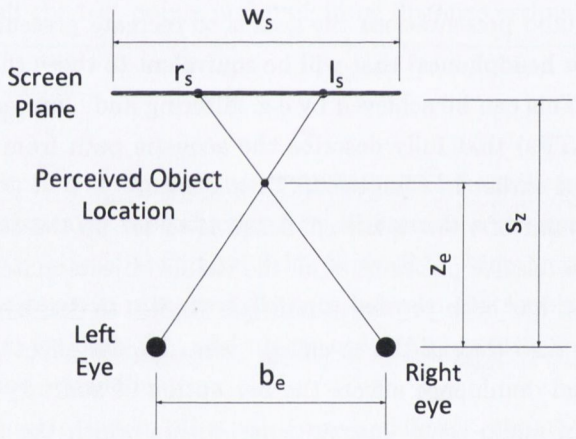


Figure 5.22: The model for depth perception when viewing stereoscopic images [243]. The perceived object location is at the intersection of two light rays between the eyes and the respective object position in the left l_s and right images r_s

top of each other on most 3-D displays, all objects in the image will have a perceived location that is in front of the screen. Further depth distortion is caused by differences between the other corresponding parameters: the camera separation b_c and eye separation b_e, the focal length f_c and viewing distance s_z, and finally the camera sensor width w_c and screen width w_s.

5.5.3 Methodology

Given a strong visual stimulus (i.e. an object, an image or a CG rendering) representing a sound source one can expect that the visual cues will dominate the perception of distance in

the audio-visual scene [37]. However, more interesting and less explored situation occurs when conflicting audio-visual cues are presented to the user, i.e. when the auditory location matches the visual component vertically and horizontally but is at the same time misaligned with it on the distance axis. The above scenario was investigated in this experiment and subjected to the perceptual evaluation. The testing was done in the context of an immersive, photo-realistic virtual environment created using stereoscopic vision and binaural sound reproduction technologies.

5.5.3.1 Participants

Twenty one subjects took part in the experiment (7 female and 14 male) and they were mainly postgraduate students and academic staff from the Department of Electronic and Electrical Engineering, Trinity College Dublin. Seven of the subjects had a background or were actively involved in audio related research and/or music production.

A larger panel of subjects was used in this experiment due to the mixture of expert and non-expert listeners as per ITU-R BS.1284-1 3 recommendation [99]. Also, because of the audio-visual nature of the experiment as well as the attempt to present the subjects with the close to real audio-visual experience, it was decided that the use of general members of the population would be beneficial.

5.5.3.2 Stimuli

In order to investigate the perceptual effect of conflicting audio-visual cues it was first necessary to faithfully recreate certain audio-visual scenes i.e. by putting them in the context of an immersive, virtual reality application. The chosen scenes consisted of stereoscopic images of a small lecture hall with a *Genelec 1029A* loudspeaker on a stand representing a sound source and accompanying audio stimulus. No other background sounds were used.

5.5.3.3 Visuals

The stereo images used in the experiment were obtained by automated translation of a camera along an axis perpendicular to the speaker axis. Using this setup rather than a conventional stereo camera rig avoids colour distortion and vertical disparity artefacts [206] and allows the eye separation distance to be adjusted and varied after image acquisition. The camera used to capture the images was a *Sony PMW-EX3* recording at a resolution of 1920 x 1080 pixels and a frame rate of 25 frames per second. A 35mm equivalent focal length (36mm sensor width) of 31mm was employed.

A two-fold solution to correct the perceived depth has been adopted. To allow objects to be perceived behind the screen the centre of the images have been offset by the amount of pixels corresponding to the eye separation b_e, with the left image shifted $b_e/2$ to the left and the right images shifted $b_e/2$ to the right. The remaining depth distortion has been fixed as follows. The

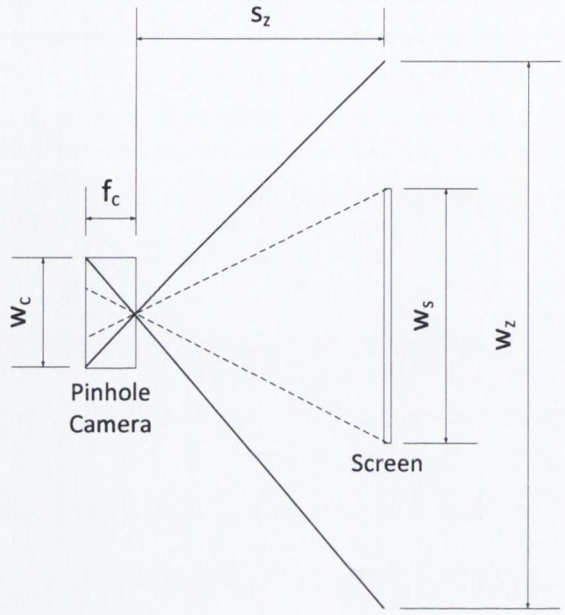


Figure 5.23: An illustration of how the crop ratio is determined. At the viewing distance the extent of the image field of view is greater than the screen width. Each image is cropped so that the field of view is the area formed between the dashed lines

camera separation was chosen to be equal to the eye separation (i.e. $b_c = b_e$). Secondly, the ratio of screen width to viewing distance has been forced to match the ratio of sensor width to focal length (i.e. $w_s/s_z = w_c/f_c$). This was achieved by cropping the image by a factor of the ratio of screen width to field of view extent w_s/w_z as depicted in Figure 5.23 and reduced the effective camera sensor width w_s. After these corrections had been applied, the presentation of the stereo image to the viewer was matched geometrically with the scene, including along the vertical and horizontal axes as well as the depth axis.

Before the images were presented to the viewer, they had been cropped as described above. A fixed eye baseline of 62mm had been defined for each viewer which is approximately the mean eye separation for adults [52]. Images 31mm to the left and right of the range midpoint were used to form the stereo pairs presented to the viewer. Furthermore, the viewing distance was chosen to be 2m. Consequently, after cropping the horizontal resolution of the images was 860 pixels. Since the display had a 16:9 aspect ratio, the vertical resolution was 484 pixels. The offset of the centres of the left and right images was achieved by translating the position of the crop window in the images horizontally. Thus, to translate the left image leftward by $b_e/2$, the centre of the window was taken 26 pixels to the right of the image centre. Using a similar argument, the centre of the crop window for the right image was moved 26 pixels to the left. After cropping the images were scaled to fit the screen area. Finally, the left and right images were presented together as a composite image, with the top half of the composite being the left

image and the bottom half being the right image. This required reducing the vertical resolution of the images by a factor of 2 and was a standard format for stereoscopic images on 3-D TVs.

5.5.3.4 Audio test signals

Prior to the test phase, B-format spatial room impulse response (SRIR) measurements of different loudspeaker locations were taken from the listener position using the exponentially swept-sine tone technique [59]. *Soundfield MKV* microphone system was used for all the recordings. From the large dataset of captured SRIRs it was decided to select a subset of the 11 measurements taken at 1, 1.5, 2, 2.5, 3, 4, 5, 6, 7, 8, 9m distances and incorporate them into the test.

Again, two different dry audio samples were used in the experiment: pink noise bursts and phonetically balanced phrases selected from the TIMIT Acoustic-Phonetic Continuous Speech Corpus database and recorded by a female reader [64]. These two sample types were selected in order to represent both unfamiliar and familiar sound sources. A sampling rate of 44.1kHz and 16bit resolution was used in both cases.

Using a convolution of dry audio samples with captured B-Format SRIRs, Ambisonic sound fields were generated. Then, sound fields of 1^{st} and 0^{th} (only 'W' component of the sound field) were decoded for binaural presentation over headphones using virtual loudspeaker approach as described in the Section 3.5.1. Regular octagonal geometry was used as the default loudspeaker layout so this time the renderings were lacking a vertical component of the sound field. Each virtual loudspeaker feed was convolved with the corresponding HRIR filter pair at runtime. Only one set of filters was used for all the subjects and was obtained from the TU Berlin on-line database [233] (measured at 1m distance and with no headphone equalisation).

5.5.3.5 Test conditions and apparatus

The test was conducted in a small, dark listening room. The monitor used to display the stereo images was an *LG 47inch 3-D TV* that has a screen width of 104cm. It uses circular polarising filters to separate the left and right images for each eye [121]. The participants were required to wear glasses with matching polarising filters. This technology has the advantage of full colour resolution (as opposed to anaglyph which uses shades of two, usually chromatically opposite colours, like red and cyan) and does not exhibit the flickering associated with using the active shutter glasses used with many 3-D TV sets [137].

Audio signals were presented binaurally. High-quality, matched, open-back headphones (*Sennheiser HD-650*) in conjunction with an infra-red head-tracking device (*NaturalPoint TrackIR 5* [161]) assured trustful and stable reproduction of acoustic images.

5.5.3.6 Procedure

In the main experiment phase, participants were simultaneously presented with a stereoscopic image (being a visual representation of a sound source) and an accompanying auditory stimulus

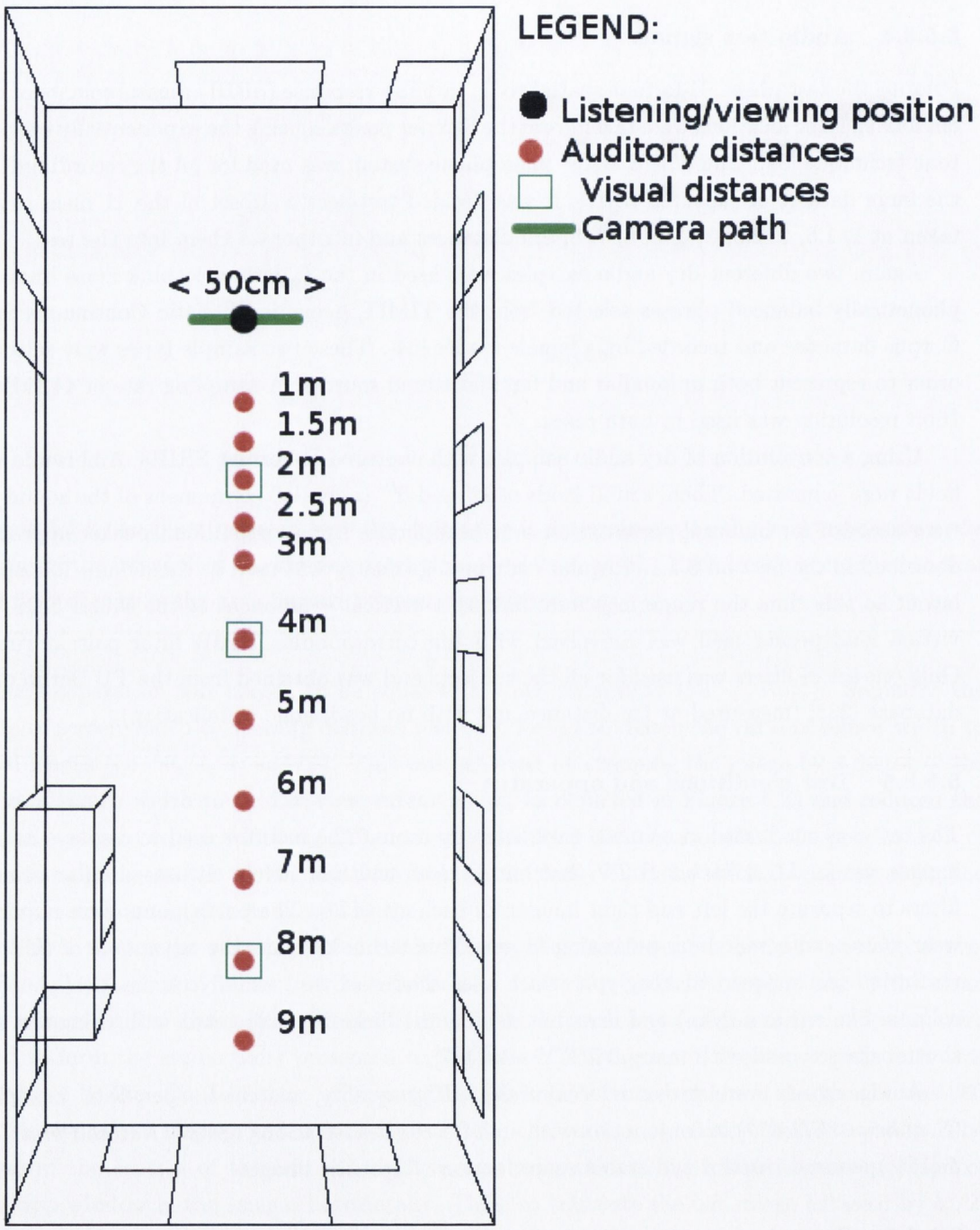


Figure 5.24: Recreated test environment with marked positions of auditory and visual measurements

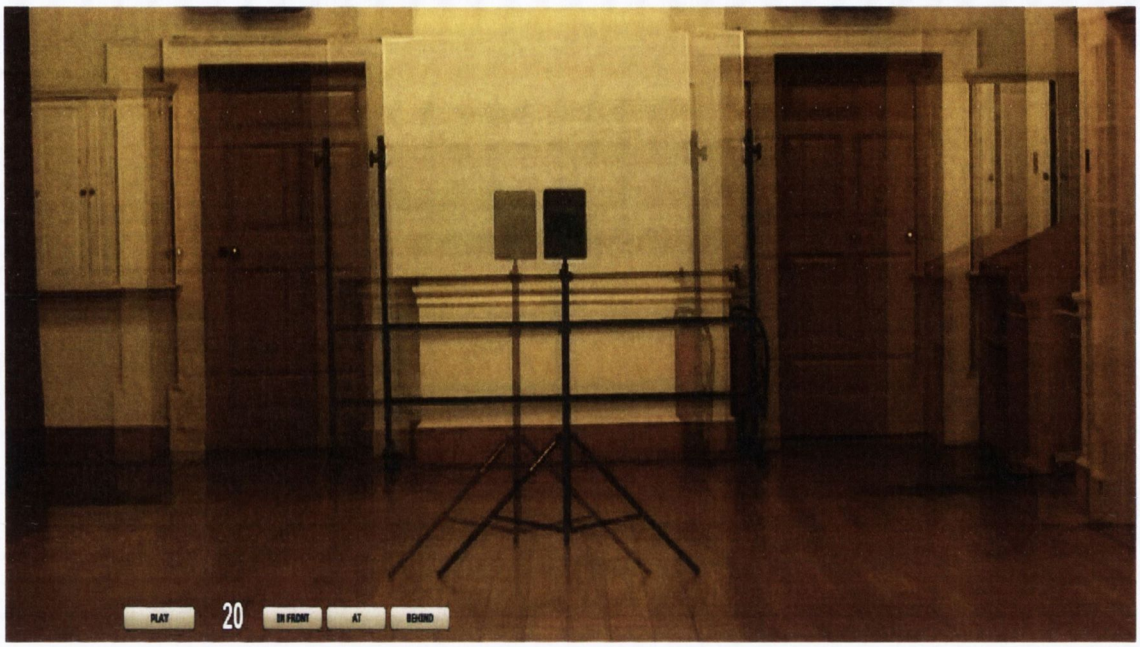


Figure 5.25: Left and right eye views of the stereoscopic image overlaid on top of each other showing a loudspeaker at 4m from the user. Simple GUI was overlaid on top of the image(s) in order to provide means of navigating through the test and saving the results

(head-tracked, binaural presentation of either pink noise bursts or female speech). Visual images were randomly presented at distances of 2, 4 and 8 meters. At the same time, the accompanying audio stimulus matched the location of the visual speaker horizontally and vertically but was randomly misaligned with it on the distance axis. After each presentation, users were asked to evaluate the spatial location of audio with respect to the visuals (e.g. *in front of*, *at the same location* or *behind* the visual loudspeaker). A similar protocol was previously used by Werner et al. in [231] to study the influence of visual feedback on vertical sound source location.

For, each visual distance, two different audio samples (speech and noise) were presented at 11 different distances (1, 1.5, 2, 2.5, 3, 4, 5, 6, 7, 8 and 9m) adding up for a total of 132 different combinations. Presentation of test items was fully randomised in order to negate any ordering effects. No feedback was given to participants at any stage. However, prior to the main test phase, each participant completed a short training session with 12 examples where the information about the relative positioning of audio and visuals was given *a priori*. During both training and main test sessions subjects were free listen to each of the test items as many times as they wished. When they finally made their decision, they used a simple Graphical User Interface (GUI) presented in Figure 5.25 overlaid on top of the images in order to save their answer and move to the next test item.

To avoid any audio/video artefacts related to uncompensated translational head movements

Table 5.5: Lower and upper bisection points for 'same distance' curves and a 50% normalised frequency threshold

Audio	Visual Distance		
	2m	4m	8m
1st Order	<1m – 3.07m	1.31m – 5.65m	6.13m - >9m
0th Order	<1m – 1.45m	1.58m – 2.41m	2.75m - >9m

Table 5.6: Lower and upper bounds of allowed audiovisual distance misalignment based on a minimum of 50% of subjects who regarded the auditory source as equidistant with its visual counterpart. * indicates that lower/upper bounds failed to be detected.

Audio	Visual Distance		
	2m	4m	8m
1st Order	*/+1.07m	-2.69m/+1.65m (4.34m)	-1.87m/*
0th Order	*/-0.55m	-2.42m/-1.59m (0.83m)	-5.25m/*

participants were instructed to refrain from translating their heads during the test. However, simple head rotations were encouraged and compensated for using the head-tracking.

5.5.4 Results

Results are presented in terms of a normalised frequency of occurrences of answers *in front of*, *at the same location* or *behind* the loudspeaker for all the combined audio stimuli (speech + noise) and various test conditions (separately for each visual loudspeaker presentation distance and spatialisation method used). Separate results for both speech and pink noise samples can be found in Appendix D. The standard error of reported percentages for the given population size ($N = 21$) is $\pm 10.9\%$ (maximum at 50%) but was omitted in the figures for clarity. From the obtained data, the lower and upper bounds of the allowed audio-visual misalignment have been inferred (Table 5.6) based on a minimum of 50% of the population that considered the auditory source as being at the same distance as the visual object.

5.5.5 Discussion

This study reveals that in many cases the incongruence of visual and auditory cues can be compensated for by human perception so that the audio-visual scene still appears to be consistent. However, the extent of the allowed misalignment between audio and vision depends on several factors, most importantly relative distance from the user and spatialisation method.

For the latter, no spatialisation at all means that the audio presentation lacks any prominent directional cues and the distance judgement has to be made only based on monaural cues, like

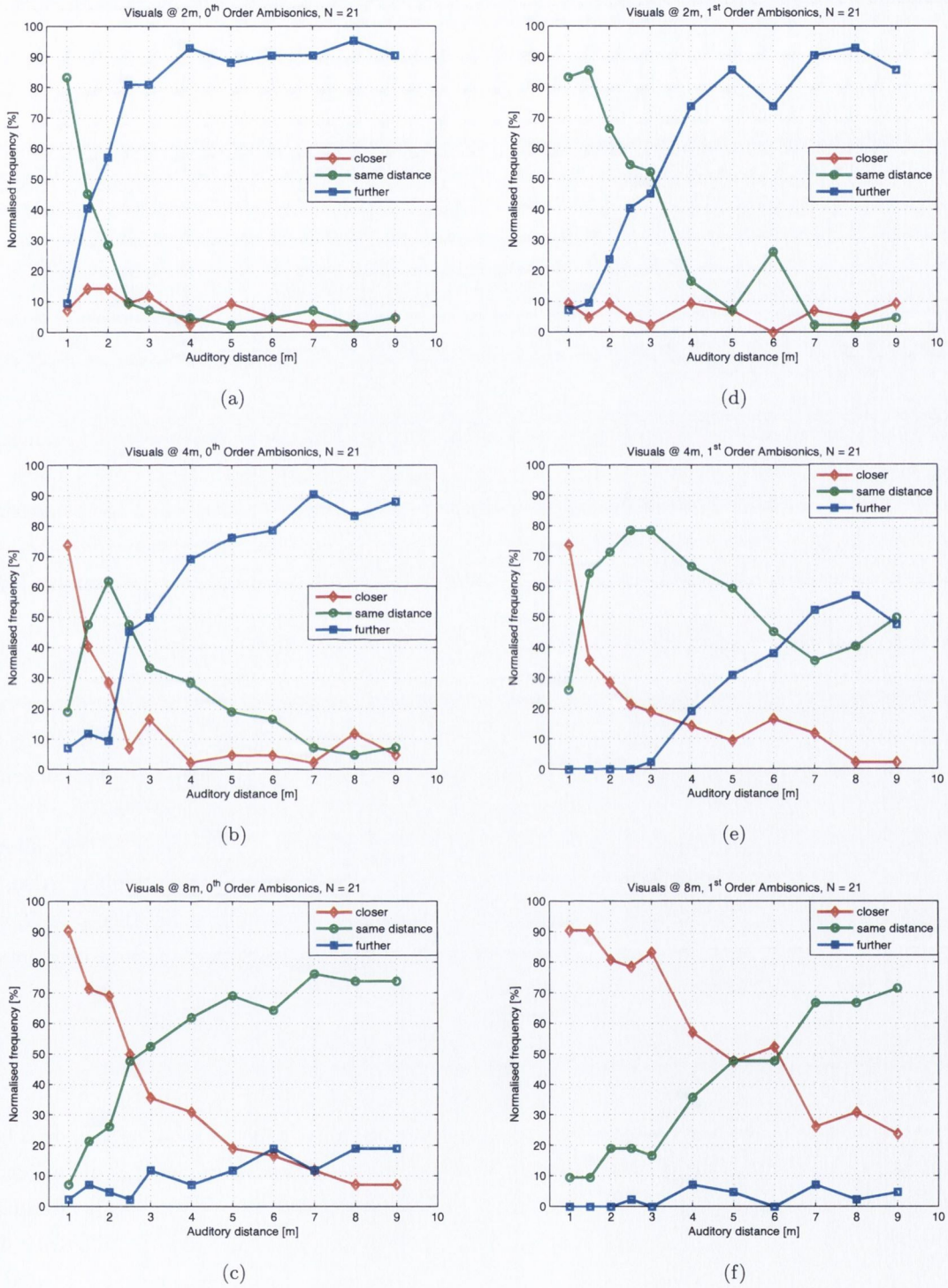


Figure 5.26: Localisation of audio sources with respect to visual objects for combined stimuli

sound pressure level and direct-to-reverberant ratio. In the case of this experiment, it is clear both from the results (Figure 5.26(a-c)) as well as from the verbal reports that it was much more difficult for the subjects to relate the directionless auditory event to the image presented on the screen, especially for the close visual sources (at 2m and 4m). In-head localisation, especially for close sound sources, was also a major problem. It is then speculated that this fact contributed to the observed higher rates of perceived audio-visual misalignment than in the case of directional (i.e. 1^{st} Order) Ambisonic renderings.

Interestingly, the opposite situation occurred when the visuals were presented further away (at 8m). When this was the case, the subjects were more inclined to accept the larger incongruence between the audio and vision than in the case of directional audio.

Regarding the Ambisonic renderings (Figure 5.26(d-f)), results obtained for 4m correlate well with the perceived distance localisation blur for this particular environment [81, 105]. For other distances the results are inconclusive in a sense that the lower and upper bounds of accepted misalignment failed to be detected. What can be observed though is that for close sources (2m) audio can be rendered at least up to 1m closer and up to 1m further away than the visual object and still be perceived as consistent with visuals. Similarly, for a faraway sources (8m) audio can be rendered up to around 2m closer and at least up to 1m further away before the unity breaks.

5.5.6 Conclusions

To conclude, this study showed that human perception puts some limitations on our ability to detect audio-visual distance incongruence. It seems that the allowed misalignment margins tend to grow with the relative distance between the source and the user but due to the lack of reliable data for very close and very faraway sources this hypothesis cannot be yet proved.

Further work should focus on improving audio-visual presentation methods by, e.g. allowing for user movements and incorporation of other distance cues like motion parallax. Also, this study revealed the need for extending the number of presented auditory distance points, especially at both extremes. Finally, the results obtained using virtual acoustics could be compared against real sound sources and tested for possible correlations.

5.6 Conclusions

The studies performed clearly show that the sensation of auditory distance can be achieved using both loudspeaker and headphone based auralisation. In headphone listening, the virtual sources can be perceived externally even if non individualised HRTFs are used. However, head-tracking aids the effect and mitigates some problems arising from the binaural rendering like front-back confusions. In loudspeaker listening, the virtual sound sources can be perceived beyond the radius of the reproduction array. Moreover, in both cases distance judgements for several tested points are no different than the same judgements obtained for real, reference sources. Thus,

no improvement in this regard has been found when increasing the directional resolution of the presented samples.

From the results it can be inferred then that monaural cues, i.e. sound intensity level and direct-to-reverberant ratio, are of greater importance to distance estimation than binaural cues, including the directional resolution of the impulse response. It is hypothesised though that the time of arrival of the first reflections can still play an important role, especially if the presentation of the source is taking place in environments different acoustically.

As proven in the Experiment I(a) and (b), the visual component is not necessary for the correct perception of auditory distance. However, as presented in the Experiment II, it can be used to influence the auditory distance so that users' estimates are "dragged" to the visual anchors. Moreover, it allows for a certain degree of incongruence in the audio-visual rendering before users start to perceive two separate sources.

6

Binaural Sound Field Stabilisation

So far we have considered the creation of VAEs that were based on dry/anechoic source recordings which were subsequently spatialised and placed in a virtual acoustic space by the means of SRIR and HRIR processing. However, in many cases, auralisations are already created for some pre-determined loudspeaker configuration and no binaural signals are available. This is often the case e.g. in video games where binaural audio is still seldom available (see Section 4.1.2.2. That is why, creating binaural auralisation from the multichannel loudspeaker signals is a worthwhile topic and it will be considered in this chapter.

The lack of binaural streams of audio can be rectified by e.g. 1) applying optimised virtual loudspeaker reproduction scheme in order to create binaural down mix of a multichannel sound track, as presented in Chapter 3; 2) applying head-tracking to the binaural reproduction in order to stabilise the resultant sound field. This second step of the approach can be very attractive, especially if one realises that there already exist relatively cheap tracking devices, like *Microsoft Kinect* [148] or *NaturalPoint TrackIR* [161], that are used in video games. As mentioned before in Section 4.1.2.2 of Chapter 4 there are other solutions on the market like myth Research Realiser A8[1] or Beyerdynamic Headzone[2]. However, because of the use of dedicated hardware and their relatively high price it is unclear whether their use with video games is significant. On the other hand, head-tracker informed sound field stabilisation can be applied efficiently in the software, which will be the topic of next sections.

In Chapter 3 it was shown that the sound field comprising spatial information may be effec-

[1] `http://www.smyth-research.com/products.html`

[2] `http://europe.beyerdynamic.com/shop/headzone-headphone-surround-system.html`

tively delivered using headphones through which binaural signals are received. This is because the binaural signals convey sufficient information to recreate a virtual sound field comprising one or more virtual sources. However, in such a situation, head movements of the user must be accounted for in order to keep the sound field stable. This is desirable, for example, if one wants to maintain a relationship, synchronisation or coincidence between the audio and the video. Failure to maintain a stable sound field may result in the perceived virtual source location offset by a degree related to the scale of movement.

There are other benefits of sound field stabilisation often recalled in the literature. For example, maintenance of a stable sound field may induce more effective externalisation of sound sources [55, 126, 229] or in other words, strengthen the sense that the audio sources are external to the listener's head. The effectiveness of this process is still not fully clear [19] however some authors suggest that the improved binaural presentation due to head-tracking results in localisation which is not significantly different from natural hearing [212]. One of the important factors that has been identified is that even small, unconscious head movements help to resolve front-back confusions [19, 212]. In binaural listening, this problem occurs the most often when non-individualised HRTFs are used. Then, it is usually difficult to distinguish between the virtual sound sources at the front and at the back of the head.

Also, it is natural to turn the head towards the the direction of the sound source [5]. It is easier then to focus on the sound source. That sound field rotation can also be used to aid this focus and facilitate the localisation. The premise for this is that whenever the user performs head rotation they have a chance of bringing the sound source into the *area* of sharper localisation (i.e. front of the head). The localisation and tracking of a sound object should then become easier.

In practice, sound field stabilisation means that the virtual loudspeakers need to be repositioned in the 3-D sound field in order to counteract user's head movements. However, it is useful to realise that this process is equivalent to applying panning functions to virtual loudspeaker feeds. That is why it is important to determine the most optimal and also cost effective panning solutions that could be used in the process of sound field stabilisation with head-tracking.

6.1 Stabilisation of Arbitrary Sound Fields

We have already shown that the whole sound field can be effectively rotated with the use of transformation matrices provided that all the sound sources in the scene were encoded into the Ambisonic domain. In this chapter, we will extend this discussion to sound fields with arbitrary spatial accuracy.

In Section 3.4.2 of Chapter 3 it has been shown that Ambisonic transformation matrices provide an easy, effective and efficient way of rotating Ambisonically encoded sound fields, i.e. changing spatial locations of all the sources in the scene in a synchronous and coherent way to counteract the head movements. Now, we will show that it is also possible to rotate sound fields

encoded into other audio formats after their prior conversion into the B-Format domain. In our investigations we will mainly refer to the popular ITU 5.1 surround sound format [101] which recommends the following loudspeaker placements (according to the coordinate system used in this work): $L = 30°$ (Left), $R = -30°$ (Right), $C = 0°$ (Centre), $Ls = 110 \pm 10°$ (Left surround), $Rs = -110 \pm 10°$ (Right surround). However, it must be emphasised that other recommendations so to the loudspeakers placement exist. For example, the Recording Academy's Producers & Engineers Wing [4] recommends positioning of the rear loudspeakers between 100° and 150° with the optimal placement ranging between 135° and 150°. What follows is that we have to accept that the lateral positioning of the sound sources in 5.1 mixes is usually uncertain unless we know exactly what loudspeaker configuration had been used at the production stage.

The conversion of any discreet multi-channel audio stream to the B-Format can be obtained by re-encoding each individual loudspeaker feed using spherical harmonic functions. This process can be seen as treating the reproducing loudspeakers of some given multi-channel audio system as *new* sound sources. Specific example of this method has been already discussed e.g. by Laitinen and Pulkki in [116] in the context of the ITU 5.1 format and its directional coding.

As already mentioned in Section 3.1.2, the .1 channel in the ITU 5.1 layout does not convey any spatial information about the sound sources and is used mostly to convey the low-frequency effects (LFE). Also, it is sometimes the case that the rear channels (Ls and Rs) are used as the effect channels as well and all the spatial information is contained only in the front of the array (e.g. spatial mixes of music performances recreating the audience point of view, where the rear loudspeakers are delivering additional reverberation). However, let us focus on the examples where both the front and the surround channels carry the spatial information about the sound sources which will be a common scenario e.g. in video games. Also, due to the lack of the vertical source information, we can omit the vertical components of the Ambisonics encoding scheme. Then, in order to re-encode 5.0 ITU channels into B-Format channels it is only necessary to perform:

$$\underbrace{\begin{bmatrix} \mathbf{w} \\ \mathbf{y} \\ \mathbf{x} \\ \mathbf{v} \\ \mathbf{u} \\ \vdots \end{bmatrix}}_{\mathbf{b}} = \underbrace{\begin{bmatrix} Y_0^0(\Phi_L) & Y_0^0(\Phi_R) & Y_0^0(\Phi_C) & Y_0^0(\Phi_{Ls}) & Y_0^0(\Phi_{Rs}) \\ Y_1^{-1}(\Phi_L) & Y_1^{-1}(\Phi_R) & Y_1^{-1}(\Phi_C) & Y_1^{-1}(\Phi_{Ls}) & Y_1^{-1}(\Phi_{Rs}) \\ Y_1^1(\Phi_L) & Y_1^1(\Phi_R) & Y_1^1(\Phi_C) & Y_1^1(\Phi_{Ls}) & Y_1^1(\Phi_{Rs}) \\ Y_2^{-2}(\Phi_L) & Y_2^{-2}(\Phi_R) & Y_2^{-2}(\Phi_C) & Y_2^{-2}(\Phi_{Ls}) & Y_2^{-2}(\Phi_{Rs}) \\ Y_2^2(\Phi_L) & Y_2^2(\Phi_R) & Y_2^2(\Phi_C) & Y_2^2(\Phi_{Ls}) & Y_2^2(\Phi_{Rs}) \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}}_{\mathbf{Y}} \underbrace{\begin{bmatrix} \mathbf{L} \\ \mathbf{R} \\ \mathbf{C} \\ \mathbf{Ls} \\ \mathbf{Rs} \end{bmatrix}}_{\mathbf{s}} \tag{6.1}$$

where $\mathbf{b}$ is the column vector containing the horizontal-only B-Format signals (in accordance with the ACN convention [9]), $\mathbf{Y}$ is the spherical harmonics encoding matrix, $\mathbf{s}$ is the column vector with the 5.0 channel signals and $Y_n^m(\Phi)$ are the spherical harmonics coefficients. Vertical

dots denote the expandability beyond the 2^{nd} order decomposition that is used in this example.

From this point onward, application of the transformation matrices to the ambisonically re-encoded 5.0 stream follows the rules outlined in Section 3.4.2. It is, as in the standard procedure, the transformed B-Format signals need to be subsequently decoded to suit a chosen loudspeaker (or virtual loudspeaker) configuration. This process will depend on factors such as what order of spherical harmonics decomposition have been used and how many loudspeakers are needed for the correct reconstruction. To address this problem, one has to decide what directional resolution, or channel separation of the 5.0 stream, needs to be preserved. Again, the higher the channel separation, the higher the order of truncation must be used and also the higher the cost will be in terms of the loudspeaker count and processing power required. From the binaural rendering point of view, each additional virtual loudspeaker in the decode results in adding two additional HRTFs which, depending on the filters' lengths and the filtering method used, may bring up the computational burden significantly. To combat this, it is useful to analyse the 5.0 mixes from the psychoacoustic point of view in order to determine what nominal localisation can be expected from different angular locations in the centre of the array. This will be the topic in one of the next sections.

6.2 Efficient Stabilisation of Non-uniform Sound Fields

One type of optimisation can be done to the system is by realising some of the time-invariant aspects thereof. Whenever the initial or original loudspeaker angles as well as the reconstruction loudspeaker angles are known *a priori*, it can be shown that the encoding, transformation (e.g. rotation) and decoding stages can be actually reduced to a single matrix operation. This approach can be directly applied to regular-to-regular and irregular-to-regular layout mappings:

$$\begin{array}{lllll} \textit{Encoding}: & \mathbf{b} & = \mathbf{sY} & & \\ \textit{Transformation}: & \mathbf{b_T} & = \mathbf{Tb} & = \mathbf{sTY} & \\ \textit{Decoding}: & \mathbf{g} & = \mathbf{Db_T} & = \mathbf{s}\underbrace{\mathbf{DTY}}_{\mathbf{G}(\Phi_H)} & \end{array} \tag{6.2}$$

where the matrix multiplication **DTY** can be replaced with a single matrix $\mathbf{G}(\Phi_H)$ whose elements will only depend on the current listener head orientation Φ_H.

In general, uniform loudspeaker configurations (e.g. diametrically opposed pairs) are preferable to work with in from the Ambisonics point of view, however this is by no means the ultimate requirement. Nonetheless, irregular-to-irregular mappings (or even regular-to-irregular mappings, should such a need occur) may prove to be more problematic since naïve decoding for irregular layouts like ITU 5.1 may result in non-optimal decoder behaviour and impaired localisation (e.g. [79]). The methods for optimising the decoding for the ITU 5.1 are numerous in the literature and can be found for example in [20, 21, 79, 153, 234].

However, it is useful to realise that in any case the rotated sound field would result (at the decoder end) in new loudspeaker gain coefficients applied to the loudspeaker signals. These modified gains, in turn, can be thought of as a weighted sum of all the original loudspeaker gains in the setup. It can be presented presented mathematically as in Equation 6.3 based on the example of a non-uniform ITU 5.1 setup and disregarding the LFE channel:

$$\begin{bmatrix} L' \\ R' \\ C' \\ Ls' \\ Rs' \end{bmatrix} = \begin{bmatrix} G_{1,1}(\Phi_H) & \dots & G_{1,5}(\Phi_H) \\ \vdots & \ddots & \vdots \\ G_{5,1}(\Phi_H) & \dots & G_{5,5}(\Phi_H) \end{bmatrix} \begin{bmatrix} L \\ R \\ C \\ Ls \\ Rs \end{bmatrix} \tag{6.3}$$

or simply

$$\mathbf{g}' = \mathbf{G}(\Phi_H)\mathbf{g} \tag{6.4}$$

where $[L, R, C, Ls, Rs]^T$ and $[L', R', C', Ls', Rs']^T$ are original and transformed 5.0 loudspeaker feeds due to head rotation by the angle Φ_H. Thus, any arbitrary decoder optimisation process can be performed by multiplying the matrix $\mathbf{G}(\Phi_H)$ with specific correction gains, e.g. in a from of a diagonal matrix $\mathbf{W}$ as presented in Equation 6.5.

$$\mathbf{g}' = \mathbf{G}(\Phi_H)\mathbf{W}\mathbf{g} \tag{6.5}$$

In order for the virtual loudspeakers to be applied to the rotated signals, each re-calculated loudspeaker gain needs to be convolved with the corresponding pair of the HRIRs. The signal flow diagram illustrating the rotation of the ITU 5-loudspeaker sound field for binaural presentation is illustrated in Figure 6.1. Interestingly, the same matrix multiplication procedure can be used when panning a virtual sound source around. It seems logical, since with every head rotation, all the audible channels in a system need to be panned in the contrary direction in order to compensate for this movement. This problem will be investigated in the next section.

6.3 Sound Field Stabilisation using Panning Functions

It has already been proposed in Section 6.2 that for more efficient processing and use with real-time applications the re-encoding, rotation and decoding stages of non-uniform sound fields is performed with a single matrix $\mathbf{G}$ containing new loudspeaker gain coefficients that are derived

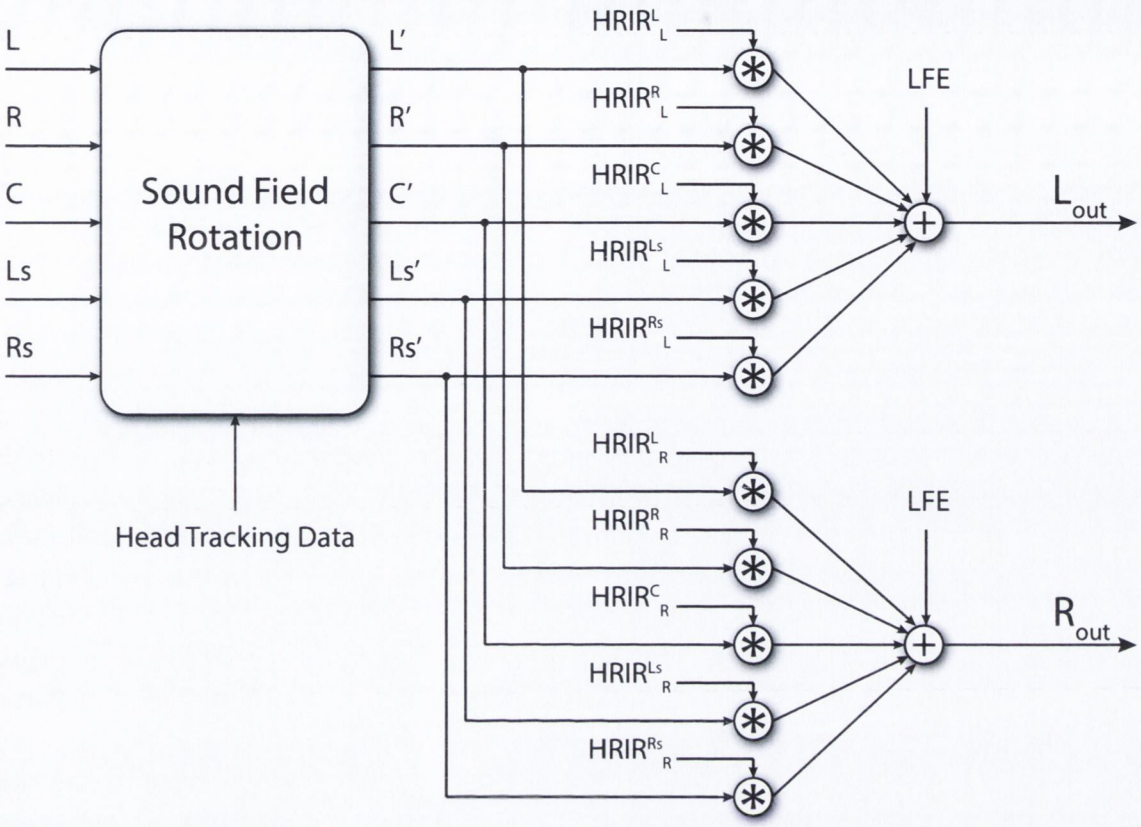


Figure 6.1: Signal flow diagram of the 5.1 rotation procedure for the binaural presentation

based on the current user's head orientation. This operation can seen as equivalent to applying a panning function $g_i(\phi_S)$ to each discrete loudspeaker feed. In order to explain the method in detail we will first refer to some more common panning techniques, primarily the state-of-the art Pairwise Constant Power Panning (PCPP) (e.g. [232]) the Ambisonic Equivalent Panning proposed by Neukom [162] and Continuous Panning Law by Craven [42].

There are also other panning techniques. An overview of many of them can be found e.g. in [232]. In 2004 Kyriakakis et al. [114] proposed a panning algorithm for irregular speaker placement (e.g. ITU 5.1, 7.1 etc.). They developed a gain correction scheme for irregular loudspeaker positioning so that the same output power is guaranteed with different panning angles. The other focus of the work on creating wide as opposed to point-like phantom image sound sources. In 2009 Tanno et al. [219] presented a method for improving phantom source imaging in the ITU 5.1 configuration. However, they work focused on the sides of the array only. Also, at the time of writing of this report, Zotter and Frank [246] also proposed a novel panning algorithm which combines VBAP with Ambisonics and which is intended to be easily scalable for different loudspeaker configuration (including "with-height" arrays). It would be an

interesting future work to include their algorithm in the comparative analysis.

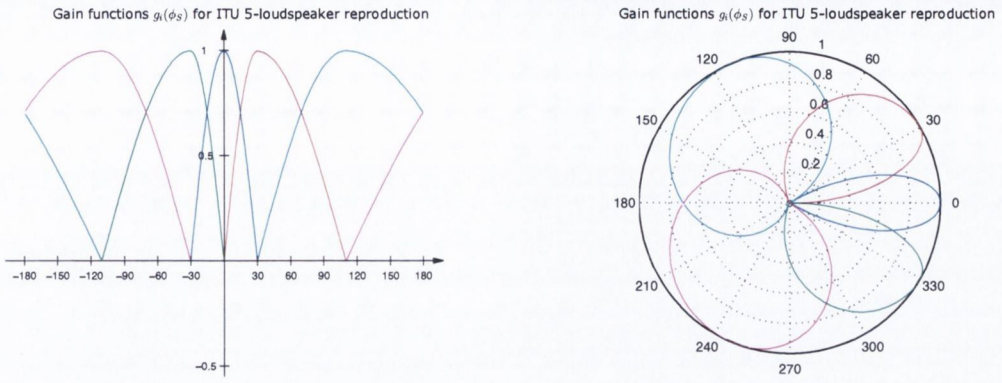
Here, however we focus on the three aforementioned panning algorithms in order to perform the detailed analysis from the point of view of the objective localisation metrics - energy and velocity vectors (as defined and explained in Section 3.3.3.1). Subsequently, a novel approach of direct unconstrained non-linear gain optimisation will be proposed as a valid, expandable method for generating transformation matrices for real-time, head-tracked audio.

6.3.1 Pairwise Constant Power Panning

Pairwise Constant Power Panning can be explained with the two following statements: (1) In the case that the sound source is coincident with a particular loudspeaker in the array, only this particular loudspeaker outputs the signal with others remaining silent; (2) In the case that the source is located in between two loudspeakers, both of them will contribute to the phantom image creation with the amplitudes resulting in the same total power of the source signal. As the name suggest, this method attempts to preserve the output power of the virtual sound source at all panning angles and thus, its emitted acoustic energy and perceived loudness. The uniform, minimised fluctuation of emitted energy seem not to be without consequences to the perception since the uniform perceived loudness with the panned sound source constitutes an important cue from the point of view of the perceived sound distance.

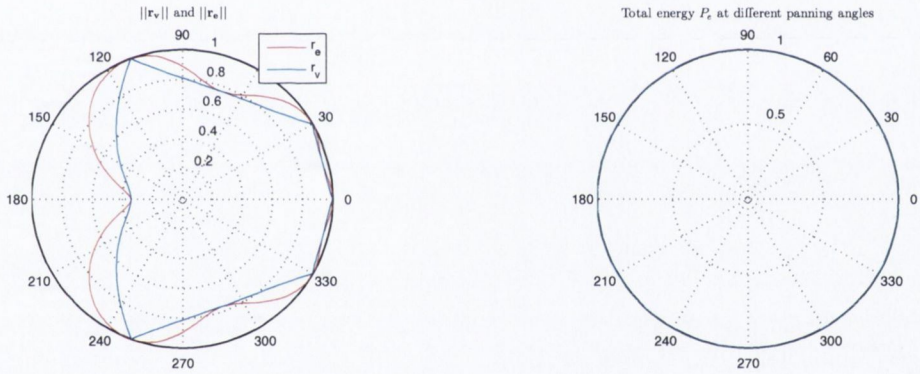
However, although the performance of the PCPP is optimal in terms of energy conservation with panning angles, localisation problems are likely to occur if the angle between loudspeakers is large (this is the so-called "hole in the middle" effect), and whenever the arc is not in front of the listener [42]. That is why, sources panned in the lateral directions or at the back of the ITU 5.1 array are in general considered unstable.

We have already shown in Section 3.3.3.1 using the example of the regular octagonal loudspeaker array that the PCPP already optimises the energy vectors and these can only be deteriorated if one departs from the PCPP [42]. In fact, both the velocity and energy vector directions agree as well as their magnitudes reach unity if and only if the sound source coincides with the reproducing loudspeaker's location. Then, it is intuitively clear that at the remaining directions the lengths of the vectors as well as their direction drop in performance reaching their minimum at the half-way between the loudspeakers. The formal analysis confirming these results are presented below. The gain functions for individual loudspeakers resulting from PCPP at different panning angles are shown in Figure 6.2. Then, Figure 6.3(a) displays the energy and velocity vectors for the PCPP and the standard ITU 5-loudspeaker layout. Fluctuations of total output energy P_e are analysed in Figure 6.3(b). Finally, Figures 6.4(a) & (b) present errors between the intended panning angle and directions of the energy and velocity vectors and Figure 6.4(c) shows the relative energy and velocity vectors angle mismatch.



(a) (b)

Figure 6.2: Gain functions $g_i(\phi_S)$ for individual loudspeakers resulting from PCPP at different panning angles



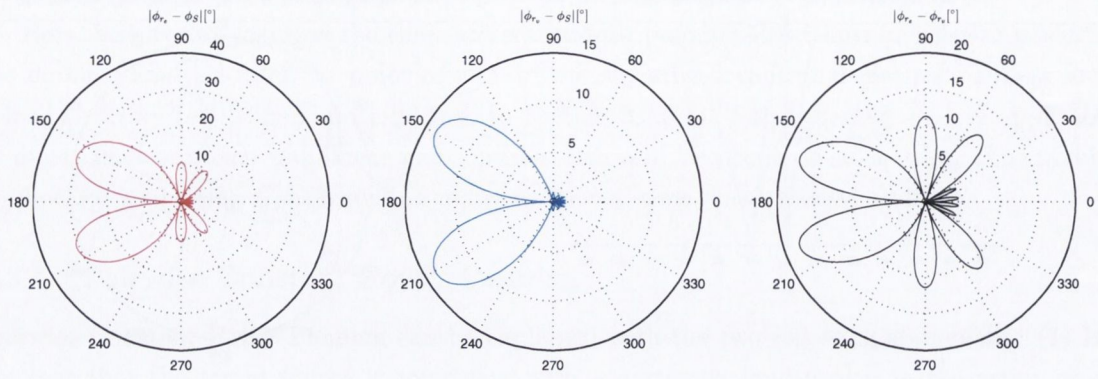
(a) (b)

Figure 6.3: (a) Analysis of the magnitudes of energy and velocity vectors in the case of PCPP; (b) Analysis of the total emitted energy P_e for different panning angles

6.3.2 Ambisonic Equivalent Panning

Amisonic Equivalent Panning was first introduced by Neukom in [162] and it makes use of Equation 6.6 in order to derive the panning functions:

$$g_{AEP}(\phi) = (0.5 + 0.5\,(\cos\,(\phi - \phi_S)))^n = \left(\cos\left(\frac{\phi - \phi_S}{2}\right)\right)^{2n} \tag{6.6}$$



(a) (b) (c)

Figure 6.4: (a) Absolute difference in [°] between the energy vector direction and the intended panning angle; (b) Absolute difference in [°] between the velocity vector direction and the intended panning angle; (c) Absolute difference in [°] between the energy vector direction and the velocity vector direction

where $-\pi \leq \phi < \pi$ and the loudspeaker gains are obtained by evaluating the function $g_{AEP}(\phi)$ at loudspeaker angles Φ_i, e.g. $g_{AEP}(\Phi_L)$, $g_{AEP}(\Phi_R)$, $g_{AEP}(\Phi_C)$, $g_{AEP}(\Phi_{Ls})$, $g_{AEP}(\Phi_{Rs})$.

This equation generates first and higher order cardioid beam patterns according to the value of n used. The behaviour of these cardioids with increasing the parameter n is somehow equivalent to the behaviour of HOA in-phase decoder with increasing the order of spherical harmonic decomposition thus the name Ambisonic Equivalent Panning. The main difference though is that the Equation 6.6 also allows for the fractional values of n which can result in a smooth transition between low and high order cardioid beam patterns. As we shall shortly show, this property can be very useful when dealing with non-uniform loudspeaker layouts like the ITU 5.1.

In order to explain the above statement in more detail, let us look at the standard 5-loudspeaker layout from the point of view of virtual microphones directed at each of the loudspeakers. Although the rear loudspeakers are typically positioned at $\pm 110°$ angles, counting from the front of the listener, the ITU-R BS.775-3 recommendation allows for a small horizontal adjustment of $\pm 10°$. It is useful then to consider for the moment the slightly adjusted loudspeaker array with $L = 30°$, $R = -30°$, $C = 0°$, $Ls = 120°$, $Rs = -120°$ can be viewed as a subgroup of loudspeakers required to form a regular array of 12 diametrically opposed pairs, as shown in Figure 6.5(a). From Section 3.4, Equation 3.66 we know that it would be the minimum required in order to perform a 5^{th} order decode, since $N \geq 2n + 2$ and thus $n \leq (N - 2)/2$. Similarly, Laborie et al. [115] claim that the optimal reconstruction order can be determined based on the two loudspeakers in the setup that form the smallest angle. This again results in

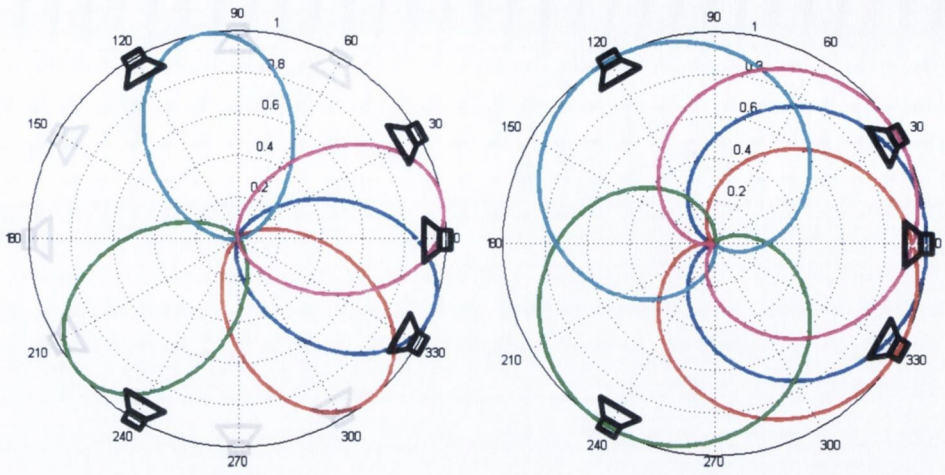


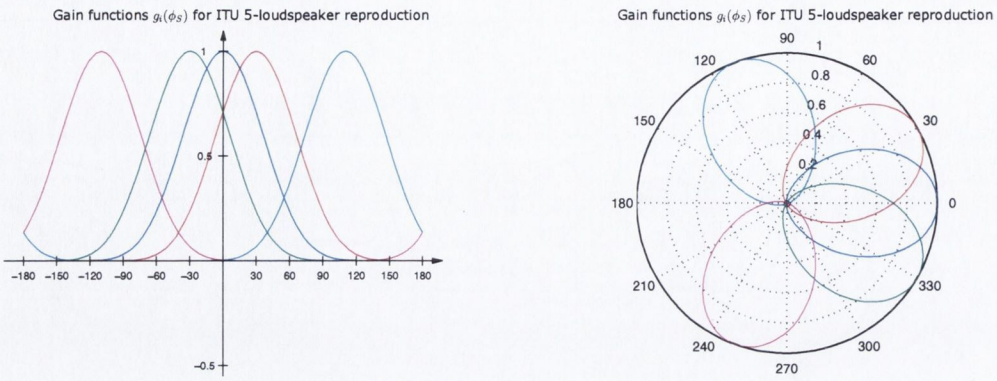
Figure 6.5: (a) Super-cardioid beam patterns used in 5^{th} order AEP (or 5^{th} order in-phase Ambisonic decode). For correct all-around reproduction, ideally 12 loudspeakers would be required, arranged in a diametrically opposed pairs as presented in the picture. (b) First order cardioid beam patterns used in 1^{st} order AEP on the 5.1 loudspeaker setup. Due to overlapping patterns at the front of the array, problems with channel separation and excessive energy at the front of the array can be predicted.

the 5^{th} order as long as the 5.1 layout is concerned. Moreover, they argue that there would be no advantage if the higher order is used.

However, for the lateral or rear source angles, the 5^{th} order cardioids are clearly much too narrow in order to reconstruct the image correctly. The problem is most apparent when the sound source is panned in-between side and rear loudspeakers as illustrated in Figure 6.5(a). More detailed analysis in Figures 6.6 - 6.8 shows the drop of localisation performance at the sides and the rear of the array as well as the energy fluctuations with the panned sound source.

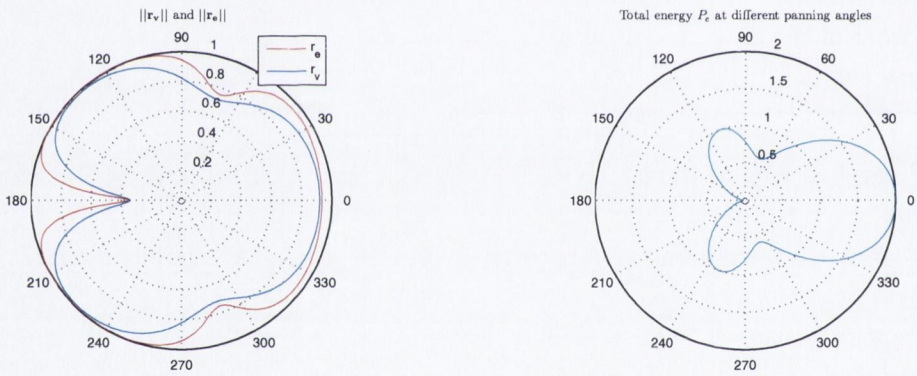
In order to account for the loudspeaker scarcity at sides and the back of the 5.1 array, lower order cardioid beams must be used as illustrated in Figure 6.5(b). However, this in turn leads to very poor resolution at the front and also excessive energy emitted from the three front loudspeakers. The case of the 1^{st} order AEP panning is again, analysed objectively in Figures 6.9 - 6.8.

Therefore, the problem that needs to be addressed with regard to the AEP and non-uniform loudspeaker layouts is really a trade-off between good separation (or spatial resolution) of the three front channels and best possible image reconstruction at the sides and at the back of the array. In order to assure both, it is necessary to vary the width of the virtual microphone beams by adjusting the exponent n according to the intended panning angle. One possible solution could be to change the exponent n dynamically with head rotations from some maximum value at the front of the array to some low value < 1 at the back, depending on the current head



(a) (b)

Figure 6.6: Gain functions $g_i(\phi_S)$ for individual loudspeakers resulting from 5^{th} order AEP panning algorithm at different panning angles

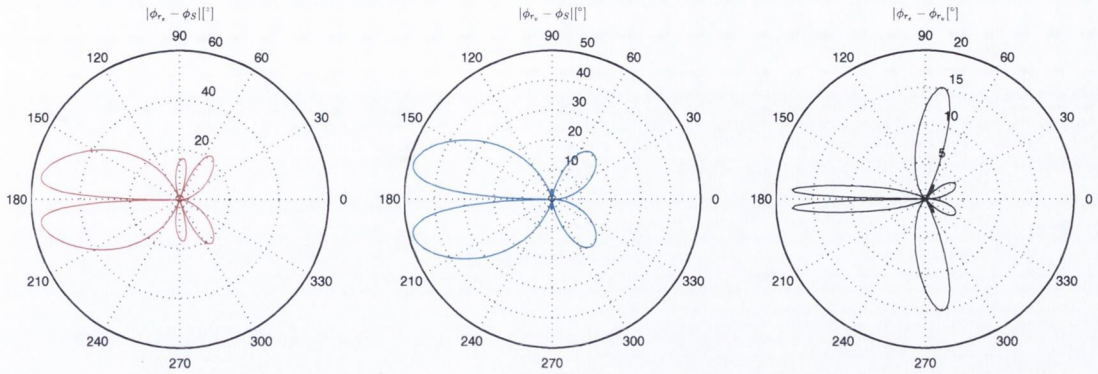


(a) (b)

Figure 6.7: (a) Analysis of the magnitudes of energy and velocity vectors in the case of 5^{th} order AEP panning algorithm; (b) Analysis of the total emitted energy P_e for different panning angles

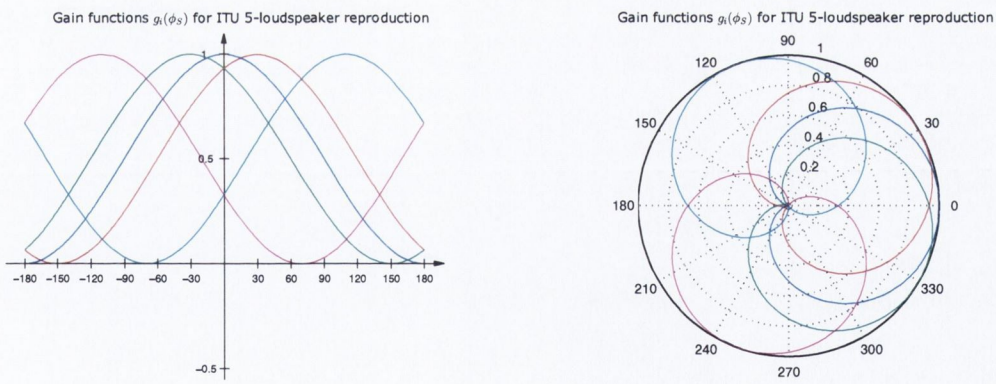
orientation and also depending on which loudspeaker feed is being currently panned. The intermediate values of n can be found e.g. based on the assumption that we want to preserve the source power whenever it appears to be in between two side or two rear loudspeakers, as in the case of the PCPP.

In this thesis, it is proposed that these values are found by pointing the virtual cardioid beam at points of symmetry between two neighbouring loudspeakers (e.g. $\Phi_L + (\Phi_{Ls} - \Phi_L)/2$). Then, using an iterative algorithm and starting from $n = 0$, n is gradually increased by some



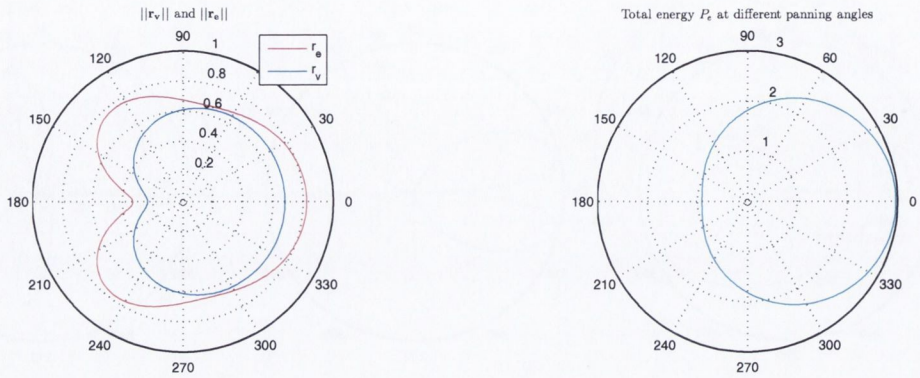
(a) (b) (c)

Figure 6.8: 5^{th} order AEP: (a) Absolute difference in [°] between the energy vector direction and the intended panning angle; (b) Absolute difference in [°] between the velocity vector direction and the intended panning angle; (c) Absolute difference in [°] between the energy vector direction and the velocity vector direction



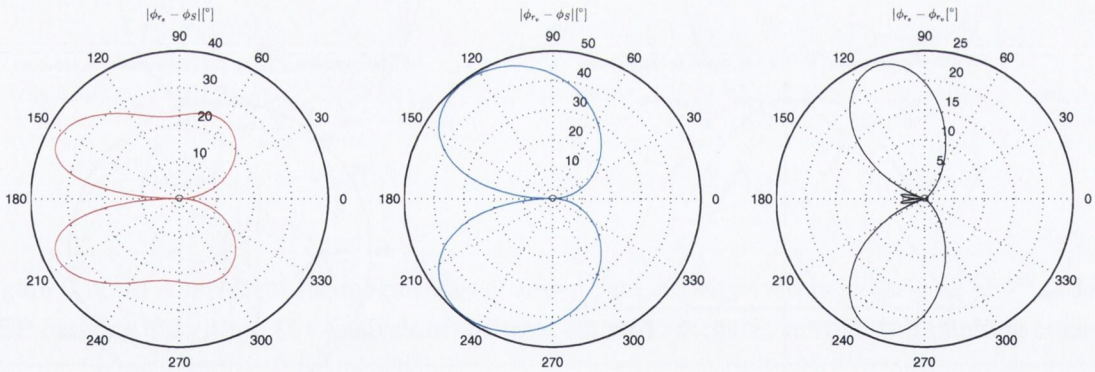
(a) (b)

Figure 6.9: Gain functions $g_i(\phi_S)$ for individual loudspeakers resulting from 1^{st} order AEP panning algorithm at different panning angles



(a) (b)

Figure 6.10: (a) Analysis of the magnitudes of energy and velocity vectors in the case of 1^{st} order AEP panning algorithm; (b) Analysis of the total emitted energy P_e for different panning angles



(a) (b) (c)

Figure 6.11: 1^{st} order AEP: (a) Absolute difference in [°] between the energy vector direction and the intended panning angle; (b) Absolute difference in [°] between the velocity vector direction and the intended panning angle; (c) Absolute difference in [°] between the energy vector direction and the velocity vector direction

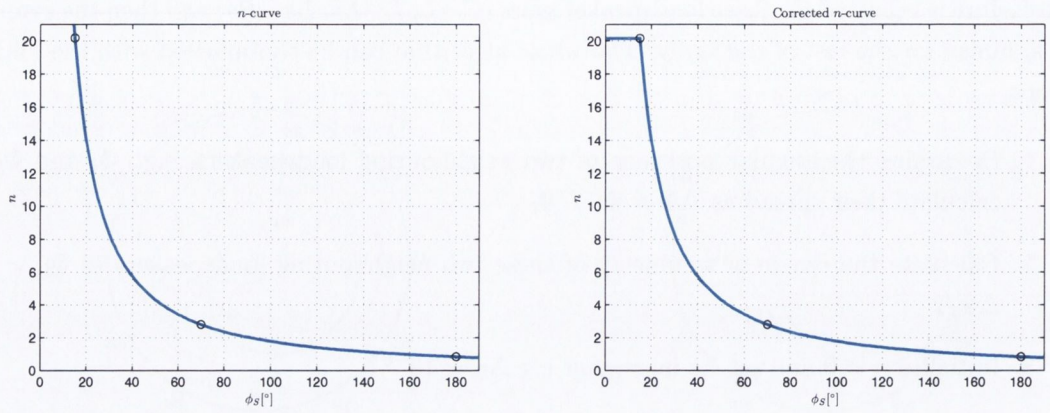
small increment Δn until the $-3dB$ gain is reached at the two loudspeakers of interests. This procedure is repeated for three loudspeaker pairs ($C-L$, $L-Ls$, $Ls-Rs$) and then the symmetry is assumed for the rest of the array. The whole algorithm can be summarised with the following steps:

1. Determine the angular positions of two neighbouring loudspeakers, e.g. Φ_l and Φ_k and calculate their spread as $\Delta\Phi = \Phi_k - \Phi_l$
2. Calculate the centre of symmetry of these two neighbouring loudspeakers as $\Phi_{lk} = \Phi_l + \Delta\Phi/2$
3. Initialise $n = 0$ and set its increment e.g $\Delta n = 10^{-6}$
4. Evaluate $g_{AEP}(\phi) = (0.5 + 0.5(\cos(\phi - \frac{\Delta\phi}{2}))^{n+\Delta n}$ at the loudspeaker angles Φ_l and Φ_k
5. Increment n and repeat until $g_{AEP}(\Phi_l) \approx g_{AEP}(\Phi_k) \approx 0.7071$ which gives $\approx -3dB$
6. Move on to the next loudspeaker pair

After the completion of the steps above, different values of n are derived for few key directions Φ_{lk}, that correspond to the situation where the panned source is in-between two neighbouring loudspeakers. For the loudspeakers angles as defined by the ITU standards the Φ_{lk} values are [15° 70° 180°] and the corresponding n values are [20.1651 2.7855 0.8687]. To obtain the n values for all the remaining angles, a power curve can be fitted to the given points $n(\Phi_{lk})$ and n can be expressed for each panning angle in the following form:

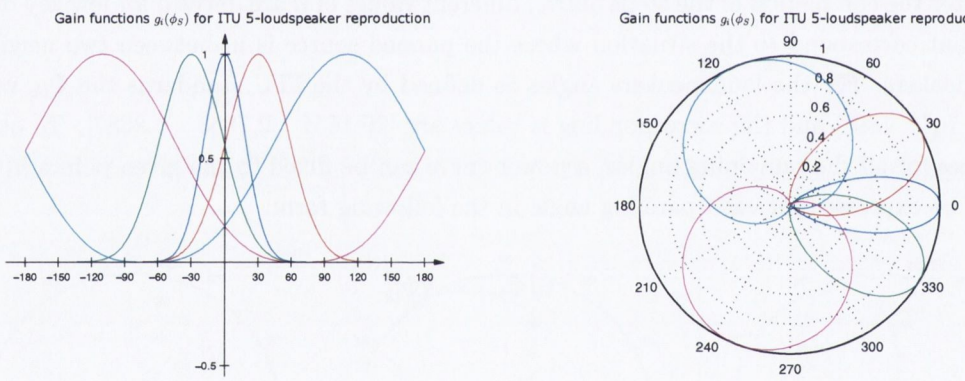
$$n(\Phi_{lk}) = A\Phi_{lk}^{B} \tag{6.7}$$

In this case, fitting the power curve with the parameters $A = 647.9$ and $B = -1.281$ proves to be the most optimal solution with $R^2 = 1$ and the root mean square error of 0.03705. The fitted curve is shown in Figure 6.12(a). Interestingly, the $3dB$ drop at the neighbouring loudspeakers at the front cannot be assured before n exceeds the 20^{th} order. In contrast, at the back of the array, n takes, as expected, values that are less than 1 which corresponds to sub-cardioid beam patterns. One obvious problem that we can immediately notice in Figure 6.12(a) is that for some panning angles at the front of less than ±15°, n takes on very large values. A proposed solution to this problem is to constrain the order n at the $0° - 15°$ arc to some fixed value, just to avoid point-like characteristics of virtual microphones at frontal angles e.g. make the width constant until the 15° threshold is reached and then start widening the cardioid according to the Equation 6.7. The corrected n-curve is presented in the Figure 6.12(b). Then again, by using the property of symmetry of the ITU 5-loudspeaker layout, for angles beyond 180°, the n-curve can be inverted so that the beam pattern narrows back to its maximum order $n = 20.1651$.



(a) (b)

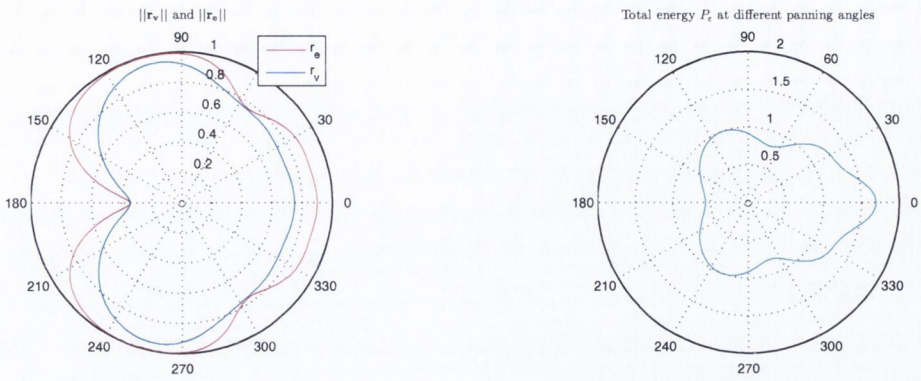
Figure 6.12: n-curve (a) and corrected n-curve (b) fitted into the 3 points of data: $n(15°) = 20.1651$, $n(70°) = 2.7855$ and $n(180°) = 0.8687$



(a) (b)

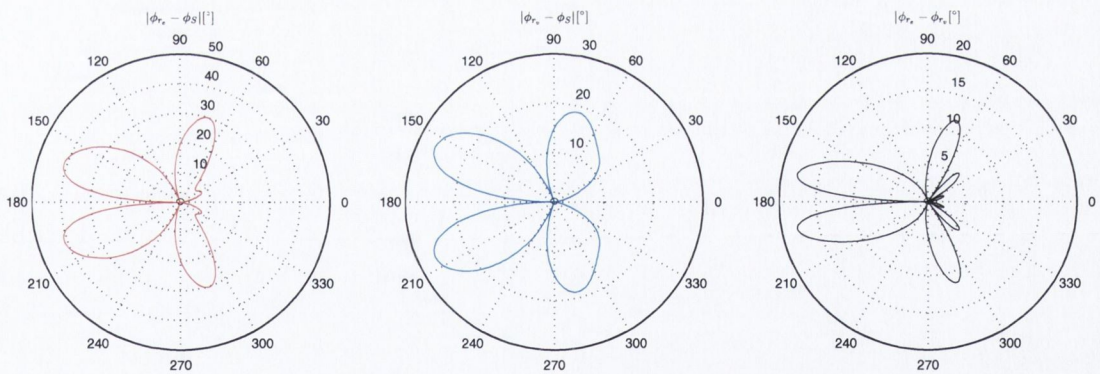
Figure 6.13: Gain functions $g_i(\phi_S)$ for individual loudspeakers resulting from variable order AEP panning algorithm at different panning angles

The panning curves resulting from the variable order AEP are presented in Figure 6.13. Their performance is again analysed objectively in Figures 6.14 and 6.15. In terms of the reproduced energy, it is clearly visible that front sources are favoured above the rear ones and the fluctuations are still significant, although much milder than in the case of fixed order AEP. It is however at the cost of the energy and velocity vectors that are slightly worse than in the case of the fixed 5^{th} order AEP. In summary, from the point of view of objective localisation analysis, AEP is unable to give comparable results to the PCPP in any aspect.



(a) (b)

Figure 6.14: (a) Analysis of the magnitudes of energy and velocity vectors in the case of variable order AEP panning algorithm; (b) Analysis of the total emitted energy P_e for different panning angles



(a) (b) (c)

Figure 6.15: Variable order AEP: (a) Absolute difference in [°] between the energy vector direction and the intended panning angle; (b) Absolute difference in [°] between the velocity vector direction and the intended panning angle; (c) Absolute difference in [°] between the energy vector direction and the velocity vector direction

6.3.3 Craven's Continuous Panning

Craven realised that introducing small amounts of anti-phase components can actually improve the performance of the panning functions at low frequencies, where velocity vectors are good predictors of the quality of localisation, without affecting too much the performance of the energy vectors, that predict the localisation at mid and high frequencies. He argues that low negative gains, when squared, do not significantly affect the energy vectors but may actually lead to better reconstruction of acoustic velocity. In order to derive necessary beam patterns, in [42] he proposed that the panning functions are expressed as a sum of circular harmonics using the following equation:

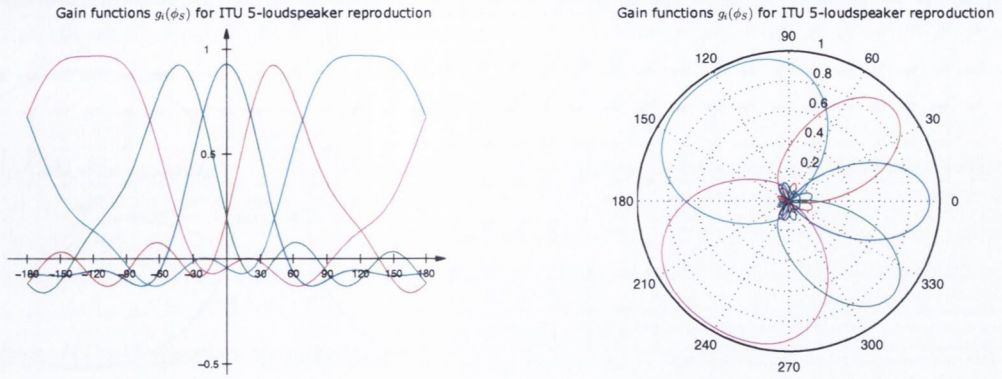
$$g_C(\Phi_i) = \sum_{m=1}^{n} (\alpha_{\Phi_i,n} \cos(m\phi_S) + \beta_{\Phi_i,n} \sin(m\phi_S)) \tag{6.8}$$

where $g_C(\phi)$ is the gain of the i^{th} for the desired panning angle ϕ_S. Coefficients $\alpha_{\Phi_i,n}$ and $\beta_{\Phi_i,n}$ are found by the means of minimisation of a cost function in order to optimise the system in light of the following objectives:

- Reproduced energy should be substantially independent of panning angle
- The velocity and energy vector directions ϕ_{r_v} and ϕ_{r_e} should be closely matched
- The angles ϕ_{r_v} and ϕ_{r_e} should be reasonably close to the panning angle ϕ_S
- Velocity vector length $||\mathbf{r_v}||$ should be close to unity
- Energy vector length $||\mathbf{r_e}||$ should be as large as possible

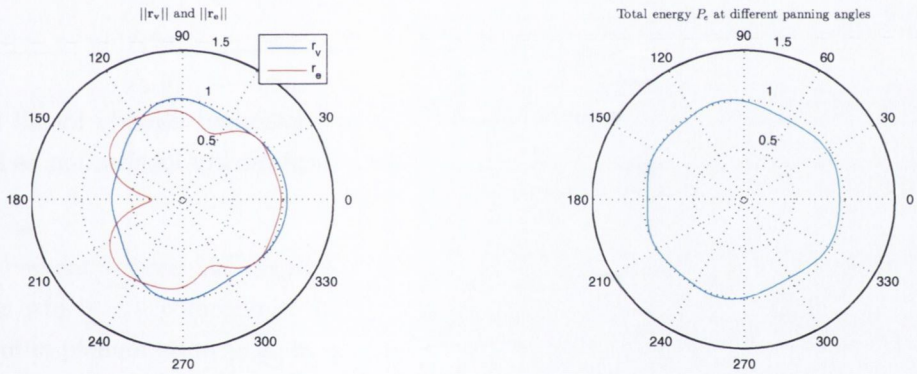
So clearly, the panning functions are optimised based on the aforementioned localisation quality criteria - energy and velocity vectors - as first proposed by Gerzon to describe the desired performance of a real Ambisonic decoder [76]. The $\alpha_{\Phi_i,n}$ and $\beta_{\Phi_i,n}$ coefficients simply scale the circular harmonic coefficients of order 1 up to 4. The optimisation is done using non-linear conjugate-gradient method and is used in order to minimise the cost function that combines into one mathematical expression the above objectives.

The resultant panning curves are presented in Figure 6.16. The objective analysis of the gain functions (Figures 6.17 - 6.18) clearly shows the superior performance of the magnitudes of velocity vectors as compared to the methods presented so far. Energy fluctuations with panning angles are minute. In his own analysis, Craven points out at the troublesome spot at $\phi_S \approx 50°$ where the substantial velocity and energy mismatch error of around 10° may cause localisation problems in the sensitive frontal stage (Figure 6.18(c)). It is a possibility that the problem is influenced by the individual mismatch of both energy and velocity vector angles as related to the panning angle, which are quite large around this region (Figure 6.18(a) & (b)). In order to



(a) (b)

Figure 6.16: Gain functions $g_i(\phi_S)$ for individual loudspeakers resulting from Craven's continuous panning algorithm at different panning angles



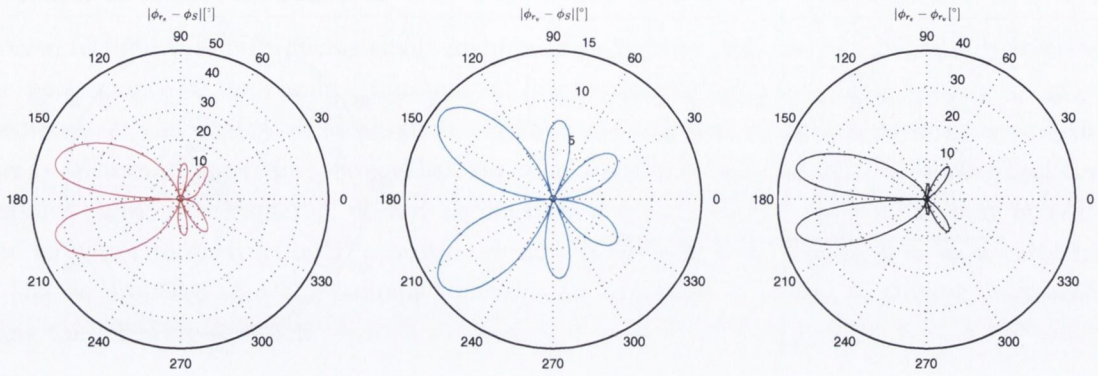
(a) (b)

Figure 6.17: (a) Analysis of the magnitudes of energy and velocity vectors in the case of Craven's continuous panning algorithm; (b) Analysis of the total emitted energy P_e for different panning angles

mitigate the problem it is hypothesised by the author that it would be worthwhile to try and adjust the gain of the rear loudspeaker feeds (e.g. by introducing slightly larger negative gains).

6.3.4 Panning by Direct Gain Optimisation

In this work it is proposed that a look-up table with gains coefficients is constructed with an azimuthal resolution of 1° for low-cost implementation of head-tracking to the ITU 5.1-to-



(a) (b) (c)

Figure 6.18: (a) Absolute difference in [°] between the energy vector direction and the intended panning angle; (b) Absolute difference in [°] between the velocity vector direction and the intended panning angle; (c) Absolute difference in [°] between the energy vector direction and the velocity vector direction

binaural mixdowns. The gains in the look-up table are optimised directly for all the panning angles ϕ_S in order to satisfy the objective predictors of best quality localisation as listed below:

1. $||\mathbf{r_e}|| \approx 1$
2. $||\mathbf{r_v}|| \approx 1$
3. $P_e \approx 1$
4. $\phi_{r_e} \approx \phi_{r_v}$
5. $\phi_{r_e} \approx \phi_S$
6. $\phi_{r_v} \approx \phi_S$

The optimisation is done using non-linear unconstrained search [128] for the minimum of the multivariable cost function $f(g) = f(g_1, g_2, g_3, g_4, g_5)$ where g_i are the loudspeaker gains. The total cost function being a sum of partial quadratic functions $f_k(g)$ is designed and analysed symbolically, and reflects in the mathematical way the set of objectives as presented above. The symbolic analysis is performed in order to derive the gradient of the cost function:

$$\nabla f(x_1, x_2, \ldots, x_n) = \left[\frac{\delta f}{\delta x_1}, \frac{\delta f}{\delta x_2}, \ldots, \frac{\delta f}{\delta x_n}\right]^T \tag{6.9}$$

and its Hessian:

$$H(f(x_1, x_2, \ldots, x_n)) = J(\nabla f(x_1, x_2, \ldots, x_n)) = \begin{bmatrix} \frac{\delta^2 f}{\delta x_1^2} & \frac{\delta^2 f}{\delta x_1 \delta x_2} & \cdots & \frac{\delta^2 f}{\delta x_1 \delta x_n} \\ \frac{\delta^2 f}{\delta x_2 \delta x_1} & \frac{\delta^2 f}{\delta x_2^2} & \cdots & \frac{\delta^2 f}{\delta x_2 \delta x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\delta^2 f}{\delta x_n \delta x_1} & \frac{\delta^2 f}{\delta x_n \delta x_2} & \cdots & \frac{\delta^2 f}{\delta x_n^2} \end{bmatrix} \tag{6.10}$$

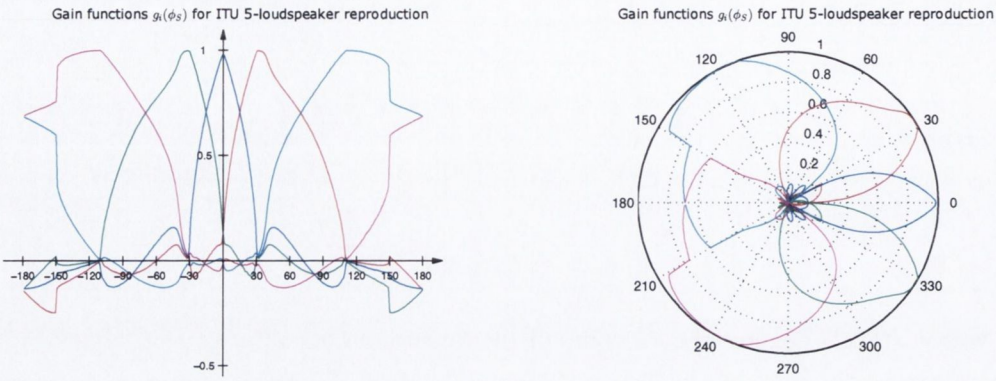
where $J(\xi(x))$ denotes the Jacobian of the function. This approach has the advantage that the gradient estimation by the means of finite differences is avoided and so is the risk of the numerical error, particularly in the estimation of the Hessian. The partial quadratic cost functions and the resultant total cost function are:

$$\begin{aligned} f_1(g) &= (1 - ||\mathbf{r_e}||)^2 \\ f_2(g) &= (1 - ||\mathbf{r_v}||)^2 \\ f_3(g) &= (1 - P_e)^2 \\ f_4(g) &= (\phi_{r_e} - \phi_{r_v})^2 \\ f_5(g) &= (\phi_{r_e} - \phi_S)^2 \\ f_6(g) &= (\phi_{r_v} - \phi_S)^2 \\ f(g) &= f_1(g) + f_2(g) + f_3(g) + f_4(g) + f_5(g) + f_6(g) \end{aligned} \tag{6.11}$$

In its current version the algorithm uses these partial quadratic cost functions with equal weightings which is a compromise between the quality of localisation for a broadband signal and ease of implementation (e.g. in game audio engines). Future work should look at different weighting schemes for the low- and mid- to high- frequency bands where more weight is given to the $f_2(g)$ and $f_6(g)$ at low frequencies and more weight is given to $f_1(g)$ and $f_5(g)$ at mid and high frequencies. For this to happen, shelf filters need to be employed in order to split the multichannel input into low and mid/high frequency streams.

The algorithm used in order to minimise the function the $f(g)$ uses the MATLAB routine `fminunc` to perform a large-scale search for the minimum of the function in the vicinity of some initial guess. The MATLAB script used is included in Appendix E of this thesis. The script expects a 5×360 matrix as an input that contains. In each column there are 5 loudspeaker gains that are used in order to position a sound source at the given angle.

In the process of optimisation it is usually a good practice to choose the initial guess so that e.g. some of the parameters are already optimised. In this vein, the PCPP gain functions computed at 1° increments seemed to be a good candidate for the further optimisation and were used used as a starting point for further optimisation. Using PCPP gain functions as an initial guess, the algorithm converges on a result after 7 iterations on average.



(a) (b)

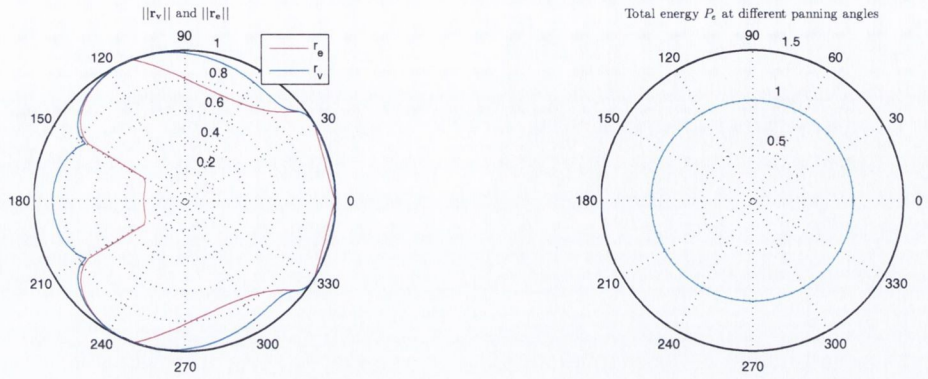
Figure 6.19: Gain functions $g_i(\phi_S)$ for individual loudspeakers resulting from the proposed panning algorithm at different panning angles

The resultant gain functions for individual loudspeakers at different panning angles after the optimisation process are shown in Figure 6.19. Then, Figure 6.20(a) displays the energy and velocity vectors for the optimised panning gains and the standard ITU 5-loudspeaker layout. The lack of fluctuations of total output energy P_e is presented in Figure 6.20(b). Finally, Figures 6.21(a) & b analyse errors between the intended panning angle and the directions of the vectors and Figure 6.21(c) shows the relative energy and velocity vectors angle mismatch.

The results obtained suggest strong performance of the panning functions, especially at the front of the array and also comparable performance to the best-so-far algorithms at the remaining sectors. Fluctuations of the total emitted energy are virtually non-existent across the whole panning domain which makes the method comparable to the PCPP in this regard. The velocity-energy vector direction mismatch at the front of the array is greatly reduced around the troublesome point of 50° (Figure 6.21) and is also smaller at the other sectors of the array.

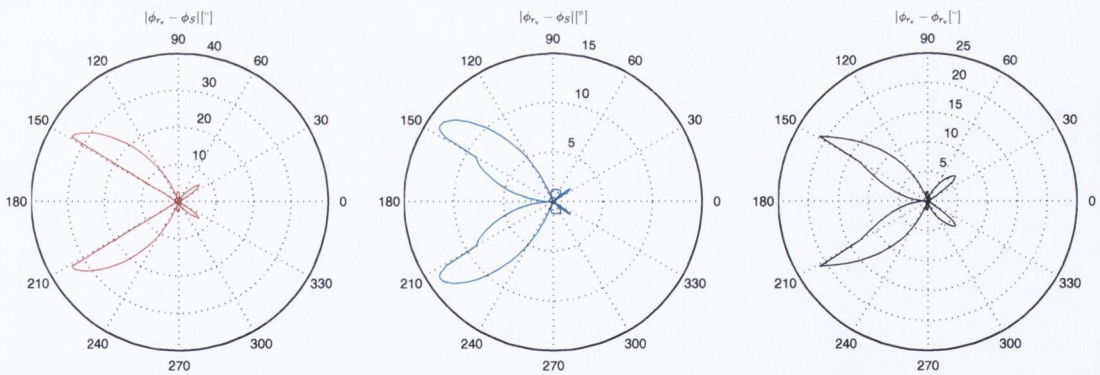
On the downside, this method can result in discontinuities or dramatic gain changes, especially in the sparse loudspeaker areas (i.e. rear of the 5.1 array). As can be seen in Figure 6.19 this is exactly the case when the virtual sound source is panned around around ±150°. The *Ls* and *Rs* loudspeaker gains drop drastically and stay low in the whole arc of around ±150°. However, for the head-tracked audio such a situation would only happen if the user turns their head by more than 40°. This seems to be quite unrealistic, especially if we accept the fact that for most of the optic-based head-trackers used in games like *Kinect* [148] or *TrackIR* [161] there is a limited range of reliable orientation tracking. For example, Kinect face-tracking is capable of tracking user's yaw up to 45° but works best up to 30° [3]. No similar data has been found for the *TrackIR* but from the authors personal experience the yaw tracking limitation of this

[3]See http://msdn.microsoft.com/en-us/library/jj130970.aspx



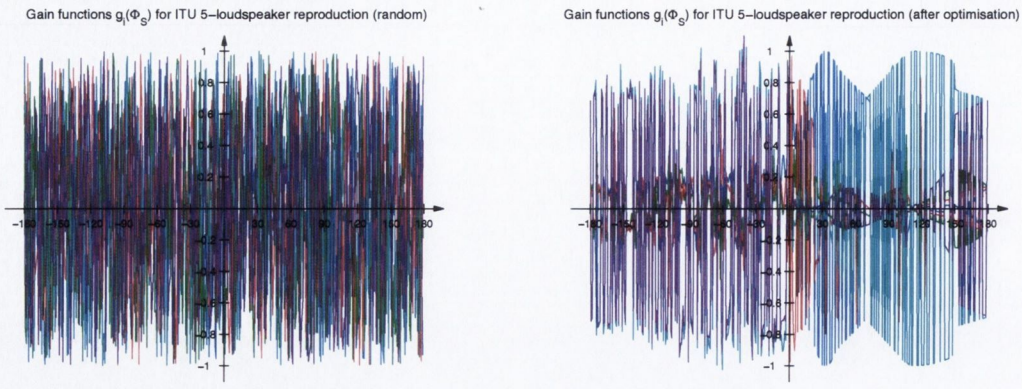
(a) (b)

Figure 6.20: (a) Analysis of the magnitudes of energy and velocity vectors in the case of proposed panning algorithm; (b) Analysis of the total emitted energy P_e for different panning angles



(a) (b) (c)

Figure 6.21: (a) Absolute difference in [°] between the energy vector direction and the intended panning angle; (b) Absolute difference in [°] between the velocity vector direction and the intended panning angle; (c) Absolute difference in [°] between the energy vector direction and the velocity vector direction



(a) (b)

Figure 6.22: (a) Randomly chosen initial loudspeaker gains for ITU 5-loudspeaker reproduction and $-180° - 180°$ panning range; (b) The same loudspeaker gains after approx. 57 iterations of the optimisation algorithm at each angle.

device is in line with the one exhibited by the *Kinect*. Only due to the fact that the clip with the LEDs is mounted on one side of the head, there is a small left-right asymmetry in the yaw tracking limits.

Also, it must be emphasised that different initial guesses may lead to radically different results of optimisation. In the most extreme example in which initial loudspeaker gains are chosen randomly, the algorithm may lead to erroneous an meaningless results. This issue is illustrated in Figure 6.22. The reason for this is that for random initial guesses there increases the risk that the algorithm arrives at the local minimum of the cost function and not the global minimum. However, the risk can be minimised if the initial guesses are chosen in a more meaningful and psychoacoustically justified way, e.g. using gain functions of some other panning method like PCPP. Also, in the case of random initial guesses it takes much longer for the algorithm to converge on a result (57 iterations on average).

Lastly, the optimisation is claimed based on the calculated objective predictors of localisation accuracy and not based on the improvement in terms of number of required operations/MACs. It must be emphasised, that the gain optimisation is done off-line and the results are stored in a lookup table. In fact, this can also be applied to other panning methods revised in this chapter so that alternative lookup tables can be constructed. However, application of the pre-computed gains for the use with head-tracking devices seems to be an attractive approach since accounting for the new user's head orientation only requires to scale the multichannel signals by the relevant gain factors read from the lookup table. Besides that, no other processing of channels is needed.

In terms of expected localisation improvement, the results suggest that the proposed panning algorithm should outperform other panning algorithms, especially in the frontal and lateral

directions. However, due to the lack of one-to-one mapping between the objective predictors and the subjective results, it is unclear at this stage whether the improvement in localisation accuracy will be significant. That is why, future work should evaluate the method proposed subjectively in order to assess the localisation performance of the panned sources with real subjects.

6.4 Conclusions

In this chapter we have explored various techniques that can be used in order to perform stabilisation of arbitrary sound fields using head-tracking. First, we have focused on sound fields encoded into the Ambisonics domain. Following from the discussion in Chapter 3 we have used the notion that other multichannel audio formats can be re-encoded into the Ambisonics domain by encoding their discrete loudspeaker feeds using spherical harmonic functions whose arguments are the angles Φ_i (and Θ_i in the case of 3-D systems) characterising these loudspeakers. With this approach, we can provide uniform rotations (where no directions are favoured above the others) but again, only if the reproduction is done using regular arrangements of loudspeakers. In other cases, the existing knowledge [20,21,79,153,234] may be very useful in order to design a decoder for irregular loudspeaker arrays that would minimise the localisation blur of the rotated sound field in all directions. From the point of view of the Ambisonics theory, infinite expansion of spherical harmonics decomposition of the re-encoded sound field may eventually lead to equivalent localisation of sources in the rotated sound field as it was in the original sound field, but never better. It may seem quite disappointing but we have to remember that as long as there is no access to the information about individual sound sources in the scene, the ambisonic reconstruction can be at most as good as the original mix. So, the scalability of this approach seems to be attractive, particularly if: (1) there is no access to the information about individual sources in the mix (source audio, spatial co-ordinates); (2) uniform and smooth panning is desired; (3) it is desired that no panning direction is favoured over the other directions; (3) processing power required is not the major constraint.

One suggested optimisation to the re-encoding scheme was to combine the re-encoding, transformation and decoding stages into one single stage. This is possible if: (1) the location of original sources is known and also time-invariant (which we assume is true if we deal with the discrete multichannel loudspeaker feeds); (2) the decoding loudspeaker setup is also known and also time-invariant. Then, the resultant sound field transformation matrix $\mathbf{G}$ will only depend on some varying run-time parameters, e.g. current user's head orientation.

In any case, the matrix $\mathbf{G}$ serves the role that maps the original channel feeds onto some new pre-determined reproduction channels. Each new loudspeaker feed is simply a linear combination of all the original channel feeds multiplied by some weighting coefficients $G_{i,k}$. It is interesting to note that for regular-to-regular but also for irregular-to-irregular loudspeaker arrangements, this matrix performs a similar role as a multichannel pan-pot. That is why, it was extremely

useful to analyse various panning functions in terms of their objective localisation performance and especially for the most popular loudspeaker configurations like the ITU 5.1.

Lastly, Section 6.3 revised the panning algorithms from the point of their usability to sound field rotation and stabilisation, e.g. in head-tracked binaural audio. Three algorithms were presented and objectively analysed, namely Pairwise Constant Power Panning, Ambisonic Equivalent Panning and the Continuous Panning Law. The analysis uncovered strengths and weaknesses of each of the algorithms as well as pointed out some possible areas for improvements. As the result, a novel set of panning functions has been developed that successfully addresses the Gerzon's predictors for good localisation [76]. The gain coefficients for each panning angle ϕ_S and 1° increment have been found by direct non-linear optimisation of the quadratic, multivariable objective function. Finally, it was proposed that for most efficient implementation of the rotator, it is suggested to compile a lookup table with all the possible loudspeaker gain coefficients for each panning angle ϕ_S and 1° increment.

7

Conclusion

7.1 Plausible, Real-Time Rendering of Virtual Auditory Environments: Summary

At the beginning of this work the main research problem was formulated, which proposed investigations into novel methodologies for plausible, real-time and interactive rendering of Virtual Auditory Environments. Of concern was to retain the level of realism offered in data-base auralisations whilst significantly minimising the computational and information storage effort. Therefore, the main hypothesis was stated as follows:

> **Convolution-based auralisation can be adapted and efficiently implemented in the context of real-time multi-source and dynamic auralisation, so that the perceptual effect is enhanced whilst maintaining a low overall computational cost.**

This statement was supported by previous research in the fields of audio signal processing techniques, acoustics and spatial hearing theory. In order to address the problem, it was proposed that the data-base auralisation paradigm can be adapted for the use in real-time and interactive applications by the means of parametrisation and simplification of certain aspect of a typical audio signal processing chain. These components were identified as those responsible for creating the spatial impression of the enclosing spaces (Room Impulse Responses) and, in the case of headphone listening, those responsible for binaural rendering of the sound fields (Head Related Impulse Responses). For the former, psychoacoustically driven optimisation has been proposed

that decomposes the filters according to their directional content. Then, incorporation of these components into the reconstructed sound fields is divided into off-line (diffuse reverberation) and on-line (direct sound and early reflections) processes. Additionally, high directional resolution can be applied only to spatialisation of the direct sound and early reflections as, from definition, no prominent directional information is contained in the diffuse reverberation tail.

In the case of binaural reproduction, Head Related Impulse Responses are simplified by decomposition into direction dependent and direction independent components. Similarly, only the components carrying the directional information are incorporated in the real-time processing stage.

In order to verify the hypothesis, investigations into various areas of the broadly defined spatial audio technologies have been performed. In particular:

- Fundamentals of spatial hearing and, specifically, the mechanisms responsible for perception of auditory source direction, distance and motion have been discussed. The human abilities and limitations regarding the spatial perception of sound have been reviewed based on former research in the field. Some important areas directly applicable to this work, such as perception of auditory and multi-modal distance, or perception of auditory motion have been identified as still relatively poorly understood. That is why further investigations have been proposed in order to gain more insight into the above issues.

- Different spatial audio reproduction methodologies have been reviewed. Particular focus has been on those techniques that are most suitable for dynamic and interactive auditory presentations for a single listener. Of particular concern ware headphone-based techniques that allow for personalisation of listening experience and minimise the influence of listening environment.

- A review of existing auralisation techniques used in different fields like architectural acoustics or virtual reality and gaming has been presented. A novel approach to adapt database auralisation to real-time and interactive scenarios has been proposed. It was followed by the discussion on implementation issues in the process of real-world acoustic recording and its subsequent virtualisation. Two applications of VAEs were presented: a traditional Irish trio recorded in a medium-size reverberant hall and performance of a choir in the Christ Church Cathedral in Dublin. Each stage of creating the audio-visual virtual reality environment was explained with a particular focus on generation (source directivity), propagation (parametrisation and adaptation of SRIRs) and reception (HRIR processing) of auditory sound fields. Finally, formal evaluation of the proposed auralisation methodology was performed.

- Further evaluation of the VAEs and, in particular, SRIR parametrisation was performed. Perception of auditory distance in headphone-based and loudspeaker-based VAEs has been investigated by the means of subjective listening tests. The virtual environments were

created based on real-world SRIRs whose directional components were subsequently re-rendered into Higher Order Ambisonics. It was concluded, that perception of auditory distance in virtual headphone and loudspeaker renderings matches the real-world experience well.

- Because of the fact that virtual auditory environments are in many cases accompanied by visuals, for the sake of completeness, the effect of visual cues on perception of distance was also addressed. In particular, the impact of incongruence in audio-visual cues on distance assessments was investigated. The results suggest that in the presence of strong visual cues the necessity of accurate auditory distance rendering is mitigated.

- Lastly, the work was extended to include binaural rendering of loudspeaker based auralisation by the means of optimised virtual loudspeaker rendering scheme and head-tracking. In order to address the problem of listener's motion in headphone listening, a unified approach to sound field rotation has been proposed that allows for efficient spatial transformations of arbitrary sound fields with head movements. Audio signal processing techniques have been devised that allow for compensation of user head movements in headphone listening. Sound field rotation was discussed as a strategy for stabilisation of headphone sound fields so that the virtual sources retain their original locations with head movements. Optimal ways of performing rotations for different audio spatialisation schemes have been sought.

The findings of this thesis can be summarised by the following list:

- Real-time and interactive rendering of Virtual Auditory Environment can be psychoacoustically optimised to minimise the signal processing effort while achieving perceptually equivalent effects.

- Binaural rendering can be significantly optimised without losing perceptual sound quality by applying factorisation to HRIR filters.

- It has been shown that perception of auditory distance is not particularly sensitive to directional accuracy of rendered sound fields. This is true for both loudspeaker and head-tracked headphone listening.

- It has been shown that perception of distance in First and Higher Order Ambisonic sound fields is not different than perception of distance of real-world sound sources

- It has been shown that presence of vision affects the perception of auditory distance. Subjects accept even substantial misalignment of auditory and visual cues before the audio-visual presentation ceases to be perceived as a unity.

- For better realism in headphone listening, rendered sound fields should be stabilised using head movement tracking. Based on the objective localisation criteria, it is possible to optimise the localisation of stabilised sound fields.

7.2 Future Work

The work presented in this thesis pertains to both loudspeaker and, to a greater extent, headphone playback modes. However, it is by no means complete or exhaustive. For example, perceptual evaluation of the full convolution and hybrid convolution methods was done based on subjective sound qualities such as perceived reverberation, source width, clarity and natural timbre. It was also based on former studies on directional discrimination (e.g. objective measures of auditory localisation) as well as novel work presented in this thesis which pertains to the perception of auditory distance. However, to enable the comparison between different auralisation methods, the evaluation was done using static source/listener locations whereas the optimised hybrid convolution method proposed in this thesis is design in particular for the dynamic, real-time and interactive scenarios. That is why, further studies should look at the problem of novel methods of evaluation of dynamic, real-time and interactive VAEs in which both the listener and sound sources are allowed to change their locations.

Also, it must be also emphasised that some assumptions have been made in order to simplify the auralisation process. In particular:

- The transducer-induced signal colouration was neglected. It applies to both direct field recordings of musical instruments/voice and measurement of RIRs. Under this assumption, loudspeaker and microphones used are treated as acoustically transparent, i.e. they do not impose any spectral changes (magnitude and phase) on the transmitted signals. In practice, inverse filtering should be applied to the source material and acoustic responses so that the colouration is minimised [62, 105]. That is why lack of compensation might have resulted in small but noticeable tonal differences between real and virtualised recordings presented in Chapter 4. It is suggested as a future work that the impact of such tonal distortions is investigated in the context of dynamic and interactive acoustic scenes.

- In the case of binaural listening, it was similarly assumed that headphones do not impose any spectral changes on the left and right ear signals. However, Adams and Boland [1] have shown that in reality it is rarely the case and various types of headphones may lead to different levels of ILD and ITD cues distortion. However, they also argue that high quality (open-back) headphones with matched transducers usually exhibit lower levels of binaural signal distortion. That is why, throughout the work performed in this thesis, high-quality open-back headphones (*Sennheiser HD-650* and *AKG K-601*) were used for all the listening tests and demonstrations of auralisations.

For these reasons, further investigations into the proposed rendering scheme are suggested in order to include the above factors and verify their perceptual effects.

Finally, only parametrisation of measured impulse responses has been considered in this work. It was mainly due to the ease of comparison between real or measured and virtual spaces. However, the method is by no means restricted to the data-based auralisation and diffuse parts

of the RIRs could be equally generated using e.g. numerical simulation schemes. If the effects were satisfactory, that could greatly extend the scope of possible applications to include the conceived spaces or spaces that exist only virtually (e.g. in video games).

7.3 Closing Remark

It is hoped that findings of this thesis have applications in spatial audio productions and in particular, those concerned with high levels of plausibility and immersion. At the same time it has to be appreciated that some audio signal processing techniques that are not feasible just yet (like highly accurate real-time simulations based on numerical acoustics) may soon become commonplace making other techniques obsolete. However, if we look at different audio coding schemes, like the popular MP3 formats, history shows that perceptually driven optimisation has always played a fundamental part in various multimedia applications. That is why, investigations into optimisation strategies, even from the point of view of gaining a better understanding of the mechanisms in spatial hearing, and being able to apply it practice, seems to be a perfectly valid justification for the effort taken.

A

Legendre Polynomial Approximation

Basis functions that are orthogonal can be found among different families of polynomials e.g. the Legendre polynomials which are real-valued functions defined in $x \in [-1, 1]$ of a form:

$$P_n = \frac{1}{2^n n!}\left[\frac{d^n}{dx^n}(x^2 - 1)^n\right] \tag{A.1}$$

Because of the presence of the higher order derivatives in its original form Equation A.1 is rather difficult to evaluate. However, computationally stable solutions can be found using several recurrence relations, e.g.

$$\begin{aligned} P_0 &= 0 \\ P_1 &= 1 \\ P_k &= ((2n-1)xP_{n-1} - (n-1)P_{n-2})/n \end{aligned} \tag{A.2}$$

Now when we have our set of orthogonal basis functions it is time to look at how can they be linearly combined together in order to approximate any arbitrary function $f(x)$ defined for $x \in [-1, 1]$. To begin with, we have to appreciate that since the basis functions are orthogonal then each weighting coefficient c_k from the equation 3.42 is simply a measure of how "similar" the current basis function is to the approximated function $f(x)$. This assessment is achieved by integrating the product of the approximated function and the current basis function over the

domain of interest (here $x \in [-1, 1]$). This process is sometimes called a *projection* of a basis function onto the function $f(x)$.

$$c_k = \int_{-1}^{1} f(x) P_k(x) dx \tag{A.3}$$

If both functions are exactly the same, from Equation 3.43 we know that the coefficient will take a value of a constant c. If they are orthogonal then the result will be 0. Anything in between will generate the results within a range $[0, c]$. It is also important to note that in the case of the Legendre polynomials the possible maximum values of c are known and can be expressed as $2/(2n + 1)$.

One of the immediate problems we can see is that to calculate the set of coefficients c_k, a symbolic integration is required. In the digital domain, in order to evaluate the integral it is necessary to apply some numerical approximation. While probabilistic methods like Monte Carlo Integration [180] can be very efficient and arbitrarily accurate especially for multi-dimensional integrals, here we can use a simple method called trapezoid integration which converts the symbolic integration into the following sum [180]:

$$\int_a^b f(x) dx \approx \frac{b-a}{N-1} \left(\frac{1}{2} f(x_1) + \sum_{k=2}^{N-1} f(x_k) + \frac{1}{2} f(x_N) \right) \tag{A.4}$$

where k represents N equally spaced points in the range $[a, b]$. The convenience of this approach is that k can simply be the elements of the N-length vector containing the $f(x)$. Also, the higher the N, the more accurate the approximation will be.

Before the coefficients can be used as weights for the n-term polynomial sum, they first need to be normalised in order to ensure the orthonormality property (that $c_k \leq 1$). For this to happen, it is convenient to construct a matrix $\mathbf{H}$ that will contain all the possible integrated products of polynomials up to order n.

$$H_{m+1,n+1} = \int_{-1}^{1} P_m(x) P_n(x) dx \tag{A.5}$$

Because of the orthogonality of the Legendre polynomials, the matrix $\mathbf{H}$ will be diagonal and easy to invert. Then, the coefficient normalisation procedure will simply comprise a multiplication of the inverse of the matrix $\mathbf{H}$ with the coefficient vector $\mathbf{c}$

$$\mathbf{b} = \mathbf{H}^{-1} \mathbf{c} \tag{A.6}$$

Lastly, the polynomials P_k need to be scaled by coefficients b_k and added together up to and including some pre-determined order n. In this way, $\tilde{f}(x)$ is obtained which constitutes an approximation of the original function $f(x)$

$$\tilde{f}(x) = \sum_{k=1}^{n} b_k P_k(x) \tag{A.7}$$

As shown, Legendre polynomials constitute a subset of orthogonal basis functions that can be used to approximate arbitrary piecewise continuous functions defined within the range $[-1, 1]$. L^2 norm can be used in order to measure the error of the reconstruction. The idea is to minimise the square of the difference between a given function $f(x)$ and its approximation. The norm of the difference between the original function and the sum of the basis functions up to and including order N is approaching 0 as N goes to infinity. It means that the series expansion offers the unique best available approximation of the function in the L^2 norm sense among all the polynomials of degree N or less.

B

Spherical Harmonics Approximation

Detailed explanation of the Spherical harmonics approximation routines are not vastly present in the literature of audio and acoustics. Luckily, there have been multiple tutorials published that employ spherical harmonics in order to approximate light intensity functions in computer graphics [84, 205], and these were used as a basis in this work.

So far we have shown the approximation of a function with a single variable using finite expansion of orthogonal basis functions, namely Legendre polynomials. Before we expand our considerations to more complex examples in the multi-dimensional space it is useful to introduce two more concepts, or rather two more sets of orthogonal basis functions. First set is used extensively in the Fourier series approximation of periodic functions and is based on sines and cosines. In terms of the orthogonality requirement it can be shown that

$$\int_{-\pi}^{\pi} sin(mx)sin(nx) = \int_{-\pi}^{\pi} cos(mx)cos(nx) = \pi\delta_{mn} \tag{B.1}$$

where δ_{mn} is a Kronecker delta which takes values of 1 if $n = m$ or 0 if $n \neq m$. Also

$$\int_{-\pi}^{\pi} sin(mx)cos(nx) = \int_{-\pi}^{\pi} sin(mx) = \int_{-\pi}^{\pi} cos(mx) = 0 \tag{B.2}$$

Another set of useful basis functions is the associated Legendre polynomials which extends the basic set of Legendre polynomials we explored in Appendix A. For each order, there are

$m = n + 1$ polynomials which are defined as:

$$P_n^m = \frac{(-1)^m}{2^n n!}\sqrt{(1-x^2)^m}\frac{d^{n+m}}{dx^{n+m}}(x^2-1)^n \tag{B.3}$$

Similarly as in the case of Legendre polynomials, because of the presence of the higher order derivatives in its original form Equation B.3 is rather difficult to evaluate. Again, computationally stable solutions can be found using several recurrence relations, e.g.

$$\begin{aligned} P_m^m &= (-1)^m(2m-1)!!(1-x^2)^{\frac{m}{2}} \\ P_{m+1}^m &= x(2m+1)P_m^m \\ (n-m)P_n^m &= x(2n+1)P_{n-1}^m - (n+m-1)P_{n-2}^m \end{aligned} \tag{B.4}$$

These two basis function sets were introduced because they will be used next in order to derive Spherical Harmonics - functions that allow us to approximate arbitrary functions defined on a 2-D surface of a sphere.

Spherical harmonics form a set of orthogonal basis functions that are defined on a surface of a unity sphere. It is useful to think of spherical harmonics in terms of spherical coordinates using longitudinal and attitudinal angles ϕ and θ respectively. Here, the following convention is used in order to define the domain of spherical harmonic functions:

$$Y_n^m(\phi,\theta), \text{where } 0 \leq \phi < 2\pi, \tfrac{-\pi}{2} \leq \theta < \tfrac{\pi}{2}, n \in R^+, -n \leq m \leq nW \tag{B.5}$$

Spherical harmonics combine together two sets of orthogonal basis functions: associated Legendre polynomials and trigonometric functions (sines and cosines). A spherical harmonic function of order n and degree m can be constructed using the following equation:

$$Y_n^m(\phi,\theta) = \begin{cases} A_n^m \cos(m\phi)P_n^m\cos(\theta) & \text{if } m > 0 \\ A_n^m \sin(m\phi)P_n^m\cos(\theta) & \text{if } m < 0 \\ \frac{1}{\sqrt{2}}A_n^0 P_n^0 \sin(\theta) & \text{if } m = 0 \end{cases} \tag{B.6}$$

where n is the order and m is the degree of the spherical harmonic and P_n^m are the associated Legendre functions. The term A_n^m is used to normalise each function Y_n^m and in this thesis is calculated using the following formula:

$$A_n^m = \sqrt{\frac{2n+1}{4\pi}\frac{(n-|m|)!}{n+|m|)!}} \tag{B.7}$$

For each order n there are $(2n+1)$ spherical harmonics and the total number of harmonics up to and including specific order n equals $(n+1)^2$. S

In a similar way as the Legendre polynomials can be used to approximate piecewise continuous functions of one variable, the spherical harmonics can be used in approximation of a function defined in three dimensions. Often, these functions are defined by two angular arguments and assuming the unit length radius, e.g. $p(\phi, \theta, 1)$ or simply $p(\phi, \theta)$. A perfect reconstruction of a function $p(\phi, \theta)$ requires the use of infinite series expansion of spherical harmonics:

$$p(\phi, \theta) = \sum_{n=0}^{\infty} \sum_{m=-\infty}^{\infty} c_n^m Y_n^m(\phi, \theta) \tag{B.8}$$

where c_n^m are the spherical harmonics coefficients resulting from projecting the original function $p(S)$ onto the spherical basis function $Y_n^m(S)$:

$$c_n^m = \int_S p(S) Y_n^m(S) dS \tag{B.9}$$

and S is some domain of interest, e.g. a surface of a unity sphere. In all other cases (when $n < \infty$) we are talking about the n^{th} order approximation of the function $p(\phi, \theta)$ which can be written as:

$$\tilde{p}_(\phi, \theta) = \sum_{n=0}^{n_{max}} \sum_{m=-n}^{n} c_n^m Y_n^m(\phi, \theta) \tag{B.10}$$

Again, as in the example of a simple approximation of 1-D functions using Legendre polynomials, for most practical uses the integral in Equation B.9 needs to be evaluated numerically. Multiple examples of numerical methods for approximation of the integration process can be found in [180]. Here we refer to a method known as *Monte Carlo Integration* [180] which has its roots in the probabilistic theory. Assuming a uniform distribution of N sample points on a unity sphere, Equation B.9 can be approximated with the following sum:

$$\tilde{c}_n^m = \frac{4\pi}{N} \sum_{n=0}^{n_{max}} \sum_{m=-n}^{n} p(S) Y_n^m(S) \Delta S \tag{B.11}$$

Since in the rest of this work we will be dealing only with numerical approximations of reconstructed functions and their spherical harmonic coefficients, we can drop the tilde sign in the $\tilde{c}_n^m$ notation. It is also useful, particularly from the point of view of performing numerical computations, to sequence the spherical harmonic coefficients in a specific order. This will create a 1-D vector of coefficients instead of a 2-D matrix. First denote:

$$Y_i(\phi,\theta) = Y_n^m(\phi,\theta), \qquad \text{where } i = n(n+1)+m \tag{B.12}$$

This convention, referred to as Ambisonic Channel Number (ACN) has been already proposed in [9]. Thus, the coefficients c_n^m can be rewritten using the simplified notation as:

$$\tilde{c}_i = \frac{4\pi}{N} \sum_{k=0}^{(n+1)^2} p(S) Y_i(S) dS \tag{B.13}$$

Similarly as in the 1-D example, before the coefficients c_i can be used for reconstruction, we have to make sure that the orthonormality property of the basis function set is met. In order to do this, we have to multiply each coefficient vector $\mathbf{c}$ with the inverse of the matrix $\mathbf{H}$ constructed from the following terms:

$$H_{i,j} = \int_S Y_i(x) Y_j(x) dx, \qquad \text{where } i,j = [0,...,(n+1)^2] \tag{B.14}$$

which again can be approximated using the Monte Carlo integration [180] as:

$$\tilde{H}_{i,j} = \frac{4\pi}{N} \sum_i \sum_j Y_i Y_j \tag{B.15}$$

Finally, the reconstructed function $\tilde{p}(S)$ can be expressed as

$$\tilde{p}(S) = \sum_{i=0}^{(n+1)^2} b_i Y_i(x) \tag{B.16}$$

where b_i is the i^{th} term of a vector $\mathbf{b}$ of normalised spherical harmonics coefficients.

$$\mathbf{b} = \mathbf{H}^{-1}\mathbf{c} \tag{B.17}$$

C

Radiation patterns of instruments used for Auralisation in Chapter 4

Table C.1: Vocal radiation patterns and their HOA approximations

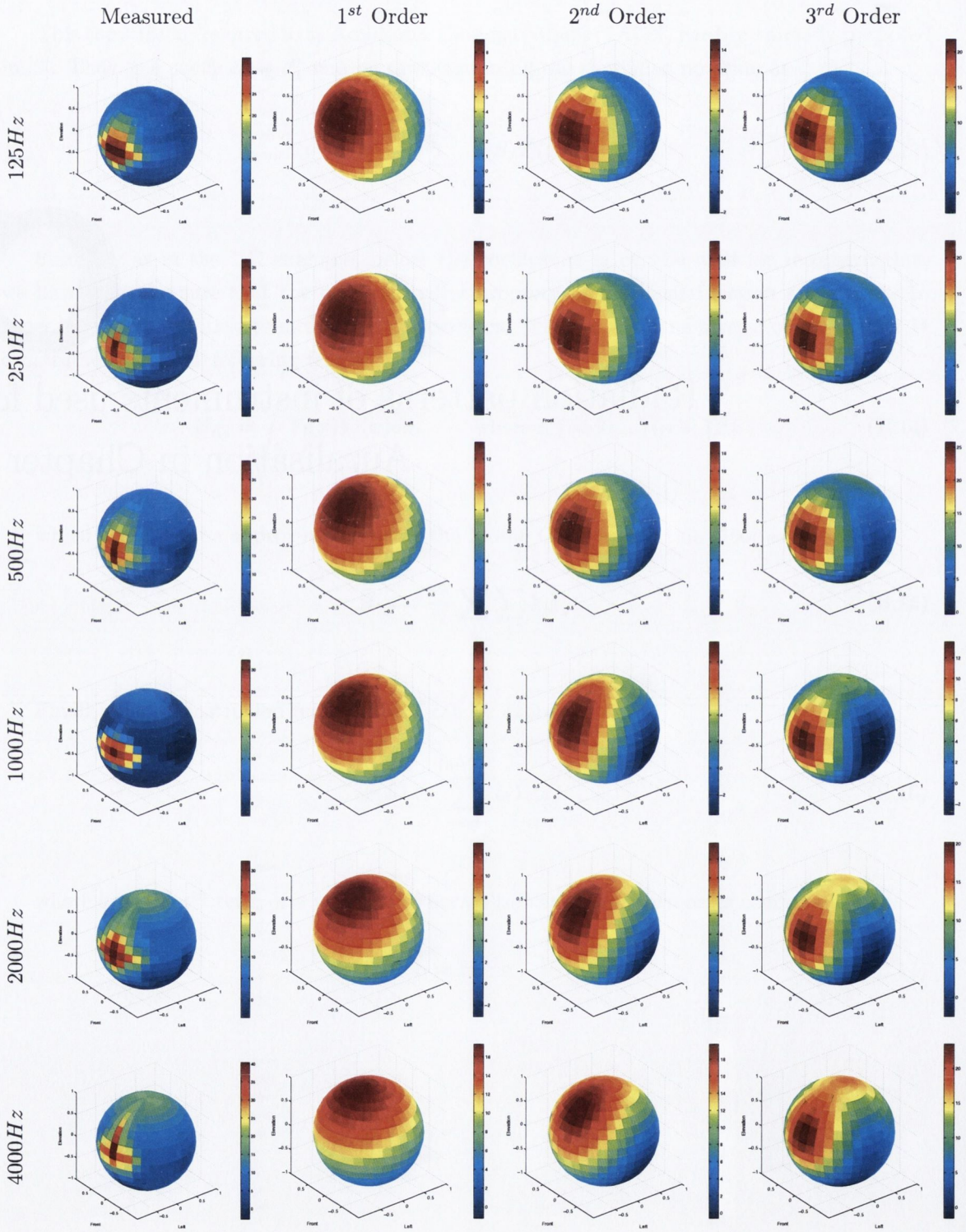
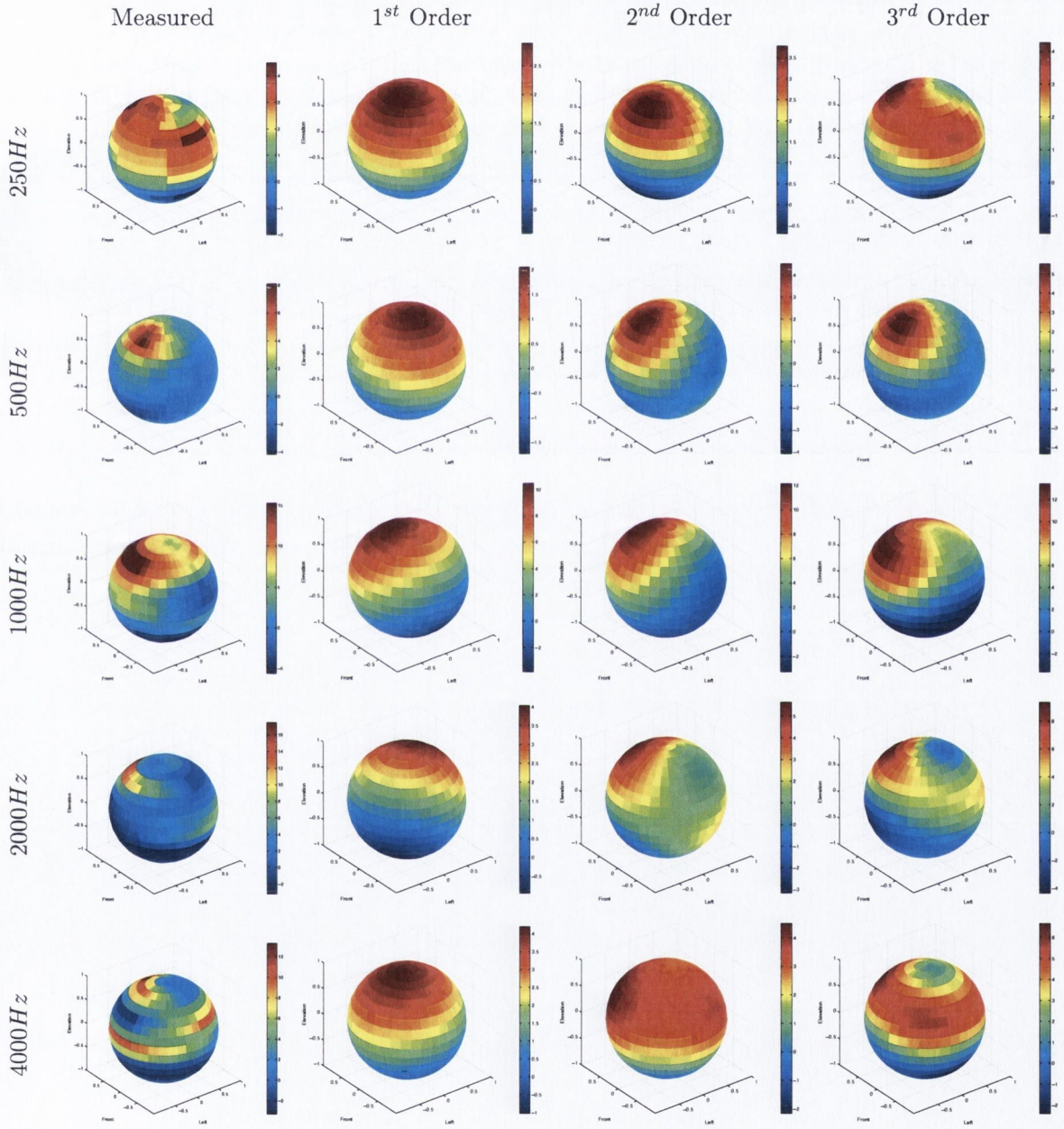


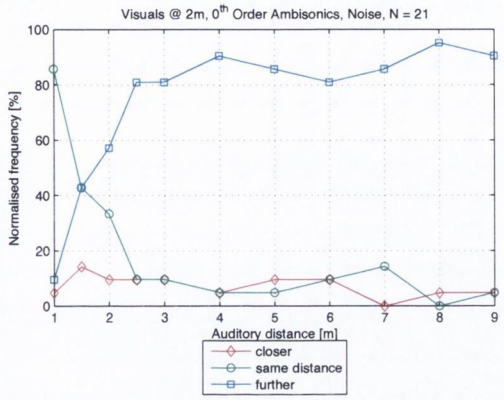
Table C.2: Violin radiation patterns and their HOA approximations



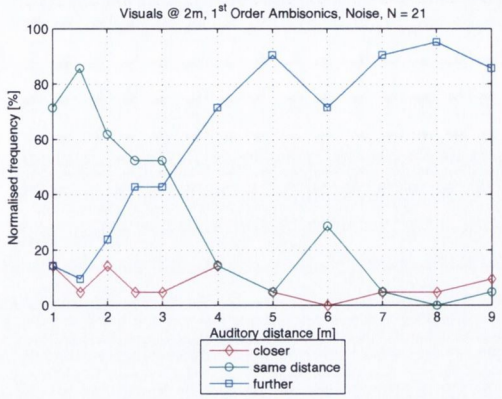
D

Results of the Audio-Visual Distance Perception Study

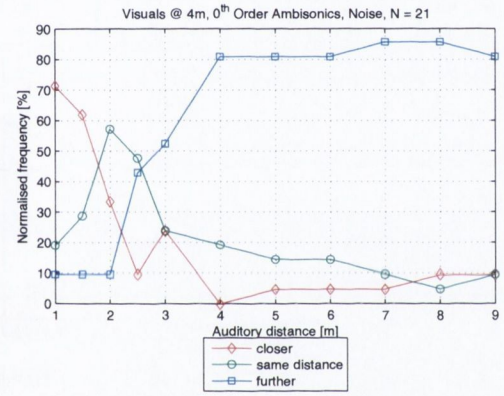
In this Appendix we present results obtained the study on the perception of audio-visual incongruence described in Chapter 5. Here, the results are presented separately for two stimulus types used in the test: female speech and pink noise bursts.



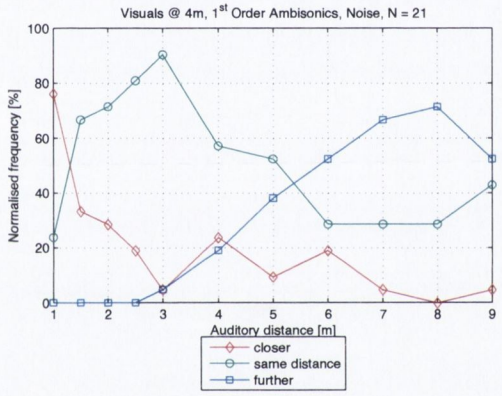
(a)



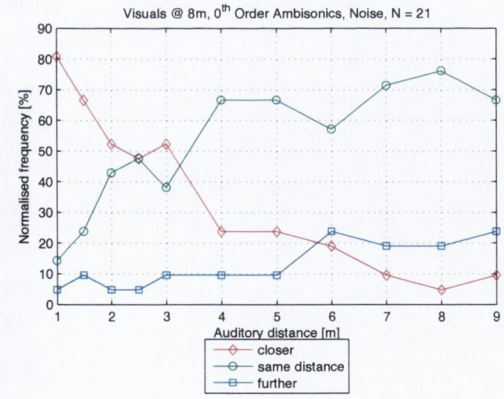
(d)



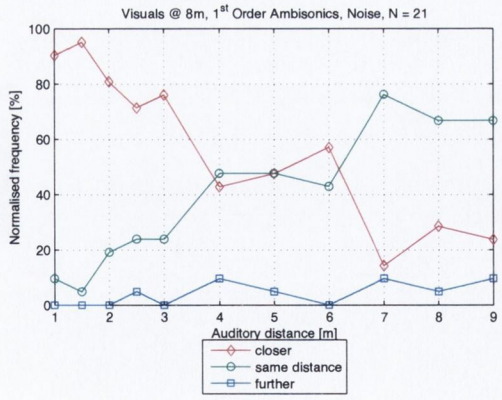
(b)



(e)

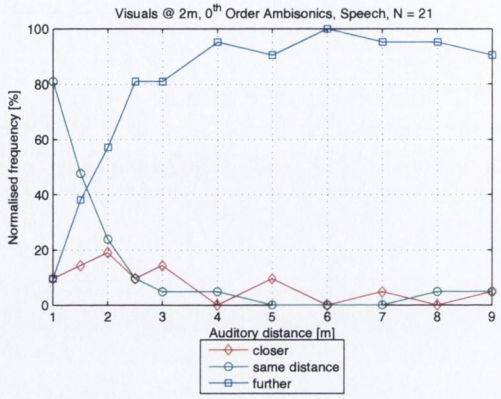


(c)

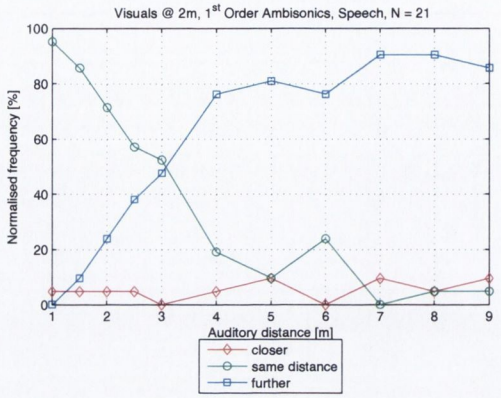


(f)

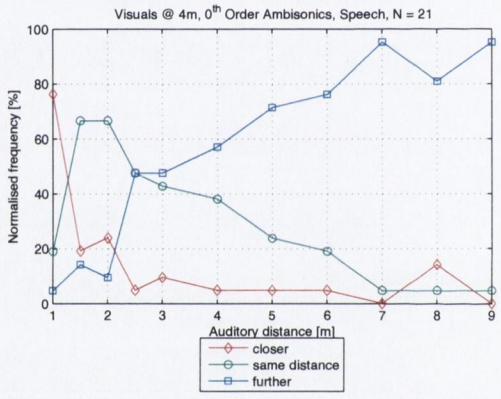
Figure D.1: Localisation of audio sources with respect to visual objects for pink noise stimulus



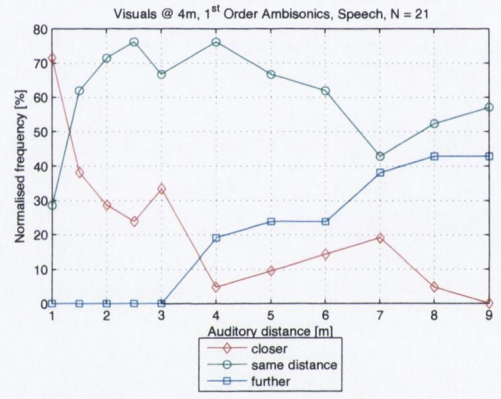
(a)



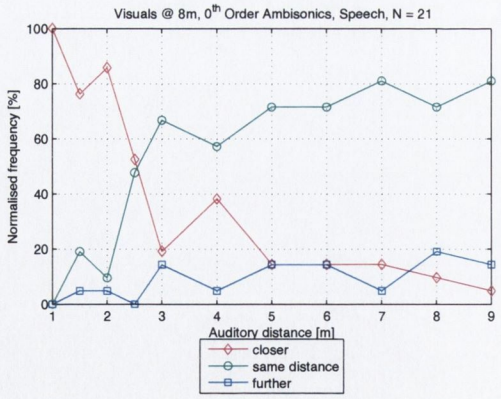
(d)



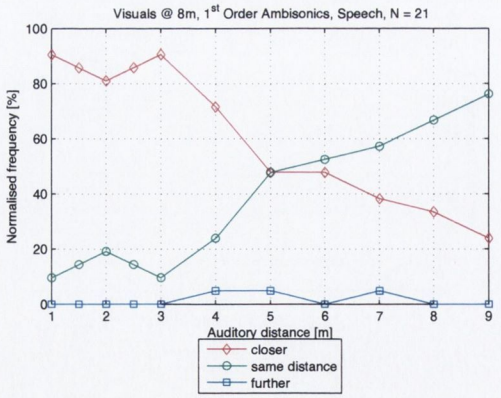
(b)



(e)



(c)



(f)

Figure D.2: Localisation of audio sources with respect to visual objects for female speech stimulus

E

MATLAB program used for gain optimisation in Chapter 6

```
%% Non-linear unconstrained Gain function optimisation for ITU 5-loudspeakr setup
% Author: Marcin Gorzel, Trinity College Dublin, October 2012
clear all
close all
clc
%matlabpool close

%% Import known coefficients (initial guesses)
importfile('g.mat'); %Use to Import 5x360 matrix with initial guesses
 %for 5 loudspekar gains and 360 angular (panning) locations
guess = g;
clear g
result = zeros(size(guess));

%% Calculate partial cost functions
PHI = [30 330 0 110 250]*pi/180; %Define loudspeaker angles
syms g1 g2 g3 g4 g5 %Define variables
g = [g1;g2;g3;g4;g5]; %Column vector of defined variables g
Pe = g1.^2+g2.^2+g3.^2+g4.^2+g5.^2; %Calculate total energy for a given angle
```

```
Pv = g1 + g2 + g3 + g4 + g5; %Calculate total velocity
ex = (g1.^2.*cos(PHI(1)) + g2.^2.*cos(PHI(2))... %x-component of the energy vector
    + g3.^2.*cos(PHI(3))  + g4.^2.*cos(PHI(4))...
    + g5.^2.*cos(PHI(5)))/Pe;
ey = (g1.^2.*sin(PHI(1)) + g2.^2.*sin(PHI(2))... %y-component of the energy vector
    + g3.^2.*sin(PHI(3))  + g4.^2.*sin(PHI(4))...
    + g5.^2.*sin(PHI(5)))/Pe;
vx = (g1.*cos(PHI(1)) + g2.*cos(PHI(2))... %x-component of the velocity vector
    + g3.*cos(PHI(3))  + g4.*cos(PHI(4))...
    + g5.*cos(PHI(5)))/Pv;
vy = (g1.*sin(PHI(1)) + g2.*sin(PHI(2))... %y-component of the velocity vector
    + g3.*sin(PHI(3))  + g4.*sin(PHI(4))...
    + g5.*sin(PHI(5)))/Pv;
phi_e = 2 * atan( (sqrt(ex.^2 + ey.^2) - ex) ./ ey ); %Energy vector direction [rad]
re = sqrt(ex.^2+ey.^2); %Energy Vector magnitude
phi_v = 2 * atan( (sqrt(vx.^2 + vy.^2) - vx) ./ vy ); %Velocity vector direction [rad]
rv = sqrt(vx.^2+vy.^2); %Velocity Vector magnitude

%% Main loop
matlabpool 4 %Optimisation for 4-core processors (MATLAB workers)
parfor i = 0:359;

%% Current panning angles and speaker gains
phi = i*pi/180; %panning angle in radians

%% Compute partial quadratic cost functions and the total cost function
a1=1; a2=1; a3=1; a4=1; a5=1; a6=1; %Weighting coefficients
J1 = a1*(phi_e-phi).^2;
J2 = a2*(phi_v-phi).^2;
J3 = a3*(1-re).^2;
J4 = a4*(1-rv).^2;%a4*(1./(rv-.08) + 1./(1.2-rv));
J5 = a5*(1-Pe).^2;
J6 = a6*(phi_e-phi_v).^2;
J = J1+J2+J3+J4+J5+J6; %Total cost fucntion we want to minimize

%% Compute Gradient and Hessian matrices
gradJ = jacobian(J,g).'; % column gradf
hessJ = jacobian(gradJ,g); %Hessian matrix
%subs(hessJ,{g1,g2,g3,g4,g5},{0,0,1,0,0});
```

```
%% Run the solver
Jh = matlabFunction(J,gradJ,hessJ,'vars',{g});
options = optimset('GradObj','on','Hessian','on', ...
    'Display','final', 'UseParallel', 'always');

[xfinal fval exitflag output] = fminunc(Jh,guess(:,i+1),options)
result(:,i+1) = xfinal; %write results to the array
i
end
matlabpool close %Close all active workers
```

Bibliography

[1] S. Adams and F. Boland. On the Distortion of Binaural Localization Cues using Headphones. In *Proceedings of the Irish Signal and Systems Conference (ISSC'10)*, Cork, Ireland, 2010.

[2] F. Adriaensen. Near Field filters for Higher Order Ambisonics.

[3] J. Ahrens and S. Spors. 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, NY, USA, October 2007.

[4] C. Ainlay, J. Chiccarelli, B. Clearmountain, F. Filipetti, L. A. Jones, R. Kaplan, J. Levison, B. Ludwig, G. Massenburg, H. Massey, H. Neuberger, P. Ramone, E. Scheiner, E. Schilling, A. Schmitt, J. Skillen, and P. Stubblebine. *Recommendations For Surround Sound Production.* The Recording Academy's Producers & Engineers Wing, 2004.

[5] V. Algazi and R. Duda. Headphone-Based Spatial Sound. *IEEE Signal Processing Magazine*, 28(1):33–42, 2011.

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

[7] V. R. Algazi, R. O. Duda, and D. M. Thompson. The use of head-and-torso models for improved spatial sound synthesis. In *Proceedings of the 113th Audio Engineering Society Convention*, Los Angeles, CA, USA, 2002.

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

[9] ambisonics.ch. Ambisonic channels, accessed September 9, 2012. `http://ambisonics.ch/standards/channels/`.

[10] J. Anderson. Introducing... the Ambisonics toolkit. Poster, June 2009.

[11] M. Anderson, D. Begault, M. Godfroy, J. D. Miller, A. Roginska, and E. Wenzel. Design and Verification of HeadZap, a Semi-automated HRIR Measurement System. In *Proceedings of the 120th Audio Engineering Society Convention*, Paris, France, May 2006.

[12] I. Arvers. Serious games. *Digitalarti Magazine #0*, 2009.

[13] D. H. Ashmead, D. L. Davis, and A. Northington. Contribution of listeners' approaching motion to auditory distance perception. *Journal of Experimental Psychology. Human Perception and Performance.*, 21(2):239–256, 1995.

[14] C. Avendano, V. Algazi, and R. O. Duda. A head-and-torso model for low-frequency binaural elevation effects. In *Proceedings of the IEEE Workshop on Application of Signal Processing to Audio and Acoustics*, pages 179–182, New Platz, NY, 1999.

[15] G. Ballou. *Handbook for Sound Engineers.* Handbook for Sound Engineers. Focal, 2008.

[16] J. Bamford. An Analysis of Ambisonics Sound Systems of First and Second Order. M.Sc. Thesis, University of Waterloo, Waterloo, Ontario, Canada, 1995.

[17] J. L. Barbour. Elevation Perception: Phantom Images in the Vertical Hemi-sphere. In *Proceedings of the Audio Engineering Society 24th International Conference: Multichannel Audio, The New Reality*, Banff, Alberta, Canada, June 2003.

[18] D. Begault, E. M. Wenzel, M. Godfroy, J. D. Miller, and M. R. Anderson. Applying Spatial Audio to Human Interfaces: 25 Years of NASA Experience. In *Proceedings of the 40th Audio Engineering Society Conference International Conference: Spatial Audio: Sense the Sound of Space*, Tokyo, Japan, Oct 2010.

[19] D. R. Begault, E. M. Wenzel, and M. R. Anderson. 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):904–917, 2001.

[20] E. Benjamin, A. Heller, and R. Lee. Design of ambisonic decoders for irregular arrays of loudspeakers by non-linear optimization. In *Proceedings of the 129th Audio Engineering Society Convention*, San Francisco, CA, USA, November 2010.

[21] E. Benjamin, R. Lee, and A. Heller. Localization in Horizontal-Only Ambisonic Systems. In *Proceedings of the 121st Audio Engineering Society Convention*, San Francisco, CA, USA, 2006.

[22] A. J. Berkhout. A Holographic Approach to Acoustic Control. *Journal of the Audio Engineering Society*, 36:977–995, 1988.

[23] S. Bertet. *Formats audio 3D hiérarchiques : caractérisation objective et perceptive des systèmes Ambionics dordres supérieurs.* PhD thesis, INSA, Lyon, France, 2009.

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

[25] M. A. Blanco, M. Flrez, and M. Bermejo. Evaluation of the rotation matrices in the basis of real spherical harmonics. *Journal of Molecular Structure: THEOCHEM*, 419(1-3):19 – 27, 1997.

[26] J. Blauert. *Spatial Hearing : The Psychophysics of Human Sound Source Localization, 2nd edition.* MIT Press, 1997.

[27] J. Blauert, H. Lehnert, J. Sahrhage, and H. Strauss. An interactive virtual-environment generator for psychoacoustic research. i: Architecture and implementation. *Acta Acustica united with Acustica*, 86(1):94–102, 2000-01-01T00:00:00.

[28] Blender Foundation. Blender 3d, accessed October 9, 2012. `http://www.blender.org`.

[29] A. D. Blumlein. Improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems. Technical report, UK Patent no. GB19310034657 19311214, 1933.

[30] M. M. Boone and E. N. G. Verheijen. Multichannel Sound Reproduction Based on Wavefield Synthesis. In *Proceedings of the 95th Audio Engineering Society Convention*, New York, NY, USA, October 1993.

[31] J. Brereton, D. Murphy, and D. Howard. A loudspeaker-based room acoustics simulation for real-time musical performance. In *25th AES UK Conference: Spatial Audio in Today's 3D World in association with the 4th International Symposium on Ambisonics and Spherical Acoustics*, York, UK, March 2012.

[32] A. W. Bronkhorst and T. Houtgast. Auditory distance perception in rooms. *Nature*, 397:517–520, February 1999.

[33] S. Browne. Hybrid Reverberation Algorithm using Truncated Impulse Response Convolution and Recursive Filtering. M.Sc. Thesis, University of Miami, Miami, Florida, USA, 2001.

[34] D. Brungart, B. D. Simpson, R. L. McKinley, A. J. Kordik, R. C. Dallman, and D. A. Ovenshire. The Interaction Between Head-Tracker Latency, Source Duration, and Response Time in the Localization of Virtual Sound Sources. In S. Barrass and P. Vickers, editors, *Proceedings of ICAD 04 Tenth Meeting of the International Conference on Auditory Display*, Sydney, Australia, July 2004. International Community for Auditory Display.

[35] D. S. Brungart and K. R. Scott. The effects of production and presentation level on the auditory distance perception of speech. *The Journal of the Acoustical Society of America*, 110(1):425–440, 2001.

[36] A. F. Carver. *The development of sacred polychoral music to the time of Schutz / Anthony F. Carver.* Cambridge University Press, Cambridge [Cambridgeshire] ; New York :, 1988.

[37] J. S. Chan, C. Maguinness, D. Lisiecka, C. Ennis, C. O'Sullivan, and F. N. Newell. Comparing audiovisual distance perception in various real and virtual environments. In *Proceedings of the 32nd European Conference on Visual Perception*, Regensburg, Germany, 2009.

[38] D. W. Chandler and D. W. Grantham. Minimum audible movement angle in the horizontal plane as a function of stimulus frequency and bandwidth, source azimuth and velocity, and number of ears. *The Journal of the Acoustical Society of America*, 87(S1):S64–S64, 1990.

[39] C. H. Choi, J. Ivanic, M. S. Gordon, and K. Ruedenberg. Rapid and stable determination of rotation matrices between spherical harmonics by direct recursion. *The Journal of Chemical Physics*, 111(19):8825–8831, 1999.

[40] T. Cooper, Duane H.; Shiga. Discrete-matrix multichannel stereo. *Journal of the Audio Engineering Society*, 20:346–360, June 1972.

[41] Core Sound. Tetramic, accessed September 24, 2012. `http://www.core-sound.com/TetraMic/1.php`.

[42] P. G. Craven. Continuous surround panning for 5-speaker reproduction. In *Proceedings of the Audio Engineering Society 24th International Conference: Multichannel Audio, The New Reality*, Banff, Alberta, Canada, June 2003.

[43] P. G. Craven and M. A. Gerzon. Coincident microphone simulation covering three dimensional space and yielding various directional outputs, 1977. U.S.Patent Number 4,042,779.

[44] E. Czerwinski, A. Voishvillo, S. Alexandrov, and A. Terekhov. Propagation distortion in sound systems: Can we avoid it? *Journal of the Audio Engineering Society*, 48(1/2):30–48, 2000.

[45] B. Dalenbäck and M. Strömberg. Real time walkthrough auralization - the first year. *Proceedings of the Institute of Acoustics*, 28(2), 2006.

[46] P. V. Damaske and B. Wagener. Richtungshörversuche über einen nachgebildeten kopf [investigations of directional hearing using a dummy head]. *Acustica*, 21:30–35, 1969.

[47] J. Daniel. *Acoustic field representation, application to the transmission and the reproduction of complex sound environments in a multimedia context (English translation).* Ph.D. Thesis, University of Paris, Paris, France, 2001.

[48] J. Daniel. Spatial Sound Encoding Including Near Field Effect: Introducing Distance Coding Filters and a Viable, New Ambisonic Format. In *Proceedings of the Audio Engineering Society 23rd Conference*, Copenhagen, Denmark, May 2003.

[49] J. Daniel, J.-B. Rault, and J.-D. Polack. Ambisonics Encoding of Other Audio Formats for Multiple Listening Conditions. In *Proceedings of the 105th Audio Engineering Society Convention*, San Francisco, CA, USA, 1998.

[50] M. DeLoura. Game Engines and Middleware, 2011.

[51] Denon Records. Anechoic orchestral music recording. Audio CD, 1995. ASIN: B0000034M9.

[52] N. Dodgson. Variation and extrema of human interpupillary distance. In *Proceedings of the SPIE Stereoscopic Displays and Applications XI*, volume 5291, pages 36–46, San Jose, California, USA, 2004.

[53] Dolby. Home Theater Sound Technologies, accessed October 21, 2012. `http://www.dolby.com/us/en/consumer/technology/home-theater/listing.html`.

[54] DTS. DTS Technologies, accessed October 21, 2012. `http://www.dts.com/professionals/sound-technologies/audio-formats.aspx`.

[55] N. I. Durlach, A. Rigopulos, X. D. Pang, W. S. Woods, A. Kulkarni, H. S. Colburn, and E. M. Wenzel. On the externalization of auditory images. *Presence: Teleoper. Virtual Environ.*, 1(2):251–257, May 1992.

[56] J. Eargle. *The Microphone Book: From Mono to Stereo to Surround - a Guide to Microphone Design and Application.* Taylor & Francis, 2004.

[57] C. F. Eyring. Reverberation Time in "Dead" Room. *The Journal of the Acoustical Society of America*, 1(2A):168–168, 1930.

[58] A. Farina. Software Implementation of B-Format Encoding and Decoding. In *Proceedins of the 104th Audio Engineering Society Convention*, Amsterdam, The Netherlands, May 1998.

[59] A. Farina. Simultaneous measurement of impulse response and distortion with a swept-sine technique. In *Proceedings of the 108th Audio Engineering Society Convention*, Paris, France, 2000.

[60] A. Farina. Acoustic quality of theatres: Correlations between experimental measures and subjective evaluations. *Journal of Applied Acoustics*, 62:889–916, 2001.

[61] A. Farina and R. Ayalon. Recording concert hall acoustics for posterity. In *Proceedings of the Audio Engineering Society 24th International Conference, Multichannel Audio: The New Reality*, Banff, Alberta, Canada, 2003.

[62] A. Farina, E. Ugolotti, A. Bellini, G. Cibelli, and C. Morandi. Inverse numerical filters for linearisation of loudspeaker's response. In *Proceedings of the First COST-G6 Workshop on Digital Audio Effects (DAFX98)*, pages 12–16, Barcelona, Spain, 1998.

[63] P. Fellgett. Ambisonics. part one: General system description. *Studio Sound*, 17(8):24–26, 28, 40, August 1975.

[64] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall. The DARPA Speech Recognition Research Database: Specifications and Status. In *Proceedings of DARPA Workshop on Speech Recognition*, 1986.

[65] H. Fletcher. Stereophonic recording and reproducing system. *Journal of the Society of Motion Picture and Television Engineers*, 61(3):355–363, 1953.

[66] S. Foster, E. Wenzel, and R. Tayior. Real Time Synthesis of Complex Acoustic Environments. In *Proceedings of the IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 1991.

[67] A. Foteinou and D. Murphy. Evaluation of the Psychoacoustic Perception of Geometric Acoustic Modeling-Based Auralization. In *Proceedings of the 130th Audio Engineering Society Convention*, London, UK, May 2011.

[68] N. V. Franssen. *Some considerations on the mechanism of directional hearing.* Ph.D. Thesis, Technische Hogeschool, Delft, The Netherlands, 1960.

[69] J. Frenette. Reducing artificial reverberation algorithm requirements using time-variant feedback delay networks. Master's thesis, University of Miami, December 2000.

[70] M. B. Gardner. Distance estimation of O° or apparent O° oriented speech signals in anechoic space. *Journal of the Acoustical Society of America*, 45(1):47–53, 1969.

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

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

[73] M. A. Gerzon. Ambisonics. Part Two: Studio techniques. *Studio Sound*, 17(8):20–22, 40, August 1975.

[74] M. A. Gerzon. The design of precisely coincident microphone arrays for stereo and surround sound. In *Proceedings of the 50th Audio Engineering Society Convention*, London, UK, 1975.

[75] M. A. Gerzon. Practical Periphony: The Reproduction of Full-Sphere Sound. In *Proceedings of the 65th Audio Engineering Society Convention*, London, UK, February 1980.

[76] M. A. Gerzon. General metatheory for auditory localisation. In *Proceedings of the 92nd Audio Engineering Society Convention*, Vienna, Austria, March 1992.

[77] M. A. Gerzon. Panpot laws for multispeaker stereo. In *Proceedings of the 92nd Audio Engineering Society Convention*, Vienna, Austria, March 1992.

[78] M. A. Gerzon. Psychoacoustic Decoders for Multispeaker Stereo and Surround Sound. In *Proceedings of the 93rd Audio Engineering Society Convention*, October 1992.

[79] M. A. Gerzon and G. J. Barton. Ambisonics Decoders For HDTV. In *Proceedings of the 92nd Audio Engineering Society Convention*, Vienna, Austria, March 1992.

[80] F. Giron. *Investigations about the directivity of sound sources.* Ph.D. Thesis, Ruhr-Universität, Bochum, Germany, 1996.

[81] M. Gorzel. Investigations into OPSI with height. M.Phil. Thesis, Trinity College Dublin, Dublin, Ireland, 2009.

[82] M. Gorzel, G. Kearney, H. Rice, and F. Boland. On the Perception of Dynamic Sound Sources in Ambisonic Binaural Renderings. In *Proceedings of the Audio Engineering Society 41st International Conference: Audio for Games*, February 2011.

[83] T. Y. Grechkin, T. D. Nguyen, J. M. Plumert, J. F. Cremer, and J. K. Kearney. How does presentation method and measurement protocol affect distance estimation in real and virtual environments? *ACM Transactions on Applied Perception*, 7(4):26:1–26:18, July 2010.

[84] R. Green. Spherical harmonic lighting: The gritty details. *Archives of the Game Developers Conference*, 2003.

[85] M. Gröhn, T. Lokki, and T. Takala. Static and Dynamic Sound Source Localization in a Virtual Room. In *Proceedings of the Audio Engineering Society 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, Espoo, Finland, June 2002.

[86] C. Guastavino and B. F. G. Katz. Perceptual evaluation of multi-dimensional spatial audio reproduction. *Journal of the Acoustic Society of America*, 116(2):1105–1115, 2004.

[87] W. Gulick. *Hearing: physiology and psychophysics.* Oxford University Press, 1971.

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

[89] J. D. Harris and R. L. Sergeant. Monoaural/binaural minimum audible angles for a moving sound source. *Journal of Speech and Hearing Research*, 14:618–629, 1971.

[90] W. M. Hartmann. Localization of sound in rooms. *The Journal of the Acoustical Society of America*, 74(5):1380–1391, 1983.

[91] W. M. Hartmann. How we localize sound. *Physics Today*, 52(11):24–29, 1999.

[92] W. M. Hartmann and B. Rakerd. Localization of sound in rooms. IV: The Franssen effect. *The Journal of the Acoustical Society of America*, 86(4):1366–1373, October 1989.

[93] B. G. Haustein and W. Schirmer. Messeinrichtung zur untersuchung des richtungslokalisationsvermoegens [a measuring apparatus for the investigation of faculty of directional localization]. *Hochfrequenztechnik und Elektroakustik*, 1970.

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

[95] P. M. Hofman, A. J. V. Opstal, and J. G. A. V. Riswick. Relearning sound localization with new ears. *The Journal of the Acoustical Society of America*, 105(2):1035–1035, 1999.

[96] J. Huopaniemi, L. Savioja, and T. Takala. DIVA virtual audio reality system. In *Proceedings of the International Conference on Auditory Display (ICAD'96)*, pages pp. 111–116, Palo Alto, CA, 1996.

[97] IRCAM. Listen HRTF database. `http://recherche.ircam.fr/equipes/salles/listen/index.html`, 2003. Accessed July 4, 2012.

[98] ISO 3382-1:2009. Acoustics – measurement of room acoustic parameters – part 1: Performance spaces. International Standards Organisation, 2009.

[99] ITU-R BS.1284-1. General methods for the subjective assessment of sound quality. International Telecommunications Union, 1997.

[100] ITU-R BS.1770-2. Algorithms to measure audio programme loudness and true-peak audio level. International Telecommunications Union, 2011.

[101] ITU-R BW.1116. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. International Telecommunications Union, 1997.

[102] J. Ivanic and K. Ruedenberg. Rotation matrices for real spherical harmonics. Direct determination by recursion. *The Journal of Physical Chemistry*, 100(15):6342–6347, 1996.

[103] J. Ivanic and K. Ruedenberg. Additions and corrections: Rotation matrices for real spherical harmonics. Direct determination by recursion. *The Journal of Physical Chemistry*, 102(45):9099–9100, 1998.

[104] R. Jacques, B. Albrecht, D. de Vries, and F. Melchior. An Approach for Multichannel Recording and Reproduction of Sound Source Directivity. In *Proceedings of the Audio Engineering Society 119th Convention*, New York, NY, USA, October 2005.

[105] G. Kearney. *Auditory Scene Synthesis using Virtual Acoustic Recording and Reproduction.* Ph.D. Thesis, Trinity College Dublin, Dublin, Ireland, 2010.

[106] G. Kearney, M. Gorzel, F. Boland, F. Smyth, D. Lennon, and H. Rice. Real-time walk-through auralisation of the acoustics of Christ Church cathedral Dublin. *Proceedings of the Institute of Acoustic*, 33:244–258, May 2011.

[107] G. Kearney and J. Levison. Virtual Vs. Actual Multichannel Acoustical Recording. In *Proceedings of the 124th Audio Engineering Society Convention*, Amsterdam, The Netherlands, May 2008.

[108] G. Kearney, C. Masterson, and F. Boland. Acoustic impulse response interpolation using dynamic time warping. In *Proceedings of the Audio Engineering Society 35th International Conference*, London, UK, February 2009.

[109] G. Kearney, C. Masterson, M. Gorzel, H. Rice, and F. Boland. Application of HRIR Factorization to Game Audio. In *Proceedings of the 41st Audio Engineering Society International Conference: Audio for Games*, London, UK, February 2011.

[110] W. D. V. Keet. The influence of early lateral reflections on the spatial impression. In *Proceedings of the 6th International Congrgress on Acoustics, E-2-4*, Tokyo, Japan, 1968.

[111] R. Kessler. An optimised method for capturing multidimensional acoustic fingerprints. In *Proceedings of the 118th Audio Engineering Society Convention*, Barcelona, Spain, 2005.

[112] M. Kleiner, B.-I. Dalenbck, and P. Svensson. Auralization - An Overview. *Journal of the Audio Engineering Society*, 41(11):861–875, 1993.

[113] J. Křivánek, J. Konttinen, S. Pattanaik, K. Bouatouch, and J. Žára. Fast approximation to spherical harmonics rotation. In *Proceedings of the 22nd spring conference on Computer graphics*, SCCG '06, pages 154:1–154:10, New York, NY, USA, 2006. ACM.

[114] C. Kyriakakis and R. Sadek. A Novel Multichannel Panning Method for Standard and Arbitrary Loudspeaker Configurations. In *Proceedings of the 117th Audio Engineering Society Convention*, Oct 2004.

[115] A. Laborie, R. Bruno, and S. Montoya. Reproducing Multichannel Sound on any Speaker Layout. In *Proceedings of the 118th Audio Engineering Society Convention*, May 2005.

[116] M.-V. Laitinen and V. Pulkki. Converting 5.1 audio recordings to B-format for directional audio coding reproduction. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'2011)*, pages 61–64, May 2011.

[117] R. Lee. SHELF FILTERS for Ambisonic Decoders. Technical report, `http://ambisonic.info/info/ricardo/shelfs.html`, May 2008.

[118] R. Lee and A. J. Heller. Ambisonic localisation - part 2. In *Proceeding of the 14th International Congress on Sound and Vibration*, Cairns, Australia, July 2007.

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

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

[121] C. Lennon. 3DTV Overview. In *IEEE Denver Signal Processing Society*, April 2010.

[122] J. Lewald, S. Peters, M. C. Corballis, and M. Hausmann. Perception of stationary and moving sound following unilateral cortectomy. *Neuropsychologia*, 47(4):962 – 971, 2009.

[123] T. Lokki. *Physically-based Auralization: Design, Implementation, and Evaluation.* PhD thesis, Helsinki University of Technology, Espoo, Finland, 2002.

[124] T. Lokki and L. Savioja. Evaluation of Auralization Results. In *Proceedings of the Forum Acusticum*, Budapest, Hungary, Aug/Sep 2005.

[125] T. Lokki, L. Savioja, and J. Hiipakka. A framework for evaluating virtual acoustic environments. In *Proceedings of the 110th Audio Engineering Society Convention*, May 2001.

[126] J. M. Loomis, C. Hebert, and J. G. Cicinelli. Active localization of virtual sounds. *The Journal of the Acoustical Society of America*, 88(4):1757–1764, 1990.

[127] J. M. Loomis, R. L. Klatzky, J. W. Philbeck, and R. G. Golledge. Assessing auditory distance perception using perceptually directed action. *Perception And Psychophysics*, 60(6):966–980, 1998.

[128] D. Luenberger and Ye. *Linear and Nonlinear Programming.* Springer, third edition, 2008.

[129] J. Mackenzie, J. Huopaniemi, V. Valimaki, and I. Kale. Low-order modeling of head-related transfer functions using balanced model truncation. *IEEE Signal Processing Letters*, 4(2):39–41, 1997.

[130] P. Majdak, P. Balazs, and B. Laback. Multiple Exponential Sweep Method for Fast Measurement of Head Related Transfer Functions. In *Proceedings of the 122nd Audio Engineering Society Convention*, Vienna, Austria, May 2007.

[131] Y. Makita. On the directional localization of sound in the stereophonic sound field. *E.B.U. Review, Part A*, 73:102–108, June 1962.

[132] D. Malham. 3-D acoustic space and its simulation using Ambisonics, 2007.

[133] Martin, Geoff and Woszczyk, Wieslaw and Corey, Jason and Quesnel, Ren. Sound Source Localization in a Five-Channel Surround Sound Reproduction System. In *Proceedings of the 107th Audio Engineering Society Convention*, New York, NY, USA, September 1999.

[134] C. Masterson. *Binaural Impulse Response Rendering for Immersive Audio.* PhD Thesis, Trinity College Dublin, Dublin, Ireland, 2011.

[135] C. Masterson, G. Kearney, and F. Boland. Acoustic impulse response interpolation using Dynamic Time Warping. In *Proceedings of the Audio Engineering Society 35th International Conference*, London, UK, 2009.

[136] C. Masterson, G. Kearney, M. Gorzel, and F. Boland. HRIR order reduction using approximate factorization. *IEEE Transactions on Audio, Speech, and Language Processing*, 20:61–71, August 2012.

[137] D. F. Mcallister. Display technology. *In: Hornak, J. 2005. Wiley Encyclopaedia on Imaging Science and Technology. Wiley Interscience: Bognor Regis, West*, pages 1327–1344, 2006.

[138] S. G. McGovern. Fast image method for impulse response calculations of box-shaped rooms. *Applied Acoustics*, 70(1):182 – 189, 2009.

[139] S. G. McGovern. The image source reverberation model in an n-dimensional space. In *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx11)*, Paris, France, September 2011.

[140] A. McKeag and D. McGrath. Sound Field Format to Binaural Decoder with Head-Tracking. In *Proceedings of the AES 6th Australian Regional Convention*, Melbourne, Australia, 1996.

[141] D. Menzies. W-panning and O-format, tools for object spatialization. In *Proceedings of the 8th International Conference on Auditory Display (ICAD2002)*, Kyoto, Japan, 2002. Advanced Telecommunications Research Institute (ATR).

[142] D. Menzies. Nearfield synthesis of complex sources with High Order Ambisonics, and binaural rendering. In *13th International Conference on Auditory Display*, Montreal, Canada, 2007.

[143] S. Merchel, A. F. Franco, L. Pesqueux, M. Rouaud, and M. O. Soerensen. Sound Reproduction By Wave Field Synthesis. Project report, Aalborg University, 2004.

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

[145] D. Mershon and L. King. Intensity and reverberation as factors in the auditory perception of egocentric distance. *Attention, Perception, & Psychophysics*, 18:409–415, 1975. 10.3758/BF03204113.

[146] D. H. Mershon, W. L. Ballenger, A. D. Little, P. L. McMurtry, and J. L. Buchanan. Effects of room reflectance and background noise on perceived auditory distance. *Perception*, 18(3):403–416, 1989.

[147] mhacoustics. The Eigenmike microphone array, accessed October 21, 2012. `http://www.mhacoustics.com/mh_acoustics/Eigenmike_microphone_array.html`.

[148] Microsoft. Kinect, accessed September 9, 2012. `http://www.xbox.com/en-IE/Kinect`.

[149] J. C. Middlebrooks and D. M. Green. Sound Localization by Human Listeners. *Annual Review of Psychology*, 42(1):135–159, 1991.

[150] E. K. Milne. *Christ Church Cathedral, Dublin: A History.* Four Courts Press, 2000.

[151] P. Mokhtari, H. Takemoto, R. Nishimura, and H. Kato. Acoustic simulation of KEMAR's HRTFs: verification with measurements and the effects of modifying head shape and pinna concavity. In *Proceedings of the International Workshop on the Principles and Applications of Spatial Hearing*, Zao, Japan, 2009.

[152] H. Møller. Fundamentals of binaural technology. *Applied Acoustics*, 36(34):171–218, 1992.

[153] D. Moore and J. Wakefield. The Design of Ambisonic Decoders for the ITU 5.1 Layout with Even Performance Characteristics. In *Proceedings of the 124th Audio Engineering Society Convention*, Amsterdam, The Netherlands, May 2008.

[154] J. A. Moorer. About this reverberation business. *Computer Music Journal*, 23(2), 1979.

[155] P. Morse and K. Ingard. *Theoretical Acoustics.* Princeton University Press, 1987.

[156] T. Mullen. *Mastering Blender.* Serious skills. John Wiley & Sons, 2009.

[157] C. Müller-Tomfelde. Low-latency convolution for real-time applications. In *Proceedings of the Audio Engineering Society 16th International Conference: Spatial Sound Reproduction*, Rovaniemi, Finland, 1999.

[158] C. Müller-Tomfelde. Time-varying Filter in Non-uniform Block Convolution. In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX01)*, Limerich, Ireland, 2001.

[159] E. Mullins. *Statistics for the Quality Control Chemistry Laboratory.* The Royal Society of Chemistry, 2003.

[160] M. Nakahara, T. Kishi, K. Kojima, T. Hanyu, and K. Hoshi. Implementation of an Interactive 3-D Reverberator for Video Games Using Statistical Acoustics. In *Proceedings of the 133rd Audio Engineering Society Convention*, October 2012.

[161] NaturalPoint. Trackir 5, accessed June 22, 2012. `http://www.naturalpoint.com/trackir/`.

[162] M. Neukom and J. C. Schacher. Ambisonics Equivalent Panning. In *International Computer Music Conference ICMC'08*, Belfast, UK, 2008.

[163] R. Nicol, M. Emerit, F. Telecom, and C. Lannion. Reproducing 3D-Sound for Videoconferencing: a Comparison between Holophony and Ambisonic. In *Proceedings of the First COST-G6 Workshop on Digital Audio Effects (DAFX98)*, Barcelona, Spain, 1998.

[164] S. Nielsen. Auditory distance perception in different rooms. *Journal of the Audio Engineering Society*, 41(10):755–755, 1993.

[165] M. Noisternig, A. Sontacchi, T. Musil, and R. Höldrich. A 3D Ambisonic Based Binaural Sound Reproduction System. In *Proceedings of the Audio Engineering Society 24th International Conference on Multichannel Audio*, Banff, Alberta, Canada, June 2003.

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

[167] N. Olaiz, P. A. an Toni Mateos, and D. Garcia. 3D-Audio with CLAM and Blender's Game Engine. In *Proceedings of the Linux Audio Conference 2009*, Parma, Italy, April 2009.

[168] M. Otani and T. Hirahara. Auditory Artifacts due to Switching Head-Related Transfer Functions of a Dynamic Virtual Auditory Display. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E91-A(6):1320–1328, June 2008.

[169] E. Ozimek. *Dźwiek i jego percepcja. Aspekty fizyczne i psychoakustyczne.* Wydawnictwo Naukowe PWN, Warszawa-Poznań, 2002.

[170] J. Pätynen. Directivities of orchestra instruments for auralization. In *Proceedings of the EAA Symposium on Auralization*, Espoo, Finland, June 1998.

[171] J. Pätynen and T. Lokki. Directivities of Symphony Orchestra Instruments. *Acta Acustica united with Acustica*, 96(1):138–167, 2010.

[172] R. S. Pellegrini. Comparison of Data- and Model-Based Simulation Algorithms for Auditory Virtual Environments. In *Proceedings of the 106th Audio Engineering Sociery Convention*, Munich, Germany, May 1999.

[173] R. S. Pellegrini. Perception-Based Room Rendering for Auditory Scenes. In *Proceedings of the Audio Engineering Society 109th Convention*, Los Angeles, CA, USA, Sep 2000.

[174] D. R. Perrott, B. Costantino, and J. Ball. Discrimination of moving events which accelerate or decelerate over the listening interval. *The Journal of the Acoustical Society of America*, 93(2):1053–1057, 1993.

[175] Physikalisch-Technische-Bundesanstalt. Directivities of musical instruments, accessed October 21, 2012. `http://www.ptb.de/cms/en/fachabteilungen/abt1/fb-16/ag-1630/room-acoustics/directivities.html`.

[176] A. Pietrzyk. Computer modeling of the sound field in small rooms. In *Audio Engineering Society 15th International Conference: Audio, Acoustics & Small Spaces*, Copenhagen, Denmark, October 1998.

[177] D. Pinchon and P. E. Hoggan. Rotation matrices for real spherical harmonics: general rotations of atomic orbitals in space-fixed axes. *Journal of Physics A: Mathematical and Theoretical*, 40(7):1597, 2007.

[178] M. A. Poletti. A unified theory of horizontal holographic sound systems. *Journal of the Audio Engineering Society*, 48(12):1155–1182, 2000.

[179] R. Preibisch-Effenberger. *Die Schallokalisationsfahigkeit des Menschen und ihre audiometrische Verwendung zur klinischen Diagnostik [The human faculty of sound localization and its audiometric application to clinical diagnostics].* PhD thesis, Technische Universität Dresden, 1966.

[180] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing, Second Edition.* Cambridge University Press, 2 edition, October 1992.

[181] M. Puckette. Pure Data, accessed June 22, 2012. `http://puredata.info/`.

[182] V. Pulkki. Virtual sound source positioning using Vector Base Amplitude Panning. *Journal of the Audio Engineering Society*, 45(6):456–466, 1997.

[183] V. Pulkki. Generic panning tools for MAX/MSP. In *Proceedings of International Computer Music Conference*, pages 304–307, Berlin, Germany, 2000.

[184] V. Pulkki. *Spatial Sound Generation and Perception by Amplitude Panning Techniques.* PhD thesis, Helsinki University of Technology, 2001.

[185] V. Pulkki. Spatial Sound Reproduction with Directional Audio Coding. *Journal of the Audio Engineering Society*, 55(6):503–516, 2007.

[186] V. Pulkki, M.-V. Laitinen, and V. P. Sivonen. HRTF measurements with a continuously moving loudspeaker and swept sines. In *Proceedings of the 128th Audio Engineering Society Convention*, London, UK, 2010.

[187] Python Software Foundation. Python programming language, accessed October 21, 2012. `http://www.python.org`.

[188] N. Raghuvanshi and M. Lin. Physically Based Sound Synthesis for Large-Scale Virtual Environments. *IEEE Computer Graphics and Applications*, 27(1):14–18, 2007.

[189] B. Rakerd and W. M. Hartmann. Localization of sound in rooms, II: The effects of a single reflecting surface. *The Journal of the Acoustical Society of America*, 78(2):524–533, August 1985.

[190] B. Rakerd and W. M. Hartmann. Localization of sound in rooms, III: Onset and duration effects. *The Journal of the Acoustical Society of America*, 80(6):1695–1706, December 1986.

[191] L. Rayleigh. On our perception of sound direction. *Philosophical Magazine*, 13:232, 1907.

[192] A. Reilly and D. McGrath. Convolution Processing for Realistic Reverberation. In *Proceedings of the 98th Audio Engineering Society Convention*, Paris, France, Feb 1995.

[193] J. H. Rindel. Evaluation of room acoustic qualities and defects by use of auralization. *The Journal of the Acoustical Society of America*, 116(4):2483–2483, 2004.

[194] J. H. Rindel, C. Lynge, G. Naylor, and K. Rishoj. The use of a digital audio mainframe for room acoustical auralization. In *Proceedings of the 96th Audio Engineering Society Convention*, Amsterdam, The Netherlands, February 1994.

[195] F. Rumsey. Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. *Journal of the Audio Engineering Society*, 50(9):651–666, 2002.

[196] M. Rychtarikova, T. V. den Bogaert, G. Vermeir, and J. Wouters. Binaural Sound Source Localization in Real and Virtual Rooms. *Journal of the Audio Engineering Society*, 57(4):205–220, 2009.

[197] W. C. Sabine. Reverberation. In *Collected papers on acoustics*. Cambridge : Harvard University Press, 1922.

[198] L. Savioja. Real-time 3d finite-difference time-domain simulation of low- and mid-frequency room acoustics. In *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*, Graz, Austria, September 2010.

[199] L. Savioja, J. Huopaniemi, T. Lokki, and R. Väänänen. Virtual environment simulation - advances in the DIVA project. In *Proceedings of the International Conference on Auditory Display (ICAD'97)*, pages 43–46, Palo Alto, CA, 1997.

[200] L. Savioja, T. J. Rinne, and T. Takala. Simulation of Room Acoustics with a 3-D finite difference mesh. In *Proceedings of the International Computer Music Confonference*, pages 463–466, Aarhus, Denmark, September 1994.

[201] L. Savioja, V. Valimaki, and J. O. Smith. Audio Signal Processing Using Graphics Processing Units. *Journal of the Audio Engineering Society*, 59(1/2):3–19, 2011.

[202] M. R. Schroeder. Digital Simulation of Sound Transmission in Reverberant Spaces. *The Journal of the Acoustical Society of America*, 47(2A):424–431, 1970.

[203] A. Silzle, H. Strauss, and P. Novo. IKA-SIM: A System to Generate Auditory Virtual Environments. In *Proceedings of the 116th Audio Engineering Society Convention*, Berlin, Germany, May 2004.

[204] G. Simonsen. Master's thesis, Technical University of Denmark, Lyngby, Denmark, 1984.

[205] P. Sloan. Stupid spherical harmonics (sh) tricks. In *Proceedings of the Game Developers Conference*, pages 320–321, San Francisco, CA, USA, 2008.

[206] A. Smolic, P. Kauff, S. Knorr, A. Hornung, M. Kunter, M. Muller, and M. Lang. Three-dimensional video postproduction and processing. *Proceedings of the IEEE*, 99(4):607–625, 2011.

[207] F. Smyth and D. Lennon. Modelling and retro-modelling: Lidar and point cloud as a tool in acoustic analysis. Working paper series, Urban Institute Ireland, September 2009.

[208] F. Smyth, D. Lennon, J. Ollie, G. Kearney, F. Boland, and H. Rice. A Comparative Study of the Acoustics of Christ Church and St. Patrick's Cathedrals. Research Report 16857, The Irish Heritage Council, Kilkenny, Ireland, 2009.

[209] Soundfield Ltd. Soundfield tetrahedral microphone capsule, accessed September 24, 2012. `http://www.soundfield.com/soundfield/soundfield.php`.

[210] A. Southern, J. Wells, and D. Murphy. Rendering walk-through auralisations using wave-based acoustical models. In *Proceedings of the 17th European Signal Processing Conference (EUSIPCO 2009)*, Glasgow, Scotland, August 2009.

[211] A. R. Sovijärvi and J. Hyvärinen. Auditory cortical neurons in the cat sensitive to the direction of sound source movement. *Brain Research*, 73(3):455 – 471, 1974.

[212] G. Spikofski and M. Fruhmann. Optimization of Binaural Room Scanning (BRS): Considering inter-individual HRTF-characteristics. In *Proceedings of the 19th Audio Engineering Society Conference International Conference: Surround Sound - Techniques, Technology, and Perception*, pages 124–134, Jun 2001.

[213] R. Stalley. *Christ Church Cathedral, Dublin: A History*, chapter George Edmund Street and the Restoration of the Cathedral, 1968 - 78. Four Courts Press, September 2000.

[214] R. Stewart and D. Murphy. A Hybrid Artificial Reverberation Algorithm. In *Proceedings of the 122nd Audio Engineering Society Convention*, Vienna, Austria, May 2007.

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

[216] G. E. Street and E. Seymour. The Cathedral of the Holy Trinity commonly called Christ Church Cathedral Dublin: an account of the restoration of the fabric. *Sharpe and Co.*, 1882.

[217] Y. Sugita, Y. Suzuki, and Others. Implicit estimation of sound-arrival time. *Nature*, 421:911, February 2003.

[218] R. Szeliski. *Computer vision: Algorithms and applications*. Springer, Berlin, Germany, 2010.

[219] K. Tanno, A. Saji, S. Ito, J. Huang, and W. Hatano. A Precise Sound Image Panning Method for Side Areas Using 5.1 Channel Audio Systems. In *Proceedings of the 35th Audio Engineering Society International Conference: Audio for Games*, Feb 2009.

[220] G. Theile. Multichannel Natural Music Recording Based on Psychoacoustic Principles (extended version). In *Proceedings of the Audio Engineering Society 19th International Conference*, Schloss Elmau, Germany, June 2001.

[221] J. Treviño, T. Okamoto, Y. Iwaya, and Y. Suzuki. Evaluation of a new Ambisonic decoder for irregular loudspeaker arrays using interaural cues. In *Proceedings of the 3rd International Symposium on Ambisonics & Spherical Acoustics*, Lexington, Kentucky, USA, June 2011.

[222] N. Tsingos. Pre-computing geometry-based reverberation effects for games. In *Proceedings of the 35th Audio Engineering Society International Conference on Game Audio*, London, UK, February 2009.

[223] J. Vilkamo. *Spatial Sound Reproduction with Frequency Band Processing of B-format Audio Signals.* PhD thesis, Helsinki University of Technology, Espoo, Finland, 2008.

[224] H. von Helmholtz and J. Southall. *Helmholtz's treatise on physiological optics.* Helmholtz's Treatise on Physiological Optics. The Optical Society of America, 1924.

[225] J. Vroomen. Perceptual effects of cross-modal stimulation: Ventriloquism and the freezing phenomenon. In *Handbook of multisensory processes.* MIT Press, 2004.

[226] D. Waller. Factors affecting the perception of interobject distances in virtual environments. *Presence: Teleoperators and Virtual Environments*, 8(6):657–670, December 1999.

[227] E. Wenzel, M. Arruda, D. Kistler, and F. Wightman. Localization using nonindividualized head-related transfer functions. *Journal of the Acoustical Society of America*, 1993.

[228] E. Wenzel, P. Stone, S. Fisher, and S. Foster. A system for three-dimensional acoustic 'visualization' in a virtual environment workstation. In *Proceedings of the First IEEE Conference on Visualization*, pages 329–337, 1990.

[229] E. M. Wenzel. What Perception Implies About Implementation of Interactive Virtual Acoustic Environments. In *Proceedings of the 101st Audio Engineering Society Convention*, November 1996.

[230] E. M. Wenzel, F. L. Wightman, and S. H. Foster. Development of a three-dimensional auditory display system. *ACM SIGCHI Bulletin*, 20(2):52–57, Oct. 1988.

[231] S. Werner, F. Klein, and A. Siegel. On the influence of visual feedback on vertical sound source localization. In *Proceedings of the 1st International Conference on Spatial Audio*, Detmold, Germany, 2011.

[232] J. R. West. *Five-Channel Panning Laws: An Analytical and Experimental Comparison.* M.Sc. Thesis, University of Miami, Miami, Florida, USA, 1998.

[233] H. Wierstorf, M. Geier, A. Raake, and S. Spors. A Free Database of Head-Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances. `www.audio.qu.tu-berlin.de`, 2011. accessed July 23, 2012.

[234] B. Wiggins. *An Investigation into the Real-Time Manipulation and Control of Three-Dimensional Sound Fields.* Ph.D. Thesis, University of Derby, Derby, UK, 2004.

[235] B. Wiggins and T. Spenceley. Distance coding and performance of the mark 5 and st350 SoundField microphones and their suitability for Ambisonic reproduction. In *Proceeding of the Institute of Acoustics*, volume 31, 2009.

[236] F. L. Wightman and D. J. Kistler. Headphone simulation of free-field listening. II: Psychophysical validation. *The Journal of the Acoustical Society of America*, 85(2):868–878, 1989.

[237] H. Wittek. *Perceptual Differences between Wavefield Synthesis and Stereophony.* Ph.D. Thesis, University of Surrey, Surrey, UK, 2007.

[238] P. Zahorik. Estimating Sound Source Distance with and without Vision. *Optometry & Vision Science*, 78(5):270–275, May 2001.

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

[240] P. Zahorik, D. S. Brungart, and A. W. Bronkhorst. Auditory distance perception in humans: A summary of past and present research. *Acta Acustica united with Acustica*, 91(3):409–420, 2005.

[241] C. Zhang and B. Xie. Platform for dynamic virtual auditory environment real-time rendering system. *Chinese Science Bulletin*, 58(3):316, 2013.

[242] C. J. Ziemer, J. M. Plumert, J. F. Cremer, and J. K. Kearney. Estimating distance in real and virtual environments: Does order make a difference? *Atten Percept Psychophys*, 71(5):1095–1106, July 2009.

[243] C. L. Zitnick, M. F. Cohen, S. B. Kang, B. Ressler, and A. Colburn. A viewer-centric editor for 3D movies. *Computer Graphics*, 31(1):20–35, 2011.

[244] D. N. Zotkin, R. Duraiswami, E. Grassi, and N. A. Gumerov. Fast head-related transfer function measurement via reciprocity. *The Journal of the Acoustical Society of America*, 120(4):2202–2215, 2006.

[245] F. Zotter. Rotation of spherical harmonics in R^3, accessed October 6, 2012. `http://www.ambisonics.iem.at/xchange/format/docs/spherical-harmonics-rotation`.

[246] F. Zotter and M. Frank. All-Round Ambisonic Panning and Decoding. *Journal of the Audio Engineering Socciety*, 60(10):807–820, 2012.