University of Southampton

University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Author (Year of Submission) "Full thesis title", University of Southampton, name of the University Faculty or School or Department, PhD Thesis, pagination.

Data: Author (Year) Title. URI [dataset]

UNIVERSITY OF Southampton

DOCTOR OF PHILOSOPHY

Higher Order Stereophony

Author:
Jacob HOLLEBON
ORCID ID:
0000-0002-4119-4070

Supervisor:
Prof. Filippo Maria FAZI

A thesis submitted in fulfillment of the requirements for the Doctor of Philosophy

in the

Institute of Sound and Vibration Research
Faculty of Engineering and Physical Sciences

March 15, 2023

Declaration of Authorship

I, Jacob HOLLEBON, declare that this report titled, "Higher Order Stereophony" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this report has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this report is entirely my own work.
- I have acknowledged all main sources of help.
- Where the report is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
- Parts of this work have been published as:
  - Hollebon, J. and Fazi, F. M. Generalised low frequency 3D audio reproduction over loudspeakers. In AES 148th Convention, 2020.
  - Hollebon, J. and Fazi, F. M. Efficient HRTF representation using compact mode HRTFs. In AES 149th Convention, 2020.
  - Hollebon, J. and Fazi, F. M. Experimental study of various methods for low frequency spatial audio reproduction over loudspeakers. In I3DA: International Conference on Immersive and 3D Audio, 2021.
  - Fazi, F. M. and Hollebon, J. The ring of silence in ambisonics and binaural audio reproduction. In J. Audio Eng. Soc (submitted, under review).

Signed:

Date:

UNIVERSITY OF SOUTHAMPTON

Abstract

Faculty of Engineering and Physical Sciences

Institute of Sound and Vibration Research

Doctor of Philosophy

Higher Order Stereophony

by Jacob HOLLEBON

This thesis presents the derivation, experimental validation and perceptual evaluation of a new technique for spatial audio reproduction named Higher Order Stereophony. The approach uses the Taylor expansion of a plane wave soundfield to represent the field as a summation of its derivatives. The soundfield is then reproduced correctly along a single axis only, which is assumed to be the listener's interaural axis, in an attempt to reproduce the correct binaural signals for 3D audio reproduction. A dynamic extension to the technique using head-tracking dynamically adapts the loudspeaker gains to ensure the reproduction axis always aligns with the listener's interaural axis, regardless of the listener's orientation.

Higher Order Stereophony is shown to be a generalisation to higher orders of classic stereo techniques such as the sine law, and is a frequency-independent solution resulting in loudspeaker panning gains. This generalisation is similar in manner to Higher Order Ambisonics, and parallels between the two are investigated including the derivation of decoders to transform from both the 3D or 2D Higher Order Ambisonics representation to Higher Order Stereophony, which first require a specific rotation of the soundfield followed by a matrix of gains. The new approach presents an alternative to Higher Order Ambisonics, with advantages in requiring only $(N+1)$ channels/loudspeakers for N-th order reproduction as well as requiring loudspeakers in front of the listener only.

Higher Order Stereophony is also applied to binaural rendering and shown to fully reproduce binaural signals with a rigid sphere Head-Related Transfer Function, due to the axisymmetric geometry of the scattering in the head model. When representing any Head-Related Transfer Function using spherical harmonics, a specific Higher Order Stereophony rotation is defined to align the interaural axis with the z axis, which is shown to reorder the energy of the spherical harmonic coefficients to those with m close to 0. Higher Order Stereophony is then demonstrated to reproduce all spherical harmonic coefficients of the Head-Related Transfer Function with $m = 0$ only. Subjective experiments comparing Higher Order Ambisonics and Higher Order Stereophony show that Higher Order Stereophony can perform similarly to Higher Order Ambisonics despite its advantages computationally and regarding loudspeaker positioning.

Dedicated to my Grandmother, Rita Hollebon. Your music lives on in us.

Acknowledgements

I would first like to deeply thank my supervisor, Prof. Filippo Maria Fazi. Through-out my doctoral training he has provided me with excellent guidance and teaching, thoughtful mentoring and unwavering support both academically and personally. He has gifted me countless hours of his time and I will always fondly recall our (often extended and overrunning...) meetings lost in mathematics! Without his inspiration this work would not exist, and I will be ever thankful to him.

I would also like to thank my colleagues in the VAAE team and at Audioscenic, for all of their advice and friendships during this time. Thank you to Vlad Paul, Dr. Falk-Martin Hoffmann, Dr. Eric Hamdan, Dr. Andreas Franck and Wilfried Gallian. A particularly special thank you to Dr. Marcos Simón Gálvez, who played a central role in my journey to undertaking a PhD, and who through-out has remained a good friend.

I would like to acknowledge Dr. Jordan Cheer, who has been an invaluable mentor and guiding voice. Many thanks also to all those at the ISVR, particularly those in the SPAH group.

On a more a personal note, my utmost gratitude goes to my wife, Sarah. Her love and support has been steadfast, her patience long, and her compassion deep. I could not have done these years without you. Thank you, my love.

Thank you to my family. To my father and mother Graham and Karen for always supporting my never-ending academic pursuits. To my brother, sister-in-law and nephew Ben, Pollyanna and Rupert, for always grounding me and bringing light relief! And to my Grandparents Rita and Cliff, both whom are the root of my love of music and audio. A wider thank you to my friends in Southampton and Sussex. Particularly to Mike, Ben, Perry, Toby and Joel. All my friends and family been extremely supportive through-out my extended time of studying, and I know all will be pleased I will now finally get a 'proper job'!

The final year of my doctoral training was full of challenges, with the loss of my Grandmother and the personal health of loved ones. Everyone in these acknowledgements were a constant source of strength and support to me during this time, thank you all.

Finally, and most importantly, I thank and praise God for his grace and faithfulness. This thesis, as with my whole life, serves to glorify him.

Contents

List of Abbreviations

ALLRAP All-Round Ambisonic Panning
ALLRAD All-Round Ambisonic Decoding
ACN Ambisonic Channel Numbering
ANOVA ANalysis Of VAriance
CAP Compensated Amplitude Panning
CTC Crosstalk Cancellation
DAW Digital Audio Workstation
DBAP Distance Based Amplitude Panning
GPWD Generalised Plane Wave Decomposition
HOA Higher Order Ambisonics
HOS Higher Order Stereophony
HRIR Head-Related Impulse Response
HRTF Head-Related Transfer Function
HWF Hergoltz Wave Function
ILD Interaural Level Difference
IPD Interaural Phase Difference
ISVR Institute of Sound and Vibration Research
ITD Interaural Time Difference
MADI Multichannel Audio Digital Interface
MDAP Multiple Direction Amplitude Panning
MUSHRA MUltiple Stimulus with Hidden Reference and Anchor
PWD Plane Wave Decomposition
RMS Root Mean Square
$\mathcal{SHT}^{-1}$ Inverse Spherical Harmonic Transform
$\mathcal{SHT}$ Spherical Harmonic Transform
SVD Singular Value Decomposition
UDP User Datagram Protocol
VBAP Vector Base Amplitude Panning
VBIP Vector Base Intensity Panning
VISR Versatile Interactive Scene Renderer

List of Symbols

Mathematical Operators and Symbols

j	Imaginary unit ($j = \sqrt{-1}$)	
$(\cdot)^*$	Complex conjugation	
$(\cdot)^\dagger$	Moore-Penrose Pseudoinverse	
$\lvert\cdot\rvert$	Absolute value	
$\lvert\lvert\cdot\rvert\rvert_2$	L^2 norm	
(a, b)	Open interval from a to b, excluding the endpoints	
$[a, b]$	Closed interval from a to b, including the endpoints	
$[a, b)$	Half-open interval from a to b, including a but excluding b	
$:=$	Definition	
$\in$	Is a member of	
$\implies$	Implies	
$\mathrm{Re}\{\cdot\}$	Operator taking the real part of a function	
$\mathrm{Im}\{\cdot\}$	Operator taking the imaginary part of a function	
$\mathcal{SHT}\{\cdot\}$	Direct Spherical Harmonic Transform	
$\mathcal{SHT}^{-1}\{\cdot\}$	Inverse Spherical Harmonic Transform	

Vectors, Matrices and Sets

$\mathbf{r}, \mathbf{x}, \mathbf{y}, \mathbf{z}$	Position vector defining a point in space	
$\hat{\mathbf{r}}, \hat{\mathbf{x}}, \hat{\mathbf{y}}, \hat{\mathbf{z}}$	Unitary vectors defining a point in space with unit norm	
$\hat{\mathbf{n}}$	Unitary vector pointing from the centre of the listener's head to the left ear, defining the interaural axis	
θ	Inclination angle (3D) or azimuth angle (2D)	rad
ϕ	Azimuth angle (3D)	rad
Ω	Unitary sphere	
$\mathcal{S}^2$	3D unitary sphere (sphere with unitary radius)	
$\mathcal{S}^1$	2D unitary sphere (circle with unitary radius)	
$\mathbb{N}$	The set of natural numbers excluding 0	
$\mathbb{N}_0$	The set of natural numbers including 0	
$\mathbb{Z}$	The set of integer numbers	

$\mathbf{p}_T$	Vector of target pressures
$\mathbf{p}_R$	Vector of reproduced pressures
$\mathbf{g}$	Vector of loudspeaker gains
Ψ	Plant matrix
$\mathbf{Y}$	Matrix of sampled spherical harmonics
$\mathbf{I}$	Identity matrix

Physical Quantities, Constants and Notation

a	Head radius	m
c	Speed of sound propagation	m s^{-1}
f	Frequency	s^{-1}
ω	Angular frequency	rad s^{-1}
k	Wavenumber	rad m^{-1}
$\mathbf{k}$	Wavevector in direction of arrival ($\mathbf{k} = k\hat{\mathbf{k}}$)	rad m^{-1}
g_ℓ	Gain of the ℓ-th loudspeaker	
t	Time	s
$p(\mathbf{x})$	Pressure field	kg m s^{-2}
ϵ	Error function	
$(\cdot)_{l,r,c}$	Subscript corresponding to the left ear, right ear or listener's head centre respectively	
$(\cdot)_i$	Subscript corresponding to an incident soundfield	
$(\cdot)_s$	Subscript corresponding to a scattered soundfield	
$(\cdot)_{rot}$	Subscript corresponding to a listener head rotation	
$(\cdot)_{tot}$	Subscript corresponding to a total soundfield	
$(\cdot)_{trunc}$	Subscript corresponding to an order truncated soundfield	
$(\cdot)_T$	Subscript corresponding to a target soundfield	
$(\cdot)_R$	Subscript corresponding to a reproduced soundfield	

Special Functions

δ_{pq}	Kronecker delta function
$h_n^{(1)}(\cdot)$	n-th order spherical Hankel function of the first kind
$h_n^{(2)}(\cdot)$	n-th order spherical Hankel function of the second kind
$h_n^{(2)\prime}(\cdot)$	First derivative of the n-th order spherical Hankel function of the second kind
$j_n(\cdot)$	n-th order spherical Bessel function of the first kind
$j_n'(\cdot)$	First derivative of the n-th order spherical Bessel function of the first kind
$y_n(\cdot)$	n-th order spherical Bessel function of the second kind
$J_n(\cdot)$	n-th order Bessel function of the first kind

$Y_n^m(\cdot)$	Complex spherical harmonic of order n and degree m
$Y_{n,m}(\cdot)$	Real spherical harmonic of order n and degree m
$P_n^m(\cdot)$	Associated Legendre polynomial of order n and degree m
$P_n(\cdot)$	Legendre polynomial of order n
$T_n(\cdot)$	n-th order Chebyshev polynomial of the first kind
$U_n(\cdot)$	n-th order Chebyshev polynomial of the second kind
$R_n(\cdot)$	n-th order rigid sphere radial filter
$H(\cdot)$	Plane wave kernel
$q(\cdot)$	Plane wave density

Chapter 1

Introduction

1.1 Motivations

Ever since the invention of stereophonic sound, spatial audio has been an essential aspect in the experience of audio, in particular through music and entertainment media. Over the years playback systems became more complex, moving from a stereo loudspeaker pair, to surround sound (5.1, 7.1, 5.1.2 to identify the most common[1]), to spherical loudspeaker arrays of increasing size and eventually, just a pair of headphones equipped with a head-tracking device.

At the heart of all of these systems has been the research question, 'how can we reproduce the illusion of an acoustic scene?'. The research question is in the art of deceiving, how must we convince the listener what they hear is real, not an illusion? Uniquely for a scientific field, the outputs of said research are often quickly turned into consumer products that might be enjoyed by the masses. At the time of writing this work, spatial, 3D, or immersive audio is quickly becoming commonplace in the everyday through the wide use of head-tracked binaural in smart devices and Virtual and Augmented Reality, as well as the increased use of audio soundbars that employ beamforming.

As the complexity of 3D audio systems increases, requiring more loudspeakers and more complex rendering methods (arguably becoming more inaccessible to the average consumer), it is important to pose the question 'what is exactly required to reproduce the desired 3D audio experience'? If a smaller, less complex system can be used, this is a big advantage not just in rendering and physical costs but also for the system to be practically adopted by consumers.

The existence of a large number of differing approaches for spatial audio reproduction gives weight to the complexity of the problem. Naturally, these techniques all begin from different starting assumptions and often employ different mathematical approaches to solve the problem. It is interesting to consider however, how similar

[1] The existence of the rectangular four loudspeaker quadraphonic system might also be noted, however never became commonplace.

some of these techniques are to each other. A key contribution of this work is to draw together such similarities and parallels between existing spatial audio techniques.

One known link is between the classical stereophonic sine law and first order Ambisonics. At low frequencies the stereo approach may be described as capturing a soundfield using the combination of an omnidirectional and a figure of eight microphone arranged along one Cartesian axis. First order Ambisonics is the 3D extension of this approach, utilising an omnidirectional microphone and three figure of eights aligned along each of the three Cartesian axes. This is well established in the literature, for example in [1]. However, whilst a higher order extension to Ambisonics exists, no-one has ever considered if a higher order extension to the stereophonic sine law is also possible. It is from this question that the work in this thesis is derived.

1.2 Contribution

The major contribution of this thesis is the presentation of a new novel spatial audio reproduction technique named Higher Order Stereophony (HOS). The approach is the generalisation of classical stereo to a higher order, allowing for the use of an arbitrary number of loudspeakers and reproduction to an increasingly higher frequency limit depending on the truncation order of the system. The technique is presented as a full theory encapsulating the encoding of a virtual source source through to the derivation of loudspeaker gains for the reproduction problem.

A second contribution is the extension of the HOS approach to binaural reproduction. The technique is shown to fully reproduce the rigid sphere Head-Related Transfer Function (HRTF) due to its axisymmetric nature. It is also shown to be applicable to rendering more generalised HRTFs in the region where the HRTF is approximately axisymmetric, which breaks down at higher frequencies.

HOS is derived using the Taylor expansion of a plane wave soundfield. The Taylor expansion is a rarely used soundfield descriptor in spatial audio, and to the author's knowledge this thesis is the first work in which it has been used as a full basis expansion for soundfield reproduction. Uniquely, the Taylor expansion is used here to accurately reproduce the soundfield along one axis only. Thus whilst soundfield reproduction in the literature often attempts accurate reproduction over a region such as a circle or sphere, here accurate reproduction is across a single line. The fundamental assumption of the technique is that this reproduction line is aligned with the interaural axis, and that reproducing the correct soundfield along this line leads to the correct binaural signals at the listener's ears. This novel approach leads to a dramatic reduction in the number of loudspeakers needed and the requirements on their positioning, in comparison to other state of the art spatial audio techniques. HOS is demonstrated to be an amplitude panning technique therefore real-time implementation is through simple frequency independent panning laws. The theory is

proven mathematically, demonstrated experimentally and also assessed perceptually through two listening tests comparing the technique to Higher Order Ambisonics (HOA).

A core assumption of the theory is that the listener's interaural axis aligns with the reproduction axis, to ensure the correct binaural signals are reproduced for the listener. To this end an extension to the original technique, titled Dynamic HOS, is introduced. Here the theory is extended to adapt for listener head movements based on a head-tracking implementation, to ensure that regardless of the listener's orientation the reproduction axis will align with the interaural axis. Dynamic HOS is therefore a dynamic panning gain approach reproducing the soundfield accurately across a dynamically adjusted reproduction line.

HOS is shown to bare many similarities to HOA, and thorough relations between the two techniques in both 2D and 3D are derived analytically and also demonstrated through experimental data. The ability to transform from the HOA representation to HOS means that all existing HOA material may be reproduced in a more efficient manner using less loudspeakers and with less stringent requirements on their positioning. The links between HOA and HOS are derived using a novel approach for expressing transformations between different soundfield representations using basis expansions.

A final minor contribution of this work is the analysis of a number of core spatial audio reproduction techniques at low frequencies. Under certain assumptions the stereo sine and stereo tangent law, the head-tracked stereo sine law and Crosstalk Cancellation (CTC) are all shown to be subsets of first order Ambisonics. This analysis is referenced through-out the thesis where appropriate and when each of these techniques links to HOS. A more thorough analysis of the generalised links between all of these techniques is described in two convention papers which are reported in Appendix A and B. It is this preliminary work which led to the discovery of HOS.

1.3 Thesis Overview

The thesis is structured as follows. Chapter 2 reviews the most popular and standard approaches for reproduction of spatial audio both over loudspeaker arrays and binaurally. Special attention is paid to any underlying assumptions the approaches make about the soundfield, virtual sound source or reproduction array. A review of spatial hearing is also presented. As the concept of HOA mode matching is key to the thesis, Chapter 3 is dedicated to reviewing the derivation of the classical mode matching reproduction equations using both the 2D and 3D HOA approach. The mode matching concept and the steps of the derivation will be analogous to the HOS derivation presented later. The chapter also covers a range of mathematical expressions that are key for representing a soundfield in 3D and using the spherical harmonics, which will be extensively used later in the thesis.

Chapters 4, 5 and 6 are each intended as the foundations of three self-contained journal papers, looking at different aspects of the new technique named HOS. Chapter 4 contains the initial derivation and verification of HOS. First the Taylor expansion is used to represent a plane wave soundfield as an expansion of the soundfield's derivatives. For a plane wave, these derivatives are shown to be trigonometric terms to the power of n for the n-th order. Truncation of the representation may then be made to a finite order N which results in only $(N+1)$ terms required for the expansion. This unique approach using the Taylor expansion results in an accurate representation of the soundfield across a single line only. It is assumed that this reproduction line is the interaural axis, so that both listener's ears lie on the line. Next, a unique reproduction approach through matching the reproduced soundfields derivatives to the target virtual field derivatives is presented. This order matching approach, analogous to mode matching, is used to define HOS loudspeaker signals for any given loudspeaker arrangement. Examples of HOS are then considered, and the classic stereo sine law is shown to be a first order HOS system, hence the new technique is the generalisation of stereo to higher orders. The concept of transforming between different soundfield representations is then presented, and decoders for use in transforming HOA material to HOS are then derived. Finally, experimental measurements using a linear and a spherical microphone array are used to confirm details about the HOS technique mathematically proven earlier.

In Chapter 5 an extension to the new technique titled Dynamic HOS is presented. Here the use of a head-tracker is included in the approach thus allowing the listener to rotate freely. The Dynamic HOS approach then uses dynamic panning gains to ensure reproduction is correct across the listener's interaural axis regardless of their head orientation. In a similar manner to Chapter 4, the mathematics of the approach is presented, followed by example systems including the demonstration of the stereo tangent law and head-tracked sine law which are special cases of first order Dynamic HOS. Also, the decoders linking HOA to HOS are extended to the dynamic case. Finally, a subjective listening test is presented aimed at comparing the ability of a number of HOA and HOS loudspeaker systems to reproduce a convincing virtual sound source to a listener.

Chapter 6 considers the application of HOS to binaural rendering. First, the Plane Wave Decomposition (PWD) is used to consider standard HOA rendering of generalised soundfields in both free field and with the inclusion of a real Head-Related Transfer Function (HRTF). Following this, a similar mathematical approach is used to consider the effects of including a real HRTF when utilising HOS. Simulations are presented to directly compare the ability of HOA and HOS to reproduce a given HRTF and a second subjective test comparing binaural rendering of a number of implementations of HOA and HOS is presented.

Finally, Chapter 7 presents an overview of the entire work and makes key conclusions. Possible directions for future work are discussed and finally the thesis is summarised in concluding remarks.

Chapter 2

Literature Review

This chapter will cover a literature review of the core background concepts necessary for the thesis and spatial audio in general. This review will include the state of the art in a wide range of techniques for spatial audio reproduction primarily using loudspeaker arrays. Where appropriate the definition of the loudspeaker signals for these techniques will be stated as well as analysis of the key underlying assumptions and targets of each approach. Next, standard binaural techniques are also covered. Finally a brief review of human spatial hearing is presented.

2.1 Spatial Audio Reproduction

The goal of a spatial audio system is to reproduce the acoustic illusion of virtual acoustic scenes to a listener. In its most basic case, this takes the form of a virtual sound source positioned somewhere in 3D space about the listener. The problem is not limited to such basic soundfields, for example, reverberation or diffuse-like soundfields might also be a rendering requirement. In the following sections, existing spatial audio reproduction techniques will be revised and the mathematical laws describing the approaches defined. The wide range of different spatial audio approaches that have been developed over the last 60 years holds testament to the complexity of the problem of creating realistic acoustic illusions. Further complications arise when considering what must be reproduced to recreate said acoustic illusion. For example, many approaches aim to physically recreate the soundfield exactly, however the final problem is a perceptual one - it is not just a case of what is physically measured at the listener's ears but also what they perceptually understand.

To tackle the problem these techniques often make different assumptions about the nature of the reproduction system, virtual sound source (or generalised soundfield), the listener themselves (for example HRTF models, listener orientation) or whether 3D or 2D space is considered. In [2] the distinction between three different categories of spatial audio reproduction methods is made;

1. **Soundfield Reconstruction:** Reproducing physical properties of the soundfield over a region of space.
2. **Panning Techniques:** Panpot laws where knowledge of the virtual source position relevant to the reproduction loudspeakers is used to define the loudspeaker gains. The panpot laws may be derived using physical or perceptual means (for example based on listening test observations).
3. **Binaural Techniques:** Reproduction of the pressure at the listener's ears directly through headphones or CTC loudspeaker systems.

Often these assumptions are overlooked in the literature and will be clearly stated where appropriate. One secondary yet novel contribution of this thesis is the consolidation of a number of these techniques at low frequencies, where they are mathematically shown to be equal or subsets of each other. This result comes about even though the techniques are derived using very different starting assumptions and fall into different sets of the categories listed above. These results are most concisely presented in the following two conference papers:

- Jacob Hollebon and Filippo Maria Fazi. "Generalised Low Frequency 3D Audio Reproduction Over Loudspeakers". AES 148th Convention. 2020.
- Jacob Hollebon and Filippo Maria Fazi. "Experimental Study of Various Methods for Low Frequency Spatial Audio Reproduction Over Loudspeakers". I3DA: International Conference on Immersive and 3D Audio. 2021.

Both of these papers are provided in Appendix *A* and B. However, all the results will appear naturally through-out the thesis, albeit separated in different chapters.

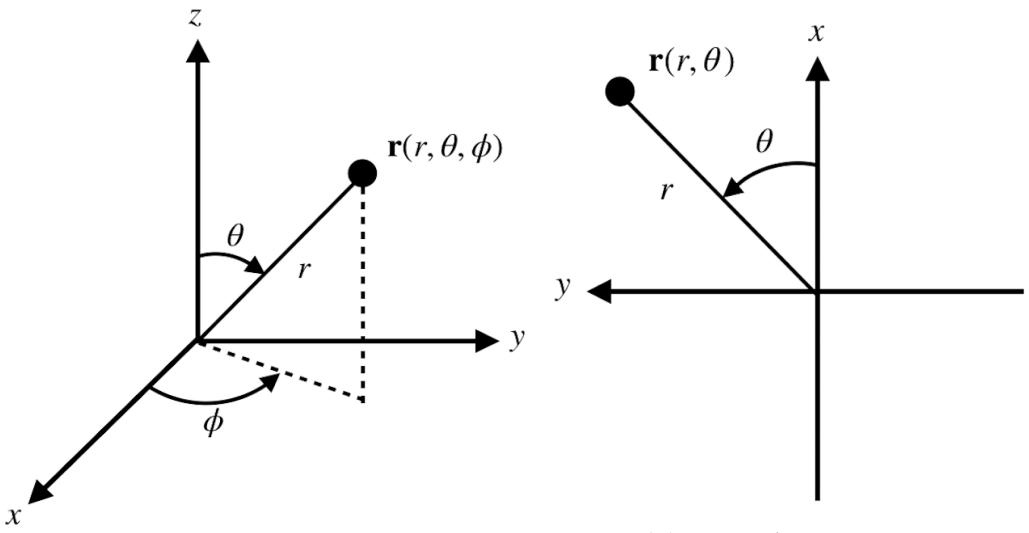
2.2 Mathematical Preliminaries

Many acoustical problems are most naturally represented in a spherical coordinate system, as opposed to a Cartesian coordinate system. This work will utilise both a 3D and a 2D coordinate system, as shown in Fig. 2.1. In 3D a point, $\mathbf{r}$, may be represented by a radius, $r = [0, \infty)$, a colatitude angle $\theta = [0, \pi]$ as measured from the positive z axis [1] and an azimuth angle $\phi = [0, 2\pi)$. The relations between spherical and Cartesian coordinates in 3D is given by

$$\mathbf{r} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = r \begin{pmatrix} \sin(\theta)\cos(\phi) \\ \sin(\theta)\sin(\phi) \\ \cos(\theta) \end{pmatrix}. \tag{2.1}$$

In 2D, $\mathbf{r}$ may be represented using only a radius, $r = [0, \infty)$ and an azimuth angle $\theta = [0, 2\pi)$. To maintain consistency with the literature and mathematical expressions,

[1] Often in spatial audio literature the elevation angle is alternatively used, measured from the $x - y$ plane and ranging between $[-\frac{\pi}{2}, \frac{\pi}{2}]$ corresponding to a more listener-centric coordinate system.



(A) 3D coordinate system.

(B) 2D coordinate system.

FIGURE 2.1: Spherical coordinate systems for 3D and 2D.

in this work θ is the *colatitude* in 3D, and the *azimuth* in 2D. The conversion from spherical to Cartesian coordinates in 2D is given by

$$\mathbf{r} = \begin{pmatrix} x \\ y \end{pmatrix} = r \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \end{pmatrix}. \tag{2.2}$$

The engineering time convention is adopted, that is simple harmonic dependence on time, t, is given by $e^{j\omega t}$ with the imaginary unit defined by $j = \sqrt{-1}$. The angular frequency, ω, is dependent on frequency, f, by $\omega = 2\pi f$ and related to the wavenumber, k, by $k = \omega/c$ where c is the speed of sound which is assumed to be uniform throughout the medium.

In the following the signal (or gain) for the ℓ-th loudspeaker is given by g_ℓ, whilst target virtual sources are denoted with the subscript T. Head rotation of a listener is represented by the subscript rot.

2.3 Stereophony

2.3.1 Stereophonic Sine Law

Stereophonic recording and reproduction introduced for the first time the concept of spatial audio. Stereo systems were first demonstrated in the 1930's, with a live transmission of a concert to a public demonstration using a three channel system [3] and subjective tests investigating two and three channel systems in [4]. Blumlein pioneered by presenting a full stereo recording and playback system, which includes the famous Blumlein stereo microphone pair [5]. The concept of artificially creating stereo was proposed by Schroeder [6], which was closely followed by the 'stereo law of sines' or simply the 'stereo sine law' [7, 8]. The stereo sine law is an application

of amplitude panning, that is panning between the two loudspeakers with regards to the amplitude of the sound source in each loudspeaker. Interestingly, panning with respective to amplitude results not in a recreated Interaural Level Difference (ILD) at the listener's ears, but rather phase differences that give rise to a recreated Interaural Time Difference (ITD). There exists multiple ways to derive the sine law, of which one approach is to try to reproduce the ITD due to a virtual plane wave source through varying the amplitude of the loudspeakers [7]. The sine law just defines the gains as a ratio, hence a gain normalisation equation is also required to solve for the different loudspeaker gains [9].

Stereo is inherently a low frequency technique, valid below 800 Hz [7]. The sine law assumes a shadowless head model, plane wave virtual sources and that the loudspeakers act as plane waves. In general, two loudspeakers are positioned equidistantly from a listener, at angles $\theta_{1,2} = \pm\theta_\ell$ either side of the listener's head. The listener is assumed to be facing forwards down the line separating the two loudspeakers. The aim is to reproduce a virtual source at an angle of θ_T. The stereo sine law gives the gains required for the two loudspeakers to reproduce the virtual source, subject to a gain normalisation, such that [7]

$$\frac{g_1 - g_2}{g_1 + g_2} = \frac{\sin(\theta_T)}{\sin(\theta_\ell)} \quad \text{subject to} \quad g_1 + g_2 = 1. \tag{2.3}$$

Hence there are two equations which can be solved to find the two unknown gains. Subsequently,

$$\begin{aligned} g_1 &= \frac{1}{2} + \frac{\sin(\theta_T)}{2\sin(\theta_\ell)} \\ g_2 &= \frac{1}{2} - \frac{\sin(\theta_T)}{2\sin(\theta_\ell)}. \end{aligned} \tag{2.4}$$

The stereo sine law may also be formulated for any generalised loudspeaker angular positioning, as well as any number of loudspeakers [10, 11]. However, the technique does require radially equidistant loudspeakers in a 2D plane. In theory, there is no limit to the position of the virtual source; it may be positioned inside or outside of the span of the loudspeakers. Out-of-span panning however requires out of phase loudspeaker signals, which in reality creates a less robust reproduction system to loudspeaker and listener perturbations. Due to this, it is very uncommon for out-of-span panning to be allowed in stereo systems.

2.3.2 Head-tracked Stereo Sine Law

Whilst the stereo sine law considers a fixed listener position, the loudspeaker gains may be formulated to adapt for any listener movements or head rotations. This is

the motivation for the head-tracked stereo sine law, also presented as Compensated Amplitude Panning (CAP) [12, 13, 14]. Naturally, this reproduction method requires tracking of the listener's head in comparison to the position of the loudspeakers. The loudspeaker gains are adapted depending on the listener's head rotation. To compensate for listener translations, delays may be applied to each of the loudspeakers so they are acoustically equidistant. Identical assumptions are made as with the stereo sine law, for example plane wave loudspeakers and virtual sources. The dynamic listener-tracking aspect of the head-tracked sine law makes it unique as few spatial audio reproduction techniques over loudspeakers include listener tracking and compensation.

Whilst the head-tracked sine law is presented here using the trigonometric formulation, the CAP technique derives the formulas using a vector based approach utilising a simplified soundfield representation - a first order Taylor expansion of the plane wave soundfield [12, 15, 16]. The original derivation considers a simple stereo loudspeaker pair but has also been expanded to the multichannel case [14] matching the result of the multichannel stereo sine law [11]. The approach has also been extended to included near-field sources however this transforms the loudspeaker signals from simple frequency independent panning gains to frequency dependent filters [13, 17]. Finally, decoding of first order Ambisonic signals to head-tracked stereo has also been validated [14, 18, 19, 20].

Consider a standard stereo loudspeaker configuration as before, for a listener positioned equidistantly from both loudspeakers however now compensating for a head rotation angle θ_{rot}. The head-tracked sine law gains are given by [12]

$$\begin{aligned} g_1 &= \frac{\sin(\theta_\ell + \theta_{rot}) + \sin(\theta_T - \theta_{rot})}{2\sin(\theta_\ell)\cos(\theta_{rot})} \\ g_2 &= \frac{\sin(\theta_\ell - \theta_{rot}) - \sin(\theta_T - \theta_{rot})}{2\sin(\theta_\ell)\cos(\theta_{rot})}. \end{aligned} \tag{2.5}$$

This gain solution will also be re-derived later in Chapter 5. When the listener is facing forwards such that $\theta_{rot} = 0$ (as required with the stereo sine law) the loudspeaker gains collapse to the standard sine law definitions. To extend the technique to higher frequencies, implementations using an intensity normalisation such that $g_1^2 + g_2^2 = 1$ have been investigated with the technique, where a crossover between $800 - 1000$ Hz is applied to transition between the two regions of differing normalisation [12].

2.3.3 Stereo Tangent Law

An alternative panning approach for stereo loudspeakers is the stereo tangent law. One starting point for the derivation of the tangent law is assuming the listener is positioned in the sweet spot of a standard stereo loudspeaker array, however now

the listener's head is assumed to be facing the position of the virtual source. In this case, the defining equations are [21]

$$\frac{g_1 - g_2}{g_1 + g_2} = \frac{\tan(\theta_T)}{\tan(\theta_\ell)} \quad \text{subject to} \quad g_1 + g_2 = 1. \tag{2.6}$$

This leads to the following gain definitions

$$\begin{aligned} g_1 &= \frac{1}{2} + \frac{\tan(\theta_T)}{2\tan(\theta_\ell)} \\ g_2 &= \frac{1}{2} - \frac{\tan(\theta_T)}{2\tan(\theta_\ell)}. \end{aligned} \tag{2.7}$$

Given that the tangent law assumes the listener is facing the virtual source, the head tracked sine law should be equal to the tangent law when $\theta_{rot} = \theta_T$. Hence, combining this condition with Eqn. 2.5 and using the identity $\sin(x + y) = \sin(x)\cos(y) + \cos(x)\sin(y)$ then

$$\begin{aligned} g_1 &= \frac{\sin(\theta_\ell)\cos(\theta_T) + \cos(\theta_\ell)\sin(\theta_T)}{2\sin(\theta_\ell)\cos(\theta_T)} = \frac{1}{2} + \frac{\tan(\theta_T)}{2\tan(\theta_\ell)} \\ g_2 &= \frac{\sin(\theta_\ell)\cos(\theta_T) - \cos(\theta_\ell)\sin(\theta_T)}{2\sin(\theta_\ell)\cos(\theta_T)} = \frac{1}{2} - \frac{\tan(\theta_T)}{2\tan(\theta_\ell)} \end{aligned} \tag{2.8}$$

therefore the tangent law is a special case of the head-tracked sine law where the listener's head rotation follows the virtual source position. Furthermore, when both θ_T, $\theta_{rot} = 0^\circ$ the approach gives identical loudspeaker gains to the stereo sine law.

The tangent law has been shown to be a 2D formulation of another popular 3D panning aproach, Vector Base Amplitude Panning (VBAP) [22].

2.3.4 Sweet Spot Enlargement

A classic issue with stereo reproduction is due to the presence of the sweet spot; the position at which the loudspeaker signals sum correctly and produce the desired virtual sound source. For a stereo pair of loudspeakers this is the position equidistant between the loudspeakers such that both loudspeakers subtend the desired angular span (classically the total loudspeaker span is 60 degrees). Deviations from this ideal listening position can lead to instabilities in the virtual source position, where large deviations lead to the sound stage collapsing to the closest reproduction loudspeaker [23, 24]. This in practice means the technique does not work properly if the listener is displaced from the sweet spot or there are multiple listeners present.

Efforts have been made to determine the size of the sweet spot [23, 25, 26, 27]. In general, these approaches characterise the sweet spot when assuming the loudspeakers are driven with equal gains to create a central virtual source. The sweet spot is then defined as the area around the central reproduction point where there is a minimal drop off in SPL, often to a -3 dB limit [23].

Approaches to artificially enlarge the stereo sweet spot have mostly focused on defining loudspeaker radiation patterns that maximise the sweet spot as per this definition. In [28] this radiation pattern was inferred from listening tests where the subjects were positioned off axis and manually altered the ratios of the loudspeaker gains so that a central virtual source was accurately reproduced. This technique is titled 'time-intensity' trading [28]. Off-centre listening positions cause the closest loudspeaker to dominate the localisation of the sound source. Time-intensity trading overcomes this by increasing the volume of the furthest loudspeaker to compensate for the output of the closest speaker which arrives both earlier and also naturally louder due to less distance attenuation. Panning with respect to inter-channel time differences also leads to another, less common form, of stereo panning [21, 29, 30]. Delay-panning is generally less robust than panning with respect to amplitude, and increasing the inter-channel delays too much might result in breaking of the virtual image into two separate sources.

In [23] a pair of dipole loudspeakers was suggested to increase the sweet spot area, where the dipoles are focused just in front of the listening position. More complex loudspeaker geometries such as small loudspeaker arrays have also been considered [31], including using loudspeaker radiation patterns again deferred from the results of listening tests considering off-axis listening positions [32].

2.3.5 Sweet Spot Adaption

An alternative approach to the issue of off-centre listening positions is to dynamically move the sweet spot to accommodate the listener movements. The purest form of this in terms of the stereo technique is the head-tracked sine law as presented earlier. However, this has only found recent popularity with its derivation through CAP. Therefore past works have considered other ways to dynamically adapt the stereo sweet spot.

The issue may be divided into compensating for listener translations and listener rotations. The effect of head rotations on the perceived placement of a virtual source using the stereo tangent law was investigated using listening tests in [24]. Using a stereo pair of loudspeakers three different head rotation angles ($\theta_{rot} = 0°, -30°, -60°$) were considered, as well as three source positions ($\theta_T = 0°, \pm 15°$) and four source signals (pink noise, speech, snare drum and chimes). The subjects were asked to adjust the ratio of the loudspeaker gains such that the virtual source most closely matched the position of a real reference source. For a forward facing listener the gain

definitions chosen by the listeners closely matched those of the tangent law, however differed as the subjects rotated their head. An empirical compensation function for the tangent law was then formulated based on the results of a binaural localisation model and the compensated approach was shown to agree with the perceptually selected gain ratios from the listening test. Thus for a full real-time implementation of this compensated method, head tracking must be employed and the gains selected from a look-up table populated as per the listening test results.

To adapt for listener translations the 'sweet spotter' approach applies listener tracking and adaptive delays to the loudspeakers to ensure they remain acoustically equidistant [33, 34, 35]. The approach is considered in a 2D plane only and has been validated using a binaural localization model [36]. Furthermore, an empirical compensation function for listener head rotations has also been implemented [33, 37]. However, as the input to the system is assumed to be channel-based stereo content no access to the individual mono virtual objects is considered, therefore the compensation aims to maintain a stable single central virtual source only and it cannot be assumed the compensation holds for all other virtual source positions.

Finally, a similar approach however now for object input to a stereo loudspeaker system has been considered [38, 39]. This approach also performs compensation of the loudspeaker array based off of listener positional tracking via a camera. The loudspeaker signals are delayed to make the array loudspeakers acoustically equidistant and attenuation due to spherical spreading of the loudspeakers is also compensated for by a distance dependent gain factor.

2.4 Higher Order Ambisonics

Higher Order Ambisonics (HOA) is a full end-to-end theory covering recording, transmitting and reproduction of a soundfield, pioneered by the work of Gerzon [40, 41, 42]. HOA relies on the expansion of a soundfield in an orthogonal basis. In 3D, this is the spherical harmonic expansion [1, 43] whilst in 2D this is a Fourier Bessel Series [44, 45, 46]. In HOA, the expansion is thought of as a modal decomposition, with the chosen orthogonal basis defining the modes[2]. This modal decomposition defines the *B-format* signals which are the weighting coefficients that give the required contribution of each mode to the total soundfield. B-format thus gives a method for recording, synthesising, modifying and transmitting a soundfield that is independent of the reproduction or capture medium. The approach is thus a scene-based technique. The modal decomposition relies on an infinite summation to fully recreate the soundfield, however for any practical system this must be truncated to a finite order N [47]. This leads to an upper frequency limit to which the approximation is valid to within a set error bound [48]. As traditionally Ambisonics was

[2]The use of the term 'mode' is not strictly accurate considering its mathematical definition, however it has been widely adopted by the Ambisonics community.

presented as a technique considering just zeroth and first order terms, the name HOA is historically used to indicate expansion beyond the first order.

For reproduction over loudspeaker arrays, decoders are defined to transform the B-format into loudspeaker signals. The most relevant for this work is mode matching, which corresponds to finding the minimum L^2 norm solution for the loudspeaker gains that matches the modes of the reproduced soundfield to the target B-format signals, up to the truncation order [42, 44]. This solution has been shown to equal a number of other soundfield reproduction methods. These include representing the decoding as a minimisation problem [48], performing pressure matching across a surface (a circle for 2D, a sphere for 3D) [43, 49] and in a similar manner a spatial sampling of the original continuous soundfield over a surface [50, 51, 52]. The mode matching decoder is sometimes called 'physical' or 'pseudoinverse' decoding. An alternative, the 'sampling decoder', pre-dates the mode matching approach and utilises a simpler matrix transpose, however both are equal when utilising special loudspeaker arrangements (equal angle circular sampling in 2D and t-designs in 3D) [1]. For the classic mode matching scenario the target virtual sources and loudspeakers are all assumed to be plane waves, and HOA is both a panning technique and a soundfield reproduction approach. However nearfield sources and nearfield loudspeakers may be considered at the cost of introducing frequency-dependent gains [50, 53].

The number and positions of the loudspeakers heavily dictate the performance of a mode matching decoder. For order N reproduction at least $(N+1)^2$ loudspeakers are required for an exact solution [48]. Poor angular sampling of the sphere leads to ill-conditioning of the problem and large loudspeaker gains when virtual sources are positioned in poorly sampled areas, which can be dealt with using Tikhonov regularisation [54]. Furthermore, only specific loudspeaker arrangements titled 't-designs' lead to panning functions with consistent overall energy of the gain definitions, for all source positions and a given order truncation [55]. Interestingly, t-designs require more than $(N+1)^2$ loudspeakers for order N reproduction, which means the loudspeaker array can control spherical harmonics to an order $N' > N$ and this leads to spectral impairment in the reproduced soundfield [56]. Whilst the input B-format modes with $N < n \leq N'$ are unspecified, the mode matching approach sets these modes to have zero energy as it utilises a minimum norm solution when $L > (N+1)^2$ [57]. Considering the spherical harmonic modes of the reproduced soundfield there is then correct reproduction up to order N, zero energy between N to N' and spatial aliasing beyond N'. Practically this leads to a reduction in energy at high frequencies/large distance steps from the reproduction point [56, 58]. This issue is described in-depth in [57] where the zero energy region is called the 'ring of silence'. Interestingly, using no more than the minimum number of loudspeakers required and allowing spatial aliasing in all modes above N can be a more favourable approach perceptually [58]. HOA mode matching is very relevant to this thesis thus

a full derivation of the approach is presented in the following chapter for both the 3D and 2D scenario.

A second classic approach to decode B-format to loudspeaker signals is psychoacoustic decoding [42]. Psychoacoustic decoding assumes the loudspeaker signals sum incoherently at the central reproduction point and hence considers controlling the energy at this position. The loudspeaker signals may also be forced to be in-phase. This decoding is more effective for high frequencies where it is more likely the incoherent summation is a physical reality [42].

HOA and VBAP have been used together in various manners, attempting to mix the advantages introduced by both techniques. VBAP is an alternative vector based amplitude panning approach that will be reviewed later in this chapter. Whilst VBAP ensures panning-invariant energy for arbitrary loudspeaker arrangements (consistent total energy of the loudspeaker gains regardless of source position), this is only the case with HOA when using t-design loudspeaker arrangements. To overcome this issue for irregular loudspeaker arrays, both the All-Round Ambisonic Panning (ALLRAP) and All-Round Ambisonic Decoding (ALLRAD) approaches use VBAP over a real and irregular loudspeaker array to create a virtual loudspeaker array positioned in an ideal HOA t-design formation. The soundfield is then recreated using standard HOA decoding approaches over this (virtual) ideal HOA arrangement [55, 59, 60]. The ALLRAP approach is defined for single virtual sources whilst the ALLRAD technique is the extension to cover reproduction of B-format signals.

Further alternatives to more traditional HOA decoding methods include sparse decoding with respect to the L^1 norm. Here the goal is to create panning functions that use minimal numbers of loudspeakers, but still recreate the spherical harmonic modes correctly and thus perform soundfield reproduction. Such techniques may lead to a perceptually better decoder, however come at the cost of requiring complicated iterative algorithms to calculate the decoders such as LASSO [61, 62].

HOA may also be decoded binaurally using two key methods. The first is decoding to virtual loudspeaker arrays with HRTF entries corresponding to the loudspeaker positions [63]. The loudspeaker positions may be fixed relative to the listeners head (head-locked), as listener head rotations may be compensated for directly using the B-format representation before decoding to loudspeaker signals. This has the advantage of a simpler real-time implementation, with a smaller dataset of HRTFs needed in the renderer. The second method is direct rendering in the spherical harmonic domain between the B-format signals and that of the spherical modal decomposition of the HRTF [64]. To calculate the spherical coefficients of the HRTF it must first be spatially sampled to give the HRTF for different source positions. If these source positions are equal to those used in a virtual loudspeaker rendering approach, both binaural decoding techniques are identical [57].

2.5 Further Panning Approaches

In this section, a range of other spatial audio panning approaches will be reviewed, which are commonly used in the field but less relevant to this work.

Vector Base Amplitude Panning (VBAP) is a 3D audio technique designed for loudspeaker arrays of more than two loudspeakers, proposed by Pulkki [65, 66]. VBAP is a popular and widely implemented technique, aided by its ability to be implemented on any irregular loudspeaker layout and also its robust implementation due to its gain non-negativity constraint. VBAP splits the loudspeaker array into sectors formulated by arcs of loudspeaker pairs (2D) or triangles of loudspeaker triplets (3D), also referred to as bases. A sparse solution is formulated such that when the virtual source position is within a given sector only this set of loudspeakers is activated.

The problem is formulated by considering positional unit vectors pointing towards each loudspeaker $\mathbf{l}_\ell$, and towards the virtual source position, $\mathbf{p}_T$. A plant matrix $\mathbf{L} = [\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3]$ is defined such that the loudspeaker gains, $\mathbf{g}$, are found by the combination of an inversion of the plant matrix and a normalisation constraint.

$$\mathbf{g} = \mathbf{L}^{-1}\mathbf{p}_T \text{ subject to } \sum_{\ell=1}^{3} g_\ell^p = C \tag{2.9}$$

where the normalisation constant is often set as $C = 1$ and $p = 1, 2$ corresponding to amplitude or intensity style normalisations respectively. When the virtual source falls on a line between only two loudspeakers, the third is not activated. Furthermore, when the source is in the centre of the triangle made by all three loudspeakers all gains are equal. Virtual source positions outside of the sector are not allowed, thus ensuring the loudspeaker gains always remain positive. Interestingly, when considering the 2D formulation of the problem the solution is equal to the stereo tangent law [65].

For the VBAP approach a single loudspeaker is activated when the virtual source falls at the loudspeaker position. This leads to a different perceived source width as the source is panned across the loudspeaker array, with maximal width when between loudspeakers and minimal width at loudspeaker positions. Whilst minimal source width is generally an advantage, the irregularity across all possible virtual source positions is a perceptual issue. The Multiple Direction Amplitude Panning (MDAP) approach was proposed as an extension to VBAP to solve this issue [67]. MDAP purposely enforces a fixed source width for all virtual source positions by always activating multiple loudspeakers. Thus whilst the source width is not always the minimal width the system could achieved, it remains consistent across all source positions.

In general panning solutions often lead to gain ratios that require a normalisation equation to ensure the loudspeaker gains can be calculated, expressed as

$$\sum_{\ell=1}^{L} g_\ell^p = C. \tag{2.10}$$

The most common solutions are amplitude normalisation ($p = 1$) and intensity normalisation ($p = 2$). A traditional approach is to use both normalisation's in different frequency bands, as seen with Vector Base Intensity Panning (VBIP) [68] and also first order Ambisonics [42]. At low frequencies, amplitude normalisation is used at it is assumed here the loudspeakers sum coherently. However, at high frequencies incoherent summation is assumed thus the intensity normalisation is more appropriate [9]. VBIP is a direct extension of VBAP employing this dual-band normalisation with a cutoff between the two normalisation approaches at 700 Hz.

Distance Based Amplitude Panning (DBAP) is a form of intensity panning where loudspeaker gains are calculated by considering the euclidean distance from the source position to each given loudspeaker [69]. All loudspeakers in the array could be activated for any source position, however weights may be applied in the formulation to force the use of a subset of the loudspeakers. In general, loudspeaker weights are applied to favour loudspeakers closest to the source position [70]. Whilst most amplitude panning approaches are limited to reproducing sources at the same radial distance as the loudspeaker array, DBAP provides a preliminary approach to render virtual sources outside of the array. In this case, the source position is mapped to the closest position on the array boundary, then further processing to create a perceptual illusion of distance is applied (for example gain attenuation, air filtering, increased reverberation) [69].

Further developments have been made to create interior panning approaches that can create virtual sources positioned inside of the loudspeaker array, of which a thorough review may be found in [71]. One advanced approach for interior panning is Auditory Distance Rendering [72]. An observation is first made regarding the main localisation cues that lead to the perception of distance, these cues are summarised as

- Spectral: For example high frequency attenuation, important for distances larger than 15 metres
- Binaural: For example nearfield ILD, important for distances less than 1 metre
- Intensity: For example due to spherical spreading
- Direct to Reverberant Ratio

Auditory Distance Rendering ignores the first two categories as these are prominent for extreme distances only. The required change in intensity is calculated based

on the requested distance. Rendering of a change in direct to reverberant ratio is based on observations using an auditory model and binaural measurements that various metrics of coherence between the two ears of a listener drop as distance increases in reverberant conditions. Rendering is then performed using a left-centre-right trio loudspeaker array. The centre loudspeaker is used for the direct sound, whilst the left and right loudspeakers use decorrelated signals weighted by a gain that increases with distance. As this fixes the virtual source position to the centre loudspeaker, the approach is extended using VBAP and any given loudspeaker array to create a virtual left-centre-right trio centred on the virtual source position.

2.6 Crosstalk Cancellation

Crosstalk Cancellation (CTC) aims to correctly reproduce the pressure at the position of the listener's ears only. To do so, soundfield control by means of inverse filtering is employed to ensure sufficient channel separation between each of the listener's ears [73, 74]. CTC differs from other loudspeaker based techniques as it is a form of binaural reproduction. Furthermore, the approach requires frequency-dependent filters as opposed to simple frequency-independent panning laws. Recent trends in the literature have favoured the use of multichannel loudspeaker arrays, particularly linear arrays [75, 76]. The approach may also be formulated for any number of listeners, although the largest demonstrated working system is currently for 3 separate listener positions [77, 78]. The CTC approach results in exact reproduction sweet spots however listener tracking may be included to create a dynamic sweet spot that moves with the listener [79, 80].

Let M be the number of listener ears, or control points, and L be the number of loudspeakers. For a given frequency the reproduced pressure at the control points, $\mathbf{p}$, a column vector of length M, may be written as [54]

$$\mathbf{p} = \mathbf{C}\mathbf{g} = \mathbf{C}\mathbf{H}\mathbf{p}_T \tag{2.11}$$

where $\mathbf{C}$ is the $M \times L$ plant matrix of acoustic transfer functions that completely describe the reproduction system, the listener's HRTF and the geometry of the problem, $\mathbf{H}$ is the $L \times M$ matrix of CTC filters and $\mathbf{p}_T$ is the length M column vector of target pressures to be reproduced at each control point. The goal in CTC is to design the set of CTC filters, $\mathbf{H}$, such that it is as close to the inverse of $\mathbf{C}$ as possible so that $\mathbf{p}_T = \mathbf{p}$. In reality, this is complicated by issues such as listener and loudspeaker perturbations and ill-conditioning of the plant matrix under inversion [54].

CTC has previously been considered a completely different approach to soundfield reconstruction methods, as suggested in [2]. However, it has been shown at low frequencies, when assuming plane wave virtual sources and loudspeakers, using a

low frequency approximation of a rigid sphere HRTF and a standard stereo loudspeaker arrangement then the CTC loudspeaker gain solutions are equivalent to the head-tracked stereo sine law [81, 82]. This means at low frequencies CTC may also be thought of as a panning technique.

2.7 Binaural Techniques

Binaural audio is a 2 channel spatial audio format resulting in specific channels for the left and the right ears of the listener. When reproduced exactly the audio mimics what the listener might hear if they were actually present in the recorded acoustic scene, thus they hear and perceive the audio spatially [83, 84]. The simplest approach to create binaural audio is by recording the 2 channels directly using a dummy head or whilst wearing a pair of binaural microphones fitted in the ears. However, whilst this approach is very simple the recording is head-locked to the orientation and position of the recorded head thus the sound scene can not be easily altered (for example it can't be rotated) after the recording has occurred. Furthermore, the HRTF used in the recording is hard encoded into the binaural audio, which can lead to issues if there is a significant mismatch between this and the listener's own HRTF [83]. The success of binaural audio is due to the inclusion of localisation cues encoded in the HRTF which the brain uses to infer the spatial position of a sound source.

Binaural audio is most commonly reproduced using headphones, however as mentioned previously CTC systems provide an alternative approach using loudspeakers. The HRTF and key localisation cues will now be reviewed, after which approaches for the synthesis of binaural audio will be presented.

2.7.1 Spatial Hearing

HRTFs

The transmission of a sound source to a listener's two ears can be modelled as a linear time-invariant system. Therefore, the filtering introduced by the physical presence of the listener, such as scattering or head shadowing, can be encoded in an acoustic transfer function referred to as the Head-Related Transfer Function (HRTF), or its time domain equivalent the Head-Related Impulse Response (HRIR) [84].

The HRTF is defined as the pressure due to a sound source measured at a position in each of the listener's ears, normalised by the pressure due to the source at the head's centre in the absence of the listener [83].

$$HRTF_{l,r}(r,\theta,\phi) = \frac{p_{l,r}(r,\theta,\phi)}{p_0(r,\theta,\phi)} \tag{2.12}$$

where $p_{l,r}$ is the pressure measured at the listener's left and right ears respectively and p_0 is the free field equalisation term with the listener absent. The HRTF is not only dependent on the angular position of a source but also the radial distance, resulting in far field and near field formulations of the HRTF [84].

The HRTF is unique to any given person, as it depends on the physical dimensions and properties of a listener's head. However, generic HRTFs remain widely used due to the difficulty in measuring or obtaining a HRTF that is individualised to a single listener. The phrases 'individualised' or 'personalised' HRTF are often used interchangeably with respect to a listener's true, unique HRTF (that has been either measured or simulated) or one that has been adapted or chosen as a closest fit to their own. Mismatches between the HRTF used for the binaural synthesis/recording and that of the listener's can lead to issues such as warped localisation cues, in-head localisation and front-back confusions [85, 86, 87]. Further work using auditory models validate that localisation cues are impaired when this HRTF mismatch occurs [88].

However, the advantages that individualised HRTFs bring is an ongoing research question, and subjective work does not always confirm these findings. In [89] a listening test revealed no significant improvement in externalisation, localisation or resolution of front-back confusions when using individualised HRTFs. In [90] subjects played a VR game aimed to assess localisation accuracy over a 2 week 7 game period, comparing individual and fitted HRTFs. Overall, whilst localisation accuracy improved over the full period, little benefit was found over using the individual HRTF compared to the best-fit selected HRTF. This suggests that influences such as training and familiarity with the game and binaural audio can also influence how well a subject localises sound when using HRTFs. A similar study [91, 92, 93] investigated HRTF individualisation through database selection in a VR shooter game and found matching results. Here no benefit was found in using an individual HRTF, except in very early rounds of the game. As the game progressed training and familiarity with both the game and whichever HRTF was used for the rendering dominated the need for individualisation. In [94] a VR listening test was conducted to investigate the 'Quality of Experience' when using a stereo audio signal, generic HRTF or individualised HRTF through anthropometric pinna measurements leading to HRTF database selection. Overall, it found no significant difference between the three audio conditions, even the non-3D stereo renderer, indicating that visuals can provide overwhelming localisation cues that are potentially stronger than audio cues.

Localisation Cues

The brain uses a number of different cues to identify the spatial location of a sound source. Furthermore, different cues may have different weightings depending on the frequency range in question [83]. Localisation cues are not only limited to acoustical

cues, for example visual cues are very powerful and can overcome acoustical localisation. Binaural cues are based on differences between the signals at the listener's ears, whilst monoaural cues may be inferred from the signal at one of the ears only. Dynamic cues rely on time-varying changes to the sound scene, for example due to the source or the listener's head moving.

The two most important binaural cues are the Interaural Time Difference and Interaural Level Difference. Both were first recognised by Lord Rayleigh as part of his *Duplex Theory* [95]. The ITD is the difference in time of arrival of an auditory event between the two ears of the listener, due to differences in the path lengths from the sound source to each of the ears. The difference in path lengths creates an Interaural Phase Difference (IPD) that can be used for localization as it is source position dependent. It is generally accepted that at low frequencies below 1.5 kHz, the ITD due to phase differences is a key cue in localization [96, 97]. At high frequencies there exists Interaural Envelope Delay Differences. This is the ITD calculated from the modulation of the envelope of the signal and can be used for localization above 1.5 kHz [98].

For any given value of ITD there is an infinite number of spatial positions from which the given ITD could arise. The possible positions are mapped to a cone radiating from the origin out to an infinite radius. This gives rise to the term the *"cone of confusion"* [96]. An extreme example of this problem arises when a sound source is positioned on the median plane ($\phi = 0°$) where the ITD is zero for all possible values of elevation.

Dynamic cues may be introduced by a moving sound source or by movements of the head such that a change in the ITD and ILD is dynamically introduced [99]. This is an important cue for overcoming front-back confusion as the time-varying analysis of the localisation cues collapses the cone of possible positions to just one point. It has been shown experimentally that rotations around a vertical axis aid in localization through the use of dynamic cues [100]. In a similar manner, head movements can also create a dynamic ILD due to dynamic head shadowing [96].

The ILD arises primarily due to the shadowing effect of the head [96]. For a sound source positioned at a point around the head, the shadowing effect results in attenuation of the pressure at the contralateral (opposite side) ear and in some cases a frequency dependent amplification at the ipsilateral (same side) ear due to a baffle effect. The shadowing is due to the absorption of the incident wave by the head, as well as the constructive or destructive addition of multipath diffracted waves which travel around the head with varying path lengths. The ILD is a frequency dependent cue. At low frequencies, when the wavelength is much larger than the head radius, the head is acoustically invisible to the incident wave and there is no ILD. Conversely, at high frequencies, when the wavelength is comparable to or smaller than the radius of the head then the scattering due to the head has a greater effect,

causing an ILD. The ILD is also distant-dependant, such that when a sound source is very close to the head there exists a low frequency ILD [101].

Monoaural cues use information native to a single ear signal, for example spectral colouration due to the pinna. The pinna causes an incident sound wave to diffract and introduces reflections in a complex manner, that makes it hard to model [102]. Furthermore, the filtering introduced is reliant on the shape and size of the pinna which changes between individuals. An important feature introduced by the pinna are large notches and peaks in the magnitude of the frequency response of the HRTF. Depending on the position of a sound source around the head, the exact location of these notches and peaks varies and hence can be used as a localization cue [103]. Pinna cues are particularly important for discerning the elevation of a sound source [104]. In a similar manner scattering and reflections introduced by the presence of a torso and shoulders aid elevation localization [105].

2.7.2 Direct HRTF Rendering

The simplest approach to render binaural audio is an object-based method where each individual sound source is directly convolved in real-time with a HRIR (or equivalently multiplied in the frequency domain with a HRTF) [83]. Head-tracking may be employed, in which case the positions of the sound sources are altered depending on the head rotation and the correct HRTF is selected from a database [89].

This rendering process must be repeated for each audio object individually which can lead to a significant computational load for complex audio scenes. Furthermore, the approach is dependent on a well-defined HRTF dataset. This dataset should have significant measurement points across the whole sphere so that sound sources can be rendered at all positions around the listener. However as a continuous measurement distribution is not a physical reality, HRTF interpolation can be used to artificially create positions in between the sampled measurement grid. Interpolation approaches include bilinear interpolation, VBAP and spherical harmonic decomposition [84].

Pre-processing of the measured HRTFs can greatly improve their performance when such interpolation methods are used. A primary issue is that due to the ITD embedded in the HRIR, the two main peaks of the HRIRs to be interpolated between do not align in time. This can lead to significant comb filtering or 'main peak widening' in the interpolated HRTF at noted in [106]. Removal of the ITD before interpolation leads to a significant improvement in the accuracy of the interpolated HRTFs magnitude [106, 107] although means the ITD must be replaced in the rendering process. Whilst the ITD can then be personalised towards a listener, this approach requires access to the position of the virtual object to infer the value of the rendered ITD. This time alignment approach also compacts the HRTF in the spherical harmonic

domain, meaning the energy of the HRTF is focused into lower order spherical harmonic channels [108, 109]. Similar issues with replacement of the ITD exist for this case however. Finally, hybrid approaches, for example Mag-LS [110, 111], which utilise forms of time alignment only at high frequencies where phase is perceptually less important have also been proposed. These are particularly relevant to HOA where the HRTF (both magnitude and phase) can be reproduced accurately up to a high frequency aliasing limit. Above this frequency only the ITD is removed, prioritising the reproduction of the magnitude of the HRTF but introducing incorrect phase information. This performs better perceptually than allowing spatial aliasing in both the magnitude and phase response of the HRTF above this frequency limit.

2.7.3 Virtual Loudspeaker Rendering

Virtual loudspeaker rendering is an approach to extend loudspeaker array based techniques to binaural reproduction. Here, as opposed to using a physical array of loudspeakers placed around a listener (which can be challenging for many practical reasons such as cost and physical space), instead an array is virtually created by using HRTFs corresponding to each desired loudspeaker position [63, 84]. The loudspeaker signals are then 'played back' through the virtual array. This approach is advantageous as it ensures any loudspeaker based spatial audio technique can be reproduced binaurally. Listener rotations may be compensated for by switching the HRTFs to align the virtual array with the new head orientation. However techniques such as HOA can compensate head rotations before defining the loudspeaker signals and thus the virtual loudspeaker array can remain head-locked in this scenario.

2.8 Chapter Review

This chapter has presented a review of the state-of-the-art in a number of relevant spatial audio techniques to this thesis. These include a range of stereo approaches including adaptive and dynamic panning, HOA, CTC and binaural rendering techniques. Where appropriate loudspeaker gain definitions for each technique and any fundamental assumptions the approach makes regarding the target soundfield, reproduction loudspeaker array or the listener were reviewed. Furthermore the key localisation cues that form the foundation of human spatial hearing were also covered.

Chapter 3

Higher Order Ambisonics Mode Matching

The goal of this chapter is to review in detail the established principles of the HOA mode matching decoder. First, the mathematical preliminaries of the underlying orthogonal basis expansions for HOA in 3D and 2D are reviewed. The mode matching solutions for both 3D and 2D systems for a single virtual plane wave source are then derived from these soundfield expansions. The approach of the derivation and the results form a foundational basis for the new Higher Order Stereophony (HOS) approach presented in later chapters, where it will also be shown that there exists fundamental links between the core HOA and HOS principles. The 3D mode matching scenario is presented first, due to the compactness of the mathematics in comparison to the 2D case.

3.1 3D Mode Matching

3.1.1 Spherical Harmonics

The spherical harmonics are a set of functions that form an orthonormal basis for square-integrable well-behaved functions over the 2-sphere, $\mathcal{S}^2$ (a unit sphere). Hence, such functions on a sphere may be expressed as a weighted linear summation of spherical harmonics.

The spherical harmonic of order n and degree m [1] may be defined in a complex form as [112]

$$Y_n^m(\theta,\phi) = \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-m)!}{(n+m)!}}P_n^m(\cos\theta)e^{jm\phi} \tag{3.1}$$

where $P_n^m(\cos\theta)$ is the associated Legendre polynomial. The real part of the spherical harmonic functions up to order four for positive m only is shown in Fig. 3.1.

[1] Outside of spatial audio literature in mainstream Physics and Mathematics, this would be defined as degree n and order m.

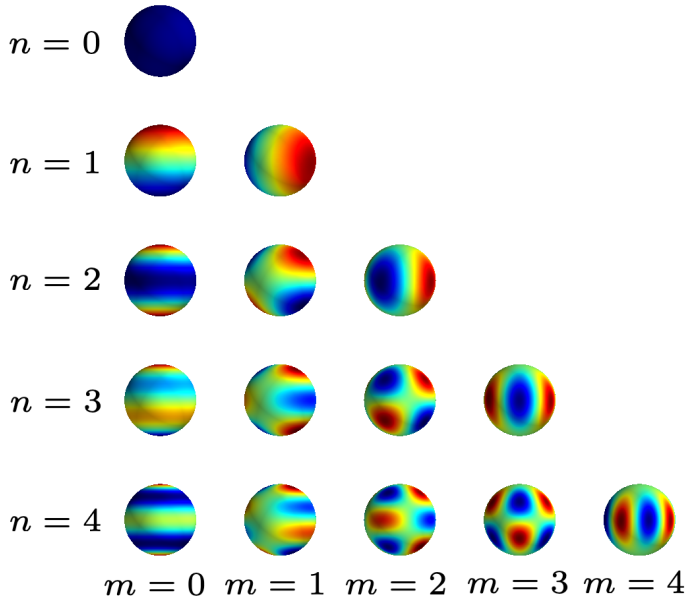


FIGURE 3.1: Real part of the complex spherical harmonics to order 4.

The spherical harmonics may also be defined in a real basis, where n, m are written as subscripts to identify as the real spherical harmonics [113]

$$Y_{n,m}(\phi,\theta) = \begin{cases} \sqrt{2}\sqrt{\frac{(2n+1)}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}}P_n^{|m|}(\cos\theta)\sin(|m|\phi) & \text{if } m < 0 \\ \sqrt{\frac{(2n+1)}{4\pi}}P_n^{|m|}(\cos\theta) & \text{if } m = 0 \\ \sqrt{2}\sqrt{\frac{(2n+1)}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}}P_n^{|m|}(\cos\theta)\cos(|m|\phi) & \text{if } m > 0. \end{cases} \tag{3.2}$$

The real spherical harmonics are useful in certain applications due to the simplified form of the azimuthal dependence ϕ. For the real spherical harmonics, this is simply $\sin(|m|\phi)$ for $m < 0$ or $\cos(|m|\phi)$ for $m > 0$. We note that for $m = 0$, the complex and real definition of the spherical harmonics are equal and therefore real. There exists a simple mapping between the real and complex spherical harmonics

$$\begin{aligned} Y_{n,m}(\phi,\theta) &= \begin{cases} \sqrt{2}\ \mathrm{Im}\left\{Y_n^{|m|}(\phi,\theta)\right\} & \text{if } m < 0 \\ Y_n^0(\phi,\theta) & \text{if } m = 0 \\ \sqrt{2}\ \mathrm{Re}\left\{Y_n^{|m|}(\phi,\theta)\right\} & \text{if } m > 0. \end{cases} \\ Y_n^m(\theta,\phi) &= \begin{cases} \frac{1}{\sqrt{2}}\left(Y_{n,|m|} - jY_{n,-|m|}\right) & \text{if } m < 0 \\ Y_{n,0}(\phi,\theta) & \text{if } m = 0 \\ \frac{1}{\sqrt{2}}\left(Y_{n,|m|} + jY_{n,-|m|}\right) & \text{if } m > 0. \end{cases} \end{aligned} \tag{3.3}$$

where the operators $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ correspond to taking the real and imaginary

parts respectively. Both the complex and real spherical harmonics form an orthonormal basis over the unit 2-sphere and as seen in Eqn. 3.3 there is a defined mapping between complex and real spherical harmonics of a given order n and degree m. Consequently we can interchangeably use either definition of the spherical harmonics and generally the key mathematical properties of the functions hold equally for both definitions, except for the application of rotations where the complex spherical harmonics are simpler [114]. Although notably, for real time signal processing using the real spherical harmonics leads to a reduction by half in the number of operations required due to using just real numbers, not complex. In much of the Ambisonics literature, the real spherical harmonics are generally chosen [1, 44, 113], whereas in soundfield control literature the complex spherical harmonics are more frequently used [43, 48]. For the remainder of this thesis, the complex definition will be adopted, however we note that the mathematics may be rewritten using the real definitions.

Finally, we may equivalently write $Y_n^m(\theta,\phi)$ or $Y_n^m(\hat{\mathbf{r}})$ where the unit vector $\hat{\mathbf{r}} = [\sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi), \cos(\theta)]^T$ with the superscript $(\cdot)^T$ representing the transpose operator.

Orthogonality

The spherical harmonics are orthonormal [112]

$$\int_\Omega Y_n^m(\theta,\phi) Y_{n'}^{m'}(\theta,\phi)^* d\Omega = \delta_{nn'}\delta_{mm'} \tag{3.4}$$

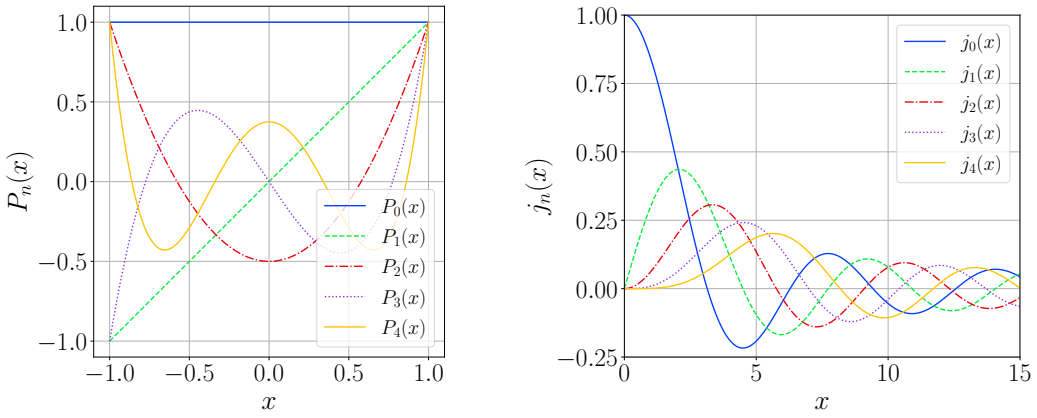
where $\int_\Omega d\Omega = \int_0^{2\pi}\int_0^{\pi} \sin(\theta) d\theta d\phi$ is integration over the $\mathcal{S}^2$ sphere and δ_{pq} is the Kronecker delta function defined as

$$\delta_{pq} = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}. \tag{3.5}$$

The Spherical Harmonic Transform

A function, $f(\theta,\phi)$ that is square integrable on $\mathcal{S}^2$ may be expressed as a weighted linear summation of spherical harmonics. This defines the Spherical Harmonic Transform ($\mathcal{SHT}$) and Inverse Spherical Harmonic Transform ($\mathcal{SHT}^{-1}$) [115]

$$\begin{aligned} f_n^m &= \int_\Omega f(\theta,\phi) Y_n^m(\theta,\phi)^* d\Omega =: \mathcal{SHT}\{f(\theta,\phi)\} \\ f(\theta,\phi) &= \sum_{n=0}^{\infty}\sum_{m=-n}^{n} f_n^m Y_n^m(\theta,\phi) =: \mathcal{SHT}^{-1}\{f_n^m\} \end{aligned} \tag{3.6}$$



(A) The Legendre polynomials up to order 4. (B) The spherical Bessel functions up to order 4.

FIGURE 3.2: Legendre polynomials and spherical Bessel functions.

where the operator $(\cdot)^*$ denotes complex conjugation. The $\mathcal{SHT}$ is a form of generalised Fourier transform, where the coefficient f_n^m gives the weighting for the order n and degree m spherical harmonic.

The Addition Theorem And The Legendre Polynomials

The addition theorem states that [112]

$$P_n(\hat{\mathbf{x}} \cdot \hat{\mathbf{y}}) = \frac{4\pi}{(2n+1)} \sum_{m=-n}^{n} Y_n^m(\hat{\mathbf{x}})^* Y_n^m(\hat{\mathbf{y}}) \text{ where } \hat{\mathbf{x}}, \hat{\mathbf{y}} \in \mathcal{S}^2. \tag{3.7}$$

$P_n(x)$ is the n-th order Legendre polynomial, defined as the subset of associated polynomials with $m = 0$, that is $P_n(x) \equiv P_n^{m=0}(x)$. $P_n(x)$ is a polynomial in x to the order n. The Legendre polynomials up to order $n = 4$ are shown in Fig. 3.2(A). The Legendre polynomials form an orthogonal basis over the region $[-1, 1]$ [115]

$$\int_{-1}^{1} P_n(x) P_{n'}(x) dx = \frac{2}{(2n+1)} \delta_{nn'} \quad \forall\, n \in \mathbb{N}_0. \tag{3.8}$$

The parity of the Legendre polynomials leads to the following useful relation

$$P_n(-x) = (-1)^n P_n(x) \implies P_n(-1) = (-1)^n. \tag{3.9}$$

The Spherical Bessel Functions

As will be seen when using the spherical harmonic expansion of a soundfield, the angular dependence of a function, (θ, ϕ), may be separated from its frequency and radial dependence, kr. The quantity kr is useful as it reflects how a change in frequency is equivalent to a change in radial distance (and vice-versa). In general, the

expansions used in this thesis rely on the spherical Bessel functions for the kr dependence. The spherical Bessel functions up to order $n = 4$ are shown in Fig. 3.2(B). Importantly, for small kr the high order spherical Bessel functions are ≈ 0. This means that for small distances about the expansion point, or equivalently low frequencies, only low order angular functions are sufficient to approximate the function to a high degree of accuracy.

The spherical Bessel functions form an orthogonal basis over the region $[-\infty, \infty]$ [116]. The orthogonality relation, which is seldom used and therefore derived in Appendix. C, is

$$\int_{-\infty}^{\infty} j_n(x) j_{n'}(x) dx = \frac{\pi}{(2n+1)} \delta_{nn'} \quad \forall\ n\ \in\ \mathbb{N}_0. \tag{3.10}$$

The following relationship holds for considering negative arguments [117]

$$j_n(-x) = (-1)^n j_n(x). \tag{3.11}$$

The 3D Jacobi-Anger Expansion

The pressure due to a plane wave incident with wavevector and wavenumber $\mathbf{k}' = k\hat{\mathbf{k}}'$ at a point $\mathbf{r}$ is given by $p(kr, \theta, \phi) = e^{-j\mathbf{k}'\cdot\mathbf{r}}$. In 3D, $\mathbf{k}' = [k'_x, k'_y, k'_z]^T$ with $k'^2 = k'^2_x + k'^2_y + k'^2_z$ and the dot product $\mathbf{k}' \cdot \mathbf{r} = k'_x x + k'_y y + k'_z z$. The unit vector $\hat{\mathbf{k}}'$ gives the *propagation direction* of the plane wave, however it is more convenient when discussing spatial audio to consider the *arrival direction*. The arrival direction, $\hat{\mathbf{k}}$, is the opposite to the propagation vector, $\hat{\mathbf{k}} = -\hat{\mathbf{k}}'$ and will be used in this work in which case $p(kr, \theta, \phi) = e^{j\mathbf{k}\cdot\mathbf{r}}$.

The 3D Jacobi-Anger expansion expands a plane wave as a summation of spherical harmonics [118]

$$p(kr, \theta, \phi) = e^{j\mathbf{k}\cdot\mathbf{r}} = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi j^n j_n(kr) Y_n^m(\theta_i, \phi_i) Y_n^m(\theta, \phi)^* \tag{3.12}$$

with the direction of the plane wave given by (θ_i, ϕ_i). Because of the symmetry of the spherical harmonics, the complex conjugate may be freely switched between the two spherical harmonic functions in Eqn. 3.12 [115]. Using the spherical harmonic addition theorem, Eqn. 3.7, the Jacobi-Anger expansion may be expressed purely in terms of Legendre polynomials

$$p(kr, \theta, \phi) = e^{j\mathbf{k}\cdot\mathbf{r}} = \sum_{n=0}^{\infty} j^n (2n+1) j_n(kr) P_n(\hat{\mathbf{k}} \cdot \hat{\mathbf{r}}). \tag{3.13}$$

The product $\hat{\mathbf{k}} \cdot \hat{\mathbf{r}} = \cos(\Theta)$, where Θ is the angle between $\hat{\mathbf{k}}$ and $\hat{\mathbf{r}}$ leading to

$$p(kr,\theta,\phi) = e^{jkr\cos(\Theta)} = \sum_{n=0}^{\infty} j^n(2n+1)j_n(kr)P_n(\cos\Theta). \tag{3.14}$$

In the special case when $\hat{\mathbf{k}}$ aligns with the z axis, such that $\mathbf{k} = k\hat{\mathbf{z}}$, then $\Theta = \theta$, which is the colatitude angle.

3.1.2 Target Soundfield

The HOA mode matching equations in 3D will now be derived. In spatial audio, the goal is often to reproduce a given target soundfield to create the acoustical illusion. Here, a virtual sound source of a plane wave representing a single object positioned in 3D space will be considered. Let a plane wave be incident from direction $\mathbf{k}_T = k\hat{\mathbf{k}}_T$, such that at a point $\mathbf{r} = r\hat{\mathbf{r}}$ the pressure due to the plane wave is $p_T(kr,\theta,\phi) = e^{j\mathbf{k}_T\cdot\mathbf{r}}$. The subscript T indicates that it is the target soundfield. Using the 3D Jacobi-Anger expansion in Eqn. 3.12 the plane wave may be expressed as

$$p_T(kr,\theta,\phi) = e^{j\mathbf{k}_T\cdot\mathbf{r}} = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} 4\pi j^n j_n(kr)Y_n^m(\theta_T,\phi_T)Y_n^m(\theta,\phi)^*. \tag{3.15}$$

3.1.3 Reproduced Soundfield

Consider a reproduction array of L loudspeakers equidistant from the origin by a distance r_ℓ. For non-equidistant loudspeakers, a delay may be applied to make them acoustically equidistant. The loudspeakers are assumed to act as propagating plane waves, an assumption that is valid for $r_\ell \gg \lambda$ [43]. The target virtual source is assumed to be at the same distance as the boundary of the loudspeaker array. As discussed previously near-field sources can be considered, at the expense of creating frequency-dependent loudspeaker gains as opposed to simple frequency-independent panning functions. The ℓ-th loudspeaker is driven by a gain g_ℓ, is situated at position (θ_ℓ,ϕ_ℓ) and has an associated wavevector $\mathbf{k}_\ell$. The reproduced field, $p_R(kr,\theta,\phi)$ with subscript R, created by the loudspeaker array is thus

$$\begin{aligned} p_R(kr,\theta,\phi) &= \sum_{\ell=1}^{L} g_\ell e^{j\mathbf{k}_\ell\cdot\mathbf{r}} \\ &= \sum_{\ell=1}^{L} g_\ell \sum_{n=0}^{\infty}\sum_{m=-n}^{n} 4\pi j^n j_n(kr)Y_n^m(\theta_\ell,\phi_\ell)Y_n^m(\theta,\phi)^*. \end{aligned} \tag{3.16}$$

3.1.4 Mode Matching

The goal is to find the loudspeaker gains which minimise $||p_T(kr,\theta,\phi) - p_R(kr,\theta,\phi)||_2^2$ with $||\cdot||_2$ the L^2 norm. For exact reproduction this leads to $p_T(kr,\theta,\phi) = p_R(kr,\theta,\phi)$,

that is so that the reproduced soundfield equals the target soundfield [48]. This corresponds to exact reproduction of the desired soundfield however requires an infinite array of loudspeakers. Later, truncation of the summation to a finite order will be performed to introduce an approximate representation, so that the problem may be formulated as a set of linear equations to solve for the loudspeaker gains. Then the impact of this finite approximation on the different types of solutions will also be discussed. For now, equate the reproduced and target soundfields

$$\begin{aligned}&\sum_{n=0}^{\infty}\sum_{m=-n}^{n} 4\pi j^n j_n(kr) Y_n^m(\theta_T,\phi_T) Y_n^m(\theta,\phi)^* \\ &= \sum_{\ell=1}^{L} g_\ell \sum_{n=0}^{\infty}\sum_{m=-n}^{n} 4\pi j^n j_n(kr) Y_n^m(\theta_\ell,\phi_\ell) Y_n^m(\theta,\phi)^*.\end{aligned} \tag{3.17}$$

Multiply both sides of the equation by a dummy variable, $Y_{n'}^{m'}(\theta,\phi)$, and integrate over the $\mathcal{S}^2$ sphere

$$\begin{aligned}&\int_\Omega \sum_{n=0}^{\infty}\sum_{m=-n}^{n} 4\pi j^n j_n(kr) Y_n^m(\theta_T,\phi_T) Y_n^m(\theta,\phi)^* Y_{n'}^{m'}(\theta,\phi) d\Omega \\ &= \int_\Omega \sum_{\ell=1}^{L} g_\ell \sum_{n=0}^{\infty}\sum_{m=-n}^{n} 4\pi j^n j_n(kr) Y_n^m(\theta_\ell,\phi_\ell) Y_n^m(\theta,\phi)^* Y_{n'}^{m'}(\theta,\phi) d\Omega\end{aligned} \tag{3.18}$$

then make use of the orthogonality of the spherical harmonics (Eqn. 3.4) to give

$$4\pi j^n j_n(kr) Y_n^m(\theta_T,\phi_T) = \sum_{\ell=1}^{L} g_\ell 4\pi j^n j_n(kr) Y_n^m(\theta_\ell,\phi_\ell), \;\; \forall\, n \in \mathbb{N}_0 \text{ and } m \in [-n,n]. \tag{3.19}$$

Finally, removing the common terms from both sides leaves the mode matching equation

$$Y_n^m(\theta_T,\phi_T) = \sum_{\ell=1}^{L} g_\ell Y_n^m(\theta_\ell,\phi_\ell) \;\; \forall\, n \in \mathbb{N}_0 \text{ and } m \in [-n,n]. \tag{3.20}$$

Hence there is a one-to-one relationship between the reproduced and target soundfields for each spherical harmonic mode of order n and degree m. As the mode matching equation is defined for each n, m this forms a set of linear equations to find the required loudspeakers gains for the total of L loudspeakers. However, as the sum over the order is to infinity this is an infinite set of equations for just L unknowns. Therefore, in practice the soundfield representation has to be truncated to a finite

order N. This means that the full set of $\sum_{n=0}^{N}(2n+1) = (N+1)^2$ mode matching equations is defined by

$$Y_n^m(\theta_T, \phi_T) = \sum_{\ell=1}^{L} g_\ell Y_n^m(\theta_\ell, \phi_\ell) \;\; \forall \; n \; \in \; [0, N] \text{ and } m \; \in \; [-n, n]. \tag{3.21}$$

The classic mode matching decoder most generally solves this set of linear equations by the use of the pseudoinverse [48]. The aim is to set $\mathbf{p}_T = \mathbf{p}_R$ with $\mathbf{p}_T$ a vector of length $(N+1)^2$ spherical harmonics up to order N sampled at the target virtual source position and $\mathbf{p}_R$ a similar vector of spherical harmonics however for the reproduced soundfield due to the loudspeaker array. Considering the forward problem that defines the reproduced field, then $\mathbf{p}_R = \mathbf{Y}\mathbf{g}$ with $\mathbf{Y}$ a $(N+1)^2 \times L$ matrix of spherical harmonics up to order N sampled at the L loudspeaker positions and $\mathbf{g}$ a vector of L loudspeaker gains. The loudspeaker gains are then defined through

$$\mathbf{p}_T = \mathbf{p}_R = \mathbf{Y}\mathbf{g} \implies \mathbf{g} = \mathbf{Y}^\dagger \mathbf{p}_T. \tag{3.22}$$

The superscript $(\cdot)^\dagger$ indicates the Moore-Penrose pseudoinverse defined as [119]

$$\mathbf{A}^\dagger = \begin{cases} \mathbf{A}^H(\mathbf{A}\mathbf{A}^H)^{-1} & \text{if } (N+1)^2 < L \\ \mathbf{A}^{-1} & \text{if } (N+1)^2 = L \\ (\mathbf{A}^H\mathbf{A})^{-1}\mathbf{A}^H & \text{if } (N+1)^2 > L \end{cases} \tag{3.23}$$

where $\mathbf{A}$ is an $(N+1)^2 \times L$ matrix and the superscript H denotes the Hermitian transpose.

The number of loudspeakers determines the type of solution [48]. When $L < (N+1)^2$ the problem is overdetermined and an exact solution can not be found. In this case the pseudoinverse gives the least-squares solution that minimises the error between the target and reproduced soundfield, that is

$$\min_{\mathbf{g}} ||\mathbf{p}_T - \mathbf{Y}\mathbf{g}||_2^2. \tag{3.24}$$

With $L \geq (N+1)^2$ an infinite number of exact solutions exists, and the pseudoinverse chooses the minimum L^2 norm solution with respect to $\mathbf{g}$,

$$\min_{\mathbf{g}} ||\mathbf{g}||_2^2 \text{ such that } \mathbf{p}_T = \mathbf{Y}\mathbf{g}. \tag{3.25}$$

Assuming hereon in that $L \geq (N+1)^2$ then using the loudspeaker gains as defined by the mode matching equation will equate to physically reproducing the soundfield about the expansion point, by recreating the spherical harmonic modes of the target soundfield up to the truncation order through the contributions of each loudspeaker to the reproduced modes. As will be discussed in the next section, the effect of order truncation is that above order N the reproduced soundfield modes will not match those of the target soundfield.

Furthermore, we note that the kr dependence of the n-th order term is purely contained in the spherical Bessel function of the n-th degree. For small arguments of kr high order spherical Bessel functions tend to 0. As the argument kr increases, the spherical Bessel functions of successive order become increasingly important. This means that for small kr, truncating the expansion to a low order is sufficient. Equally, for accurate reproduction at higher frequencies or across a large region, a higher order reproduction is required. It is important to note that when $kr > 0$, $j_n(kr) \neq 0$. Thus there is always an error for any step away from the reproduction point so when considering reproduction of the soundfield across a region, it is always accurate *to an error bound*. Due to the spherical nature of the problem, the reproduction region is that of a sphere centered on the origin.

A final consequence of the kr dependence being contained only in the spherical Bessel functions is that no frequency terms are present in the final mode matching equations, and thus the resulting loudspeaker gains. These means the loudspeaker gains are frequency-independent and define simple panning functions. The result of this is the loudspeaker gains are straightforward to implement in a practical reproduction system through simple amplitude panning.

3.1.5 Error Analysis

There exists two fundamental sources of error in the mode matching approach - not considering any practical sources of error from setting up a reproduction system in reality such as misalignment of the loudspeakers, or the effect of room reverberation. The first source of error is due to truncation of the infinite summation with respect to the order, n. As per the Jacobi-Anger expansion in Eqn. 3.12, the plane wave is only correctly recreated at all frequencies and across the whole reproduction region if the infinite sum of all spherical harmonics is correctly reproduced. As discussed, for any practical system this infinite sum must be truncated to a finite order. Therefore, before any reproduction system is even considered there is an error, $\epsilon(kr, \theta, \phi)$, due to truncation found between the original soundfield, $p(kr, \theta, \phi)$ and the target soundfield, $p_T^N(kr, \theta, \phi)$ which is truncated to order N as per the superscript. The truncation error for the 3D problem is thus defined as

$$\epsilon^{3D}_{trunc}(kr,\theta,\phi) = \frac{|p(kr,\theta,\phi) - p^N_T(kr,\theta,\phi)|^2}{|p(kr,\theta,\phi)|^2} = \frac{\left|\sum_{n=N+1}^{\infty}\sum_{m=-n}^{n} 4\pi j^n j_n(kr) Y^m_n(\theta,\phi)^* Y^m_n(\theta_T,\phi_T)\right|^2}{\left|\sum_{n=0}^{\infty}\sum_{m=-n}^{n} 4\pi j^n j_n(kr) Y^m_n(\theta,\phi)^* Y^m_n(\theta_T,\phi_T)\right|^2}. \tag{3.26}$$

This error is the remaining terms of the expansion that are not considered due to the truncation, normalised by the contributions of all the expansion terms (the total energy at point r,θ,ϕ).

A second major source of error is due to spatial aliasing of the loudspeaker array [120]. Practically, this may not be easily separated from the truncation error, as by physically using a loudspeaker array an order truncation must be made. Therefore the total error which includes truncation and spatial aliasing error may be defined as

$$\epsilon^{3D}_{tot}(kr,\theta,\phi) = \frac{|p(kr,\theta,\phi) - p^N_R(kr,\theta,\phi)|^2}{|p(kr,\theta,\phi)|^2} \tag{3.27}$$

and assuming a mode matching solution has been performed such that the loudspeaker gains satisfy Eqn. 3.21, then the total error reduces to

$$\epsilon^{3D}_{tot}(kr,\theta,\phi) = \frac{\left|\sum_{n=N+1}^{\infty}\sum_{m=-n}^{n} 4\pi j^n j_n(kr) Y^m_n(\theta,\phi)^* \left[Y^m_n(\theta_T,\phi_T) - \sum_{\ell=1}^{L} g_\ell Y^m_n(\theta_\ell,\phi_\ell)\right]\right|^2}{\left|\sum_{n=0}^{\infty}\sum_{m=-n}^{n} 4\pi j^n j_n(kr) Y^m_n(\theta,\phi)^* Y^m_n(\theta_T,\phi_T)\right|^2}. \tag{3.28}$$

As seen in Eqn. 3.28, the reproduced soundfield is still defined at modes higher than the truncation order. However, as the mode matching is truncated these modes are not correctly reproduced. Similarly, the summation over order in the numerator is from $N+1$ only as the mode matching ensures that the term $Y^m_n(\theta_T,\phi_T) - \sum_{\ell=1}^{L} g_\ell Y^m_n(\theta_\ell,\phi_\ell) = 0$ for $n \leq N$ with $L \geq N$.

In [48], a rule of thumb is derived for the correct truncation order to reproduce within an error bound to a given value of kr. This considers the error due to truncation only (not including spatial aliasing) but normalised across the whole unit sphere.

Through this derivation, a normalised definition of the total truncation error over the sphere is used and defined as

$$\epsilon^{3D}_{trunc,norm}(kr) = \frac{\int_\Omega |p(kr,\theta,\phi) - p_T^N(kr,\theta,\phi)|^2 d\Omega}{\int_\Omega |p(kr,\theta,\phi)|^2 d\Omega} \tag{3.29}$$

which is shown in [48] to equal

$$\epsilon^{3D}_{trunc,norm}(kr,\theta,\phi) = 1 - \sum_{n=0}^{N} (2n+1)[j_n(kr)]^2 \tag{3.30}$$

that is the normalised total error depends only on kr. This form of error reveals how the mode matching solution requires higher orders for increasing reproduction accuracy with regards to kr. Finally, setting the normalised error to a bound of 4%, equivalent to -14 dB, this gives the important rule of thumb that for reproduction for a given kr, the minimum truncation order required is

$$N = kr \tag{3.31}$$

to reproduce to the -14 dB error limit. This is an extremely useful guide for estimating up to what frequency or radial limit a given reproduction system is valid and will come in use throughout the thesis.

3.2 2D Mode Matching

3.2.1 Fourier Series

The analogous case considering 2D will now be presented. Consider the 2D coordinate system as introduced in Fig 2.1(B). A function $p(kr,\theta)$ may be expressed as a spatial Fourier series [115]

$$p(kr,\theta) = \frac{a_0(kr)}{2} + \sum_{n=1}^{\infty} a_n(kr)\cos(n\theta) + \sum_{n=1}^{\infty} b_n(kr)\sin(n\theta). \tag{3.32}$$

Here $a_0(kr)$, $a_n(kr)$ and $b_n(kr)$ are coefficients found utilising the orthogonality relationships for $\cos(\theta)$ and $\sin(\theta)$, which form an orthogonal basis over the 1-sphere, $\mathcal{S}^1$ (a unit circle). Integration over $\mathcal{S}^1$ is given by $\int_\Omega d\Omega = \int_0^{2\pi} d\theta$. The necessary orthogonality relationships are [121]

$$\begin{aligned}\frac{1}{\pi}\int_{\Omega}\cos(n\theta)\cos(m\theta)d\theta &= \delta_{nm}\\ \frac{1}{\pi}\int_{\Omega}\sin(n\theta)\sin(m\theta)d\theta &= \delta_{nm}\\ \int_{\Omega}\cos(n\theta)\sin(m\theta)d\theta &= 0.\end{aligned} \tag{3.33}$$

Due to the symmetry of the problem, the above holds for any interval $\theta_0 \leq \theta \leq \theta_0 + 2\pi$. Hence the coefficients may be found by [121]

$$\begin{aligned}a_0(kr) &= \frac{1}{2\pi}\int_{\Omega} p(kr,\theta)d\theta\\ a_n(kr) &= \frac{1}{\pi}\int_{\Omega} p(kr,\theta)\cos(n\theta)d\theta\\ b_n(kr) &= \frac{1}{\pi}\int_{\Omega} p(kr,\theta)\sin(n\theta)d\theta.\end{aligned} \tag{3.34}$$

The 2D Jacobi-Anger Expansion

In 2D, a plane wave $p(kr,\theta) = e^{j\mathbf{k}\cdot\mathbf{r}}$ has $\mathbf{k} = [k_x, k_y]^T$, with $k^2 = k_x^2 + k_y^2$. The dot product $\mathbf{k}\cdot\mathbf{r} = k_x x + k_y y = \cos(\Theta)$. With θ_i the arrival direction of the plane wave, considering an arbitrary azimuthal evaluation point θ, then $\Theta = \theta_i - \theta$. The 2D variant of the Jacobi-Anger expansion states [121]

$$\begin{aligned}p(kr,\theta) = e^{j\mathbf{k}\cdot\mathbf{r}} &= e^{jkr\cos(\Theta)}\\ &= \sum_{n=-\infty}^{\infty} j^n J_n(kr)e^{jn\Theta}\end{aligned} \tag{3.35}$$

with $J_n(x)$ the n-th order Bessel Function of the first kind. Making use of the identity $J_{-n}(x) = (-1)^n J_n(x)\ \forall\ n \in \mathbb{Z}$ the expansion may be re-expressed as [112]

$$e^{jkr\cos(\Theta)} = J_0(kr) + 2\sum_{n=1}^{\infty} j^n J_n(kr)\cos(n\Theta). \tag{3.36}$$

3.2.2 Target Soundfield

The target soundfield is again assumed to be that of a plane wave such that $p_T(kr,\theta) = e^{j\mathbf{k}_T\cdot\mathbf{r}}$. The scalar product $\mathbf{k}_T\cdot\mathbf{r} = kr\cos(\Theta)$ with Θ the angle between $\mathbf{k}_T$ and $\mathbf{r}$, which in 2D is simply $\Theta = \theta_T - \theta$, then the plane wave may be expanded using the 2D Jacobi-Anger expansion in Eqn. 3.36

$$p_T(kr,\theta) = e^{jkr\cos(\theta_T-\theta)} = J_0(kr) + 2\sum_{n=1}^{\infty} j^n J_n(kr)\cos(n[\theta_T-\theta]). \tag{3.37}$$

Making use of the trigonometric identity $\cos(A-B) = \cos(A)\cos(B)+\sin(A)\sin(B)$, the expansion takes the form of a Fourier series as per Eqn. 3.32

$$\begin{aligned} p_T(kr,\theta) &= J_0(kr) + 2\sum_{n=1}^{\infty} j^n J_n(kr)\cos(n\theta_T)\cos(n\theta) + 2\sum_{n=1}^{\infty} j^n J_n(kr)\sin(n\theta_T)\sin(n\theta) \\ &= \frac{a_0(kr)}{2} + \sum_{n=1}^{\infty} a_n(kr)\cos(n\theta) + \sum_{n=1}^{\infty} b_n(kr)\sin(n\theta). \end{aligned}$$
$$\text{with } a_0(kr) = 2J_0(kr),\ a_n(kr) = 2j^n J_n(kr)\cos(n\theta_T),\ b_n(kr) = 2j^n J_n(kr)\sin(n\theta_T). \tag{3.38}$$

3.2.3 Reproduced Soundfield

As before, consider a reproduction array of L equidistant plane wave loudspeakers. The ℓ-th loudspeaker is driven by a gain g_ℓ and is situated at position θ_ℓ. The reproduced field, $p_R(kr,\theta)$, due to the loudspeaker array is thus

$$\begin{aligned} p_R(kr,\theta) &= \sum_{\ell=1}^{L} g_\ell e^{j\mathbf{k}_\ell\cdot\mathbf{r}} \\ &= \sum_{\ell=1}^{L} g_\ell J_0(kr) + 2\sum_{\ell=1}^{L} g_\ell \sum_{n=1}^{\infty} j^n J_n(kr)\cos(n\theta_\ell)\cos(n\theta) \\ &\qquad + 2\sum_{\ell=1}^{L} g_\ell \sum_{n=1}^{\infty} j^n J_n(kr)\sin(n\theta_\ell)\sin(n\theta) \end{aligned} \tag{3.39}$$

3.2.4 Mode Matching

Begin by equating the target and reproduced soundfields, $p_T(kr,\theta) = p_R(kr,\theta)$, thus

$$\begin{aligned} &J_0(kr) + 2\sum_{n=1}^{\infty} j^n J_n(kr)\cos(n\theta_T)\cos(n\theta) + 2\sum_{n=1}^{\infty} j^n J_n(kr)\sin(n\theta_T)\sin(n\theta) \\ &= \sum_{\ell=1}^{L} g_\ell J_0(kr) + 2\sum_{\ell=1}^{L} g_\ell \sum_{n=1}^{\infty} j^n J_n(kr)\cos(n\theta_\ell)\cos(n\theta) + 2\sum_{\ell=1}^{L} g_\ell \sum_{n=1}^{\infty} j^n J_n(kr)\sin(n\theta_\ell)\sin(n\theta). \end{aligned} \tag{3.40}$$

As with the 3D mode matching problem, the next step is to multiply by a dummy

variable, integrate in this case over the $\mathcal{S}^1$ unit sphere and make use of the orthogonality relations for $\sin(nx)$ and $\cos(nx)$ expressed in Eqn. 3.33. The dummy variable is defined in turn as $\cos(n'\theta)$ then $\sin(n'\theta)$ to cover all conditions. Equally, the Fourier series could have been expressed in exponential form, in which case a single exponential dummy variable could have been used. Whilst this may compact the mathematics, the trigonometric variant is used in this report to match the literature, for example [45, 47]. Performing this procedure gives a set of equations

$$\begin{aligned} J_0(kr) &= \sum_{\ell=1}^{L} g_\ell J_0(kr) && \text{for } n = 0 \\ \frac{j^n}{\pi} J_n(kr)\cos(n\theta_T) &= \sum_{\ell=1}^{L} g_\ell \frac{j^n}{\pi} J_n(kr)\cos(n\theta_\ell) \;\; \forall\, n \in \mathbb{N} \\ \frac{j^n}{\pi} J_n(kr)\sin(n\theta_T) &= \sum_{\ell=1}^{L} g_\ell \frac{j^n}{\pi} J_n(kr)\sin(n\theta_\ell) \;\; \forall\, n \in \mathbb{N} \end{aligned} \tag{3.41}$$

where we note that the n-th order retains kr dependence in now not a spherical Bessel function, but a standard Bessel function. These functions exhibit a similar dependence on kr as with the spherical Bessel functions, such that for small kr higher order Bessel functions tend to zero. Hence at low frequencies or small distances from the expansion point a low truncation order may be used. However, increasing kr shows higher order truncation is required as the higher order modes contribute more significantly.

Finally, removing the common terms leaves the 2D mode matching equations

$$\begin{aligned} 1 &= \sum_{\ell=1}^{L} g_\ell && \text{for } n = 0 \\ \cos(n\theta_T) &= \sum_{\ell=1}^{L} g_\ell \cos(n\theta_\ell) \;\; \forall\, n \in \mathbb{N} \\ \sin(n\theta_T) &= \sum_{\ell=1}^{L} g_\ell \sin(n\theta_\ell) \;\; \forall\, n \in \mathbb{N} \end{aligned} \tag{3.42}$$

Thus there is a one-to-one relationship between the reproduced and target soundfields for each sine and cosine mode of order n. The above defines a linear set of equations that may be solved as before to find the required loudspeaker gains. Truncating the expansion to a finite order N shows that a minimum of $L = (2N+1)$ loudspeakers are required for exact reproduction of the specified modes. The system of linear equations may be solved using the pseudoinverse as described in Section 3.1.4 except using the 2D basis functions. The 2D formulation reproduces a circle of correct soundfield centred on the reproduction point, to an error bound. As with

the 3D case, the loudspeaker gains are independent of frequency and are therefore simple panning functions.

3.2.5 Error Analysis

As before, the two main sources of error are due to truncation of the infinite expansion to a finite number, and also spatial aliasing due to the loudspeaker array. The 2D truncation error, which considers just the effect of truncating the target soundfield, is formulated as

$$\begin{aligned}\epsilon^{2D}_{trunc}(kr,\theta) =& \frac{|p(kr,\theta) - p^N_T(kr,\theta)|^2}{|p(kr,\theta)|^2} \\ =& \frac{\left|2\sum_{n=N+1}^{\infty} j^n J_n(kr)\cos(n[\theta_T-\theta])\right|^2}{\left|J_0(kr)+2\sum_{n=1}^{\infty} j^n J_n(kr)\cos(n[\theta_T-\theta])\right|^2}.\end{aligned} \tag{3.43}$$

Assume that the loudspeaker gains are defined correctly, such that they satisfy the mode matching equations in Eqn. 3.42 up to order N with $L \geq 2N+1$. Then the total error that includes truncation error and spatial aliasing due to the loudspeaker array is given by

$$\begin{aligned}\epsilon^{2D}_{tot}(kr,\theta) =& \frac{|p(kr,\theta) - p^N_R(kr,\theta)|^2}{|p(kr,\theta)|^2} \\ =& \frac{\left|2\sum_{n=N+1}^{\infty} j^n J_n(kr)\left[\cos(n[\theta_T-\theta]) - \sum_{\ell=1}^{L} g_\ell \cos(n[\theta_\ell-\theta])\right]\right|^2}{\left|J_0(kr)+2\sum_{n=1}^{\infty} j^n J_n(kr)\cos(n[\theta_T-\theta])\right|^2}.\end{aligned} \tag{3.44}$$

The identical rule of thumb from Eqn. 3.31 defining the required truncation order for a given kr limit also holds for the 2D case.

3.3 Chapter Review

This chapter has revised and re-derived the classic HOA mode matching approach in both 3D and 2D. First, the mathematical preliminaries required to described the two orthogonal basis expansions used as soundfield descriptors were presented. In 3D, the soundfield is expressed using the spherical harmonics whilst in 2D a spatial Fourier series is used.

A target soundfield consisting of a single virtual plane wave was considered and expanded using the relevant basis. A reproduction array of radially equidistant plane wave loudspeakers was assumed. The mode matching approach aims to reproduce the target soundfield by reproducing exactly the decomposition of the target field on to each basis function, up to order N. A minimum of $(N+1)^2$ loudspeakers are required to ensure exact reproduction to order N. Under these assumptions the loudspeaker gain solutions are simple panning laws dependant on the virtual source position. The loudspeaker gains are defined by formulating the mode matching equations as a set of linear equations. This set of equations is solved by using the pseudoinverse of a matrix of basis functions sampled at the loudspeaker positions.

Chapter 4

Higher Order Stereophony

The Taylor expansion is an approach for approximating a function about a point, by considering the weighted summation of the function's derivatives evaluated at the expansion position. The Taylor expansion is therefore a basis expansion with respect to the functions derivatives. It is most commonly considered in 1D only, however is also valid for higher dimensional representations [121]. The Taylor expansion has found little use in soundfield reproduction literature, mainly with the work by Dickins at the turn of the century [122, 123]. Here, both the full 3D Taylor expansion and the spherical harmonic expansion of a soundfield were considered. When restricting to physical soundfields that satisfy the wave equation and considering that both descriptions contain the same amount of information, the Taylor expansion was found to be over-specified resulting in a greater number of terms on truncation to the N-th order when compared to the spherical harmonic expansion. Interestingly, this is true only when $n \geq 2$, because the zeroth and first order for both expansions contain equal number of terms. It was therefore concluded that the spherical harmonic expansion was more compact and thus the most convenient soundfield descriptor when considering soundfields in 3D. This has likely led to many preferring to use the spherical harmonic expansion over the Taylor expansion, for example as noted in the seminal work by Poletti [43].

Relying on the fact that the spherical harmonics are actually most commonly expressed in Cartesian coordinates, mappings between spherical harmonic directivities and soundfield derivatives (originally presented by Dickins [122]) have been used to create efficient Finite Difference Time Domain methods where the evaluation grids are expressed in Cartesian coordinates, but spherical directivities may be considered [124, 125]. Such derivative operators have also been used in the fast multipole method [126]. This leads to a representation of the spherical harmonic coefficients of the soundfield as a series of derivatives evaluated at a given point in space.

These mappings have also been utilised in the microphone array literature. Cotterell presented the relations up to second order between spherical harmonics and

soundfield derivatives in the context of Ambisonics [127]. Relating back to 2D Ambisonics, Kolundzija considered the design of differential microphone arrays up to third order along with the transformation from the soundfield derivatives to the circular harmonic coefficients [128, 129]. Interestingly, this transformation made use of the Chebyshev polynomials, which will also be used later in this work. This leads to a relevant field of literature regarding differential microphone arrays, which are designed to measurement the gradients of a soundfield based on using the Taylor expansion representation. A thorough analysis of such microphone arrays including derivations of directivity patterns and advantages and limitations is given by Elko in [130, 131]. Further work by Cray has also used such directional microphone arrays to measure the gradients of a soundfield at a point to the N-th order [132].

Thus, the literature shows that the Taylor expansion is a rarely used representation of a soundfield. This is likely due to its over-specification, requiring a larger number of terms to a given order truncation than the spherical harmonic expansion which is more compact. Despite this in certain situations considering derivatives of the soundfield may be more natural, for example in capturing a soundfield. However in the literature this is often an intermediate step towards a more common basis representation such as the spherical harmonics. Regarding soundfield reproduction, the Taylor expansion has been rarely explored.

This chapter introduces a new novel technique named Higher Order Stereophony (HOS), a technique for reproducing a soundfield accurately across a single line through the use of the single variable Taylor expansion. The goal is to develop a more efficient soundfield reproduction approach, that only recreates the soundfield correctly at the listener's ears. The technique aims to do this by soundfield reproduction along the interaural axis assuming that this will lead to the correct binaural signals. Therefore, the soundfield is reproduced accurately across a single line with respect to kr, with k the wavenumber and r the distance step away from the head centre. This means reproduction of the soundfield at a fixed distance along the line is accurate for all frequencies, or conversely when fixing the frequency then reproduction is accurate for all distance steps along the line from the reproduction centre. Each of these scenarios may be thought of as a reproduced line with respect to frequency or distance. This is achieved using an amplitude panning solution and increasing the number of loudspeakers increases the upper limit of accurate reproduction, with respect to kr. This differs to existing approaches such as CTC, which aims to reproduce all frequencies at two fixed positions. The CTC solution results in frequency dependent inverse filtering, whilst increasing the number of loudspeakers only increases the robustness of the system. The advantage over other soundfield reproduction approaches such as HOA is that less loudspeakers and less stringent requirements on their positioning is required.

The chapter is arranged as follows. First, the theory of the technique is presented, resulting in a set of order matching equations (analogous to mode matching) that

define the necessary loudspeaker gains for any given loudspeaker array. Next, the classic stereo sine law is derived through the new HOS framework and it is demonstrated that the technique generalises the traditional stereo technique to higher orders and any number of loudspeakers. Next, the link between Higher Order Ambisonics (HOA) and HOS is explored and decoders are derived to transition between the two representations. Finally, validation of the approach is presented through numerical soundfield simulations utilising experimental measurements of various microphone arrays.

4.1 The Taylor Expansion

4.1.1 The Single Variable Taylor Expansion

The Taylor expansion expresses a well-behaved function about an expansion point as a infinite summation of its derivatives evaluated at the expansion point. In 1D, if $p(x)$ is an infinitely differentiable function at a point x_0, then the function may be written as [121]

$$p(x) = \sum_{n=0}^{\infty} \frac{(x - x_0)^n}{n!} \frac{d^n p(x_0)}{dx^n}. \tag{4.1}$$

Hence the function may be approximated about the point x_0 by calculating the weighted sum of its derivatives at x_0. The n-th order term depends on the n-th derivative. Practically, the infinite summation must be truncated to a finite order N introducing an error[1] into the representation. In this case, increasing the order N results in a better approximation for larger arguments of $(x - x_0)$ and thus the approximation's accuracy further away from the expansion point x_0.

4.1.2 The Multi-Variable Taylor Expansion

Whilst the Taylor expansion is often used for single variable functions, it may also be expressed for multiple variable functions. Let the function $p(\mathbf{r})$ be an infinitely differentiable function at a point $\mathbf{r}_0$ where $\mathbf{r} = [x, y, z]^T$. Then the multi-variable Taylor expansion is given by [121]

$$p(\mathbf{r}) = \sum_{n=0}^{\infty} \frac{[(\mathbf{r} - \mathbf{r}_0) \cdot \nabla]^n}{n!} p(\mathbf{r}_0). \tag{4.2}$$

Therefore the Taylor expansion is also valid for the 3D scenario. Equally, the 2D case is formulated simply by negating the z terms. As before the expansion must be truncated to a finite order N, and increasing the order of the truncation results in a more accurate expansion for larger values of $(\mathbf{r} - \mathbf{r}_0)$. However, there is increased

[1] In the literature, the residual terms not included due to the truncation are also named the *remainder*.

complexity when considering the derivatives of the function. Define the evaluation step of the expansion to be a vector of length a and direction $\hat{\mathbf{n}}$ such that $(\mathbf{r} - \mathbf{r}_0) = a\hat{\mathbf{n}}$. Then the product in the numerator becomes

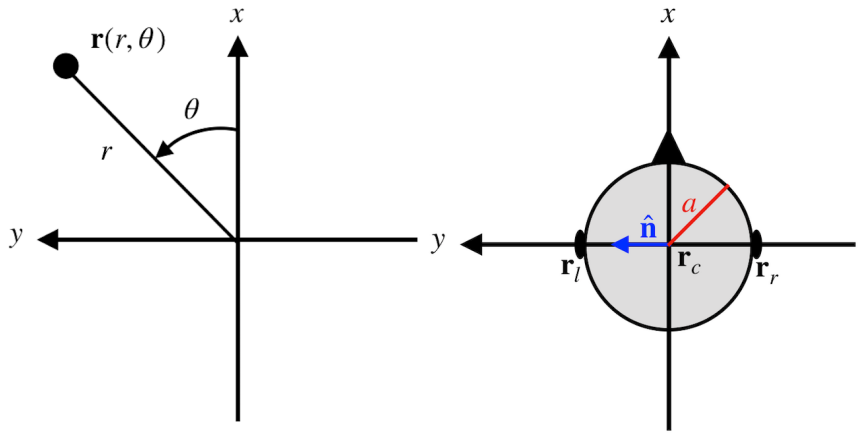
$$\begin{aligned}[(\mathbf{r} - \mathbf{r}_0) \cdot \nabla]^n &= [a\hat{\mathbf{n}} \cdot \nabla]^n \\ &= a^n \left[n_x \frac{\partial}{\partial x} + n_y \frac{\partial}{\partial y} + n_z \frac{\partial}{\partial z} \right]^n .\end{aligned} \tag{4.3}$$

Beyond $n = 1$ this product results in many cross-derivatives which complicate the evaluation. However, if the expansion is considered along one axis only, that is $\hat{\mathbf{n}} = [1, 0, 0]^T$, $[0, 1, 0]^T$ or $[0, 0, 1]^T$ then the multi-variable Taylor expansion collapses to the single variable expansion as expected, as only derivatives with respect to a single axis are required. Notably, as all that matters is the derivatives in direction $\hat{\mathbf{n}}$ the coordinate system can always be rotated such that $\hat{\mathbf{n}}$ defines one axis and the single variable expansion can simply be used.

4.1.3 Expansion Of A Plane Wave Soundfield

The goal of a 3D audio reproduction system is to reproduce a given soundfield by recreating the correct binaural signals at the listener's ears. The HOS approach is to suggest that this can be achieved, under certain conditions and assumptions, by accurately reproducing the soundfield along the listener's interaural axis only. The interaural axis is that which the listener's ears lie upon. This differs to other soundfield reproduction methods, such as HOA, that aim to reproduce the soundfield accurately over a region of space, not just a single axis. This will be shown to be preferable to HOA as it leads to less stringent requirements on the number of audio channels, loudspeakers and the loudspeaker positions. The approach is preferable to alternatives such as CTC which consider the soundfield at the two ear positions only, as these techniques lead to more complicated frequency-dependent loudspeaker filters as opposed to simple frequency-independent panning gains. Furthermore, CTC makes explicit assumptions about the listener's HRTF which are not required when using soundfield reproduction approaches.

Hereon in the analysis is simplified to 2D by considering the horizontal plane only, using a coordinate system as per Fig. 4.1(A). As seen above the Taylor expansion can be utilised for the full 3D case however this will be introduced later, as the 2D scenario leads to symmetry which greatly simplifies the 3D case. Consider a listener with their head centred at the origin as in Fig. 4.1(B). Let $\hat{\mathbf{n}}$ be the unitary vector pointing from the head centre, $\mathbf{r}_c$, to the left ear, $\mathbf{r}_l$, thus defining the interaural axis. Assuming the listener's ears are diametrically opposed across the head and that the head radius is given by a, the two ear positions are $\mathbf{r}_{l,r} = \pm a\hat{\mathbf{n}}$. It is important to note that while the context is considering reproduction across the listener's interaural



(A) 2D coordinate system.

(B) Example head orientation. For this specific head orientation, the interaural axis is aligned along the y axis.

FIGURE 4.1: Coordinate system and geometry layout.

axis, at the moment in the mathematics no complex HRTF is included, a common assumption when deriving spatial audio panning laws. As the acoustical effects of the HRTF will dictate the performance of the technique, later the mathematics will be extended to include HRTFs and their effect evaluated. It may also be noted that a valid low frequency approximation of the plane wave rigid sphere HRTF model is a shadowless head model with enlarged head radius and therefore may be modelled as two points in free field [133], which suggests at low frequencies for plane wave sources the HRTF has minimal effect. The consequence for now is that the listener's head orientation is purely used to define the position of the reproduction line and free field conditions are assumed. For simplicity fix the listener's head orientation such that the ears are always aligned along the y axis, which means $\hat{\mathbf{n}} = \hat{\mathbf{y}}$.

The incident sound source is assumed to act as a plane wave. This is a good approximation when the distance from the source to the evaluation point is considerably larger than the wavelength in question. Plane wave sources are a common assumption in the literature, and form the basis for deriving the stereo sine and stereo tangent law as well as HOA mode matching, the validity of this approximation will be further discussed later. The soundfield due to a plane wave incident with $\mathbf{k}_i = k[\cos(\theta_i), \sin(\theta_i)]^T$ measured at $\mathbf{r} = [x, y]^T$ is

$$\begin{aligned} p(\mathbf{r}) &= e^{j\mathbf{k}_i \cdot \mathbf{r}} \\ &= e^{jk[x\cos(\theta_i) + y\sin(\theta_i)]}. \end{aligned} \tag{4.4}$$

Next, the Taylor expansion will be used to expand the soundfield about the centre of the listener's head, along the interaural axis. Consider the multi-variable Taylor expansion in Eqn. 4.2. The interaural axis to be expanded along is defined by $\hat{\mathbf{n}}$ as in Eqn. 4.3. As here the listener's head is aligned along the y axis ($\hat{\mathbf{n}} = \hat{\mathbf{y}}$), the

expansion collapses to the single variable expansion as per Eqn. 4.1. The expansion is about the head centre $\mathbf{r}_c = [x_c, y_c]^T$. The n-th order term of the Taylor expansion is dependent on the n-th order derivative of the function, thus for a plane wave source the derivatives with respect to y are

$$\begin{aligned} \frac{\partial}{\partial y} p(\mathbf{r}_c) &= jk \sin(\theta_i) p(\mathbf{r}_c) \\ \frac{\partial^2}{\partial y^2} p(\mathbf{r}_c) &= [jk \sin(\theta_i)]^2 p(\mathbf{r}_c) \\ &\vdots \\ \frac{\partial^n}{\partial y^n} p(\mathbf{r}_c) &= [jk \sin(\theta_i)]^n p(\mathbf{r}_c) \end{aligned} \tag{4.5}$$

hence the Taylor expansion of the plane wave along the y axis with step size $(y - y_c)$, which reduces to being a function dependent on y only, is

$$p(y) = \sum_{n=0}^{\infty} \frac{[jk(y - y_c)\sin(\theta_i)]^n}{n!} p(\mathbf{r}_c). \tag{4.6}$$

Finally, apply the expansion to be to the listener's two ears as per the head orientation and definition in Fig. 4.1(B), so that the step size is simply the head radius, $(y - y_c) = a$. Let the head be centred at the origin. Setting the plane wave to have unitary amplitude at the centre of the head implies $p(\mathbf{r}_c) = 1$. Therefore the pressure at the listener's ear positions, $\mathbf{r}_{l,r} = \pm a\hat{\mathbf{y}}$, is given by

$$p(\pm a) = \sum_{n=0}^{\infty} \frac{[jk(\pm a)\sin(\theta_i)]^n}{n!}. \tag{4.7}$$

This formulation is the result of an expansion of the plane wave soundfield along a line using the Taylor series as the expansion basis. The n-th order term is defined by the n-th order derivative of the soundfield, which for the simple case of a plane wave and expansion along the y axis results in sine terms to the power of n. Interestingly, defining the expansion to be across the x axis results in a similar representation except considering cosine terms to the power of n. Both approaches are equally valid, and can be transformed between by applying a rotation of the reference system. The cosine formulation is given by

$$p(\pm a) = \sum_{n=0}^{\infty} \frac{[jk(\pm a)\cos(\theta_i)]^n}{n!}. \tag{4.8}$$

The spatial/frequency quantity is ka, relating to the expansion step a from the origin along the line. For the context of reproduction of binaural signals for a listener,

the derivation has considered the geometry of a listener's head including a head radius a and interaural axis $\hat{\mathbf{n}}$. However, the approach remains generalised to the expansion across a line utilising any finite step size or direction from the expansion centre. That is we are considering the soundfield across a line, motivated by the interaural axis. To accurately represent the soundfield to a higher value of ka, higher order terms of the expansion are required. On truncation of the series this fixes the ka value to which accurate representation can be achieved. Here the soundfield may be considered as a line in frequency (k) or spatially (a) with the ka value setting a bound on accurate reproduction of the soundfield in either domain.

The expansion terms are not strictly modes as they do not necessarily form an orthogonal basis over the expansion space. Thus from the presented Taylor expansion of a plane wave, the HOS *order* matching (as opposed to 'mode matching') equations will now be derived.

4.2 Higher Order Stereophony Order Matching

4.2.1 Target Soundfield

The target soundfield is that which the loudspeaker array aims to reproduce, leading to the definition of a specific set of loudspeaker gains. The target soundfield, $p_T(ka)$, is simply a plane wave as defined in Eqn. 4.7:

$$p_T(ka) = \sum_{n=0}^{\infty} \frac{[jka\sin(\theta_T)]^n}{n!}. \tag{4.9}$$

A target of a single plane wave is considered for the derivation. The solution for a soundfield consisting of a summation of plane waves comprises of a linear superposition of the individual plane wave contributions. Furthermore, most soundfields can be represented as a summation of plane waves, considering the plane wave density representation [134]. This reasoning follows that of HOA mode matching derivations.

4.2.2 Reproduced Soundfield

The reproduced soundfield is that due to the summed contributions of each individual loudspeaker in the reproduction array. Consider an array of L loudspeakers radially equidistant to the origin, which all act as plane waves. If the loudspeakers are not equidistant, a delay may be applied to them such that they are acoustically equidistant. The ℓ-th loudspeaker is situated at an angle θ_ℓ and driven by a gain g_ℓ. Considering Eqn. 4.7, the reproduced soundfield along the y axis is given by

$$p_R(ka) = \sum_{\ell=1}^{L} g_\ell \sum_{n=0}^{\infty} \frac{[jka\sin(\theta_\ell)]^n}{n!}. \tag{4.10}$$

4.2.3 Order Matching

The goal is to find the loudspeaker gains which minimise $||p_T(ka) - p_R(ka)||_2^2$. For exact reproduction this leads to the condition $p_T(ka) = p_R(ka)$ which requires that the number of loudspeakers must be infinite. Later, the effects of truncation will be considered.

$$\sum_{n=0}^{\infty} \frac{[jka\sin(\theta_T)]^n}{n!} = \sum_{\ell=1}^{L} g_\ell \sum_{n=0}^{\infty} \frac{[jka\sin(\theta_\ell)]^n}{n!}. \tag{4.11}$$

Apply the order matching principle, such that the terms of the two expansions are matched for each order n. Traditionally an orthogonality condition is applied to lead to the matching of each order. However, directly equating the n-th orders will still result in the correct overall summation even though it may not be the only possible solution. Therefore the order matching requirement is that

$$\frac{[jka\sin(\theta_T)]^n}{n!} = \sum_{\ell=1}^{L} g_\ell \frac{[jka\sin(\theta_\ell)]^n}{n!} \;\; \forall\, n \;\in\; \mathbb{N}_0. \tag{4.12}$$

This reveals that the ka dependence of the n-th term is given by $(ka)^n$. Thus for small ka, only low order terms are required. Increasing the value of ka leads to higher order terms becoming significant. Removing all remaining common terms gives the HOS order matching equation

$$\sin^n(\theta_T) = \sum_{\ell=1}^{L} g_\ell \sin^n(\theta_\ell) \;\; \forall\, n \;\in\; \mathbb{N}_0. \tag{4.13}$$

This means order matching with respect to powers of $\sin(x)$ leads to an accurate reproduction of the soundfield along a line which in this case is the y axis. Furthermore, truncating the expansion to a finite order N, termed N-th order stereo, only requires $L \geq N+1$ loudspeakers, less than the HOA approach. Importantly, as there is no frequency dependence in the order matching equations, the loudspeaker gains are real-valued and define simple amplitude panning laws.

4.2.4 Loudspeaker Gain Definitions

To formulate the set of linear equations to find the loudspeaker gains, assume truncation to the N-th order. Let $\mathbf{p}_T$ be a length $(N+1)$ vector of target signals, Ψ be an

$(N+1) \times L$ plant matrix[2] and $\mathbf{g}$ be a length L vector of loudspeaker gains. To define the loudspeaker gains an inverse problem is formulated:

$$\begin{aligned}
\mathbf{p}_T = \Psi \mathbf{g} \implies \mathbf{g} = \Psi^{\dagger} \mathbf{p}_T \\
\text{where } \mathbf{p}_T = \begin{bmatrix} 1 \\ \sin(\theta_T) \\ \sin^2(\theta_T) \\ \vdots \\ \sin^N(\theta_T) \end{bmatrix} \\
\Psi = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ \sin(\theta_1) & \sin(\theta_2) & \sin(\theta_3) & \dots & \sin(\theta_L) \\ \sin^2(\theta_1) & \sin^2(\theta_2) & \sin^2(\theta_3) & \dots & \sin^2(\theta_L) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sin^N(\theta_1) & \sin^N(\theta_2) & \sin^N(\theta_3) & \dots & \sin^N(\theta_L) \end{bmatrix} \\
\mathbf{g} = \begin{bmatrix} g_1 \\ g_2 \\ g_3 \\ \vdots \\ g_L \end{bmatrix}.
\end{aligned} \tag{4.14}$$

The plant matrix formulates the contribution of each loudspeaker to each order, which is dictated by the angular position of the loudspeakers. The superscript $(\cdot)^{\dagger}$ indicates the Moore-Penrose pseudoinverse as described in Section 3.1.4. When $(N+1) < L$ the problem is overdetermined, an exact solution cannot be found and the pseudoinverse gives the least-squares solution that minimises the error between the target and reproduced soundfield. When $(N+1) \leq L$ an infinite number of exact solutions exist, and the pseudoinverse chooses the minimum norm solution with respect to the L^2 norm. Practically the solution may be calculated quickly numerically as there is no frequency dependence, or even by hand for low orders.

So far, only the sine representation from expansion over the y axis has been considered. As explained in Section 4.1.3 considering reproduction across the x axis leads to a similar style solution except using a cosine formulation, for example

[2] It is important to note that here the plant matrix is frequency independent and contains trigonometric terms dependent on the loudspeaker angular positions only. This differs from some uses in the soundfield reproduction literature where the plant matrix is a frequency dependent matrix containing the transfer functions from a loudspeaker array to a set of microphone positions.

$$\mathbf{p}_T = \begin{bmatrix} 1 \\ \cos(\theta_T) \\ \cos^2(\theta_T) \\ \vdots \\ \cos^N(\theta_T) \end{bmatrix}$$
$$\Psi = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ \cos(\theta_1) & \cos(\theta_2) & \cos(\theta_3) & \dots & \cos(\theta_L) \\ \cos^2(\theta_1) & \cos^2(\theta_2) & \cos^2(\theta_3) & \dots & \cos^2(\theta_L) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \cos^N(\theta_1) & \cos^N(\theta_2) & \cos^N(\theta_3) & \dots & \cos^N(\theta_L) \end{bmatrix} \tag{4.15}$$

where the set of resulting loudspeaker gains now leads to reproduction across the x axis, not the y axis.

This derivation has used the approximation of sources and loudspeakers which act as plane waves. In practise, realistic sources (particularly a loudspeaker) may better modelled as a monopole source accounting for the spherical spreading due to the wave propagation. Utilising a monopole model for the derivation (and therefore allowing for placement of the source with respect to radial distance) is left for future work, noting that this extension exists for HOA in a similar manner [53] and is expected to be applicable to HOS. However, this inclusion will likely lead to frequency dependent loudspeaker filters instead of panning gain as with the HOA case (note that if the virtual source is positioned at the same radius as the loudspeaker array, in HOA the filters collapse back to frequency independent panning gains). A key justification to use the plane wave model other than mathematical simplicity is that a plane wave is a far field approximation of a monopole source, however formally defining the exact far field region is not a simple task, and it is often considered when $kr \gg 1$ [135]. For illustrative purposes, setting $kr = 10$ for a radius of 3 metres leads to a frequency of 182 Hz, which might be considered a reasonable regime to assume to be in the monopole's far field. This suggests for large radius loudspeaker arrays the plane wave approximation might be suitable.

4.2.5 The Instability Condition

The contribution of each loudspeaker is governed by the sine or cosine of its angular position. Each loudspeaker must therefore be uniquely positioned to ensure they contribute an additional degree of freedom to the system. Consider an N-th order system utilising the minimum required number of $N + 1$ loudspeakers. For a pair of loudspeakers i and j situated at angles θ_i, θ_j respectively, the scenario when $\sin(\theta_i) = \sin(\theta_j)$ results in both loudspeakers contributing to the reproduction axis identically and thus the system views both as identical loudspeakers. That is only N degrees of

freedom are available and an exact solution can no longer be achieved. This scenario should be avoided when designing the loudspeaker array to ensure the minimum number of loudspeakers are used. This issues arises from the cone of confusion[3], as the soundfield recreated by the loudspeaker across a single axis only is considered. This means for a given loudspeaker at θ_i then $\theta_j \neq \pi - \theta_i$.

In this limit where $\theta_j \to \pi - \theta_i$, the plant matrix can become ill-conditioned leading to large loudspeaker gain definitions. To overcome the ill-conditioning, an additional loudspeaker can be added at a more appropriate angular position. Alternatively, a practical method to combat the large loudspeaker gains whilst retaining the use of only $N+1$ loudspeakers is to employ Tikhonov regularisation when inverting the plant matrix [54], such that

$$\mathbf{A}^{\dagger}_{reg} = (\mathbf{A}^H\mathbf{A} + \beta\mathbf{I}_L)^{-1}\mathbf{A}^H \tag{4.16}$$

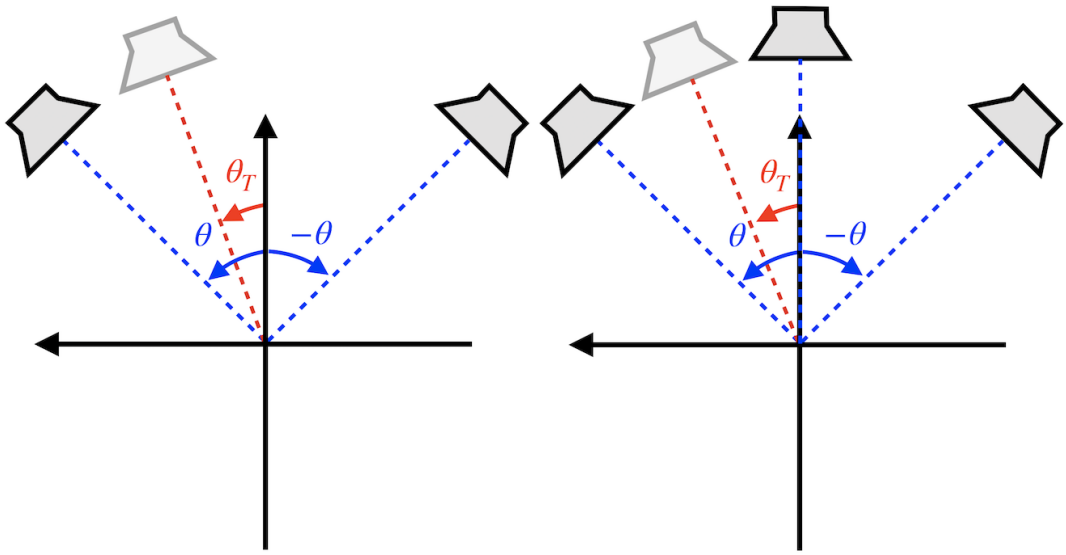
with the regularisation parameter β a non-negative real scalar and $\mathbf{I}_x$ a size $x \times x$ identity matrix. When using Tikhonov regularisation the left and right formulations of the pseudoinverse are identical, as proven in Appendix D. This approach seeks the solution that minimises both the error between the reproduced and the target signals as well as the energy of the loudspeaker gains, weighted by the regularisation parameter. Practically this will apply a limit to the loudspeaker gains however at the cost of introducing further error into the solution. However, if the problem tends to the overdetermined scenario the exact solution already cannot be found. Therefore, this approach provides a simple practical solution to stabilise the loudspeaker gains, without adding more loudspeakers but at the cost of allowing errors in the solution.

4.3 Example Higher Order Stereophony Systems

HOS may be used to derive loudspeaker gains for a range of existing loudspeaker arrays. In this section the HOS gains will be derived analytically for a classic stereo loudspeaker setup, revealing the link of the technique to existing stereo systems. This motivates the naming of the technique as *Higher Order Stereophony*, where it is the generalisation of the stereo theory. Next, a specific second order system will also be considered analytically.

To find the loudspeaker gains, it is a case of using Eqn. 4.14 to create the plant matrix, Ψ, and target signals, $\mathbf{p}_T$, for that given loudspeaker rig. Then the loudspeaker gains, $\mathbf{g}$ are found using the pseudoinverse such that $\mathbf{g} = \Psi^{\dagger}\mathbf{p}_T$. For low orders, the pseudoinverse of the plant matrix may be computed by hand to give an analytical expression for the loudspeaker gains.

[3] In 2D this is not a cone but two points reflected across the reproduction axis, however the cone of confusion is used as it remains the more familiar and intuitive terminology.



(A) A standard symmetric first order stereo loudspeaker arrangement.

(B) A second order stereo LCR arrangement.

FIGURE 4.2: First and second order example loudspeaker arrays.

4.3.1 First Order Stereo

To begin with, consider HOS performed to just the first order. The minimum number of loudspeakers required is $L = N + 1 = 2$. As shown in Fig. 4.2(A), let the two loudspeakers be positioned as a standard stereo pair at $\pm\theta$, with the aim to reproduce the virtual source positioned at θ_T. In this case the target pressure vector and plant matrix are

$$\begin{aligned} \mathbf{P}_T &= \begin{bmatrix} 1 \\ \sin(\theta_T) \end{bmatrix} \\ \Psi &= \begin{bmatrix} 1 & 1 \\ \sin(\theta) & \sin(-\theta) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \sin(\theta) & -\sin(\theta) \end{bmatrix} \\ \Psi^{+} &= \frac{-1}{2\sin(\theta)} \begin{bmatrix} -\sin(\theta) & -1 \\ -\sin(\theta) & 1 \end{bmatrix} \end{aligned} \tag{4.17}$$

where the identity $\sin(-x) = -\sin(x)$ has been used. This leads to the loudspeaker gain definitions

$$\mathbf{g} = \frac{-1}{2\sin(\theta)} \begin{bmatrix} -\sin(\theta) - \sin(\theta_T) \\ -\sin(\theta) + \sin(\theta_T) \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 + \frac{\sin(\theta_T)}{\sin(\theta)} \\ 1 - \frac{\sin(\theta_T)}{\sin(\theta)} \end{bmatrix}. \tag{4.18}$$

This is traditional stereo sine law as defined in [7, 8]. Hence by using the Taylor expansion, the classic stereo sine law has been derived and is a first order Taylor

approximation of reproducing the actual plane wave target soundfield across a line, with the assumption the reproduction line is that of the interaural axis. The stereo sine law is defined as a low frequency approach, HOS therefore both generalises and expands the stereo theory to any given order, for any given loudspeaker array and reproduction across any frequency or spatial range (as restricted by the loudspeaker array and truncation order).

4.3.2 Second Order Stereo

With the link between classic stereo and HOS established, it is interesting to now consider a second order system. A logical step would be to consider the loudspeaker gains for a standard LCR (left-centre-right) loudspeaker setup [136]. This loudspeaker arrangement is shown in Fig. 4.2(B), the LR loudspeakers are a standard symmetric stereo pair, whilst the C loudspeaker is centred in front of the listener. Thus $\theta_1 = \theta_L = \theta$, $\theta_2 = \theta_C = 0$ and $\theta_3 = -\theta_L = -\theta$. This is the frontal half of a standard surround sound system (for example a 5.1, 7.1 or 5.1.2 system).

Traditionally, the LR pair is used as a stereo arrangement whilst the C loudspeaker is a separate channel purely defined to reinforce the vocals and other important events in the audio mix [137]. However, using the second order stereo gain definitions, all three loudspeakers may be used for the spatial content instead of just the LR pair. For this setup with $N = 2$ the target pressure vector and plant matrix may be written as

$$\begin{aligned}
\mathbf{P}_T &= \begin{bmatrix} 1 \\ \sin(\theta_T) \\ \sin^2(\theta_T) \end{bmatrix} \\
\Psi &= \begin{bmatrix} 1 & 1 & 1 \\ \sin(\theta) & 0 & \sin(-\theta) \\ \sin^2(\theta) & 0 & \sin^2(-\theta) \end{bmatrix} \\
\Psi^{\dagger} &= \begin{bmatrix} 0 & \frac{1}{2\sin(\theta)} & \frac{1}{2\sin^2(\theta)} \\ 1 & 0 & -\frac{1}{\sin^2(\theta)} \\ 0 & -\frac{1}{2\sin(\theta)} & \frac{1}{2\sin^2(\theta)} \end{bmatrix}
\end{aligned} \tag{4.19}$$

thus the loudspeaker gains are given by

$$\mathbf{g} = \frac{1}{2} \begin{bmatrix} \frac{\sin(\theta_T)}{\sin(\theta)} + \left(\frac{\sin(\theta_T)}{\sin(\theta)}\right)^2 \\ 2 - 2\left(\frac{\sin(\theta_T)}{\sin(\theta)}\right)^2 \\ -\frac{\sin(\theta_T)}{\sin(\theta)} + \left(\frac{\sin(\theta_T)}{\sin(\theta)}\right)^2 \end{bmatrix}. \tag{4.20}$$

This specific scenario is interesting due to how each loudspeaker contributes to each order of the reproduction. For this specific setup, the centre loudspeaker fully controls the zeroth order (the pressure at the head centre). Next, the LR pair fully recreate the first order contributions, those given by the sine terms to the power of 1. For the first order contributions each of the LR pair have equal magnitude but opposite phase. Finally, the second order terms, those that are sine squared, are controlled by all three loudspeakers, however the LR pair works at equal magnitude in phase whilst the C loudspeaker requires a magnitude that equals the sum of the LR contributions, working in opposite phase to the LR loudspeakers.

This second order system is demonstrated because it is a readily available loudspeaker arrangement, used throughout the audio industry. Thus the HOS technique could be easily implemented without any new major loudspeaker arrangements needing to be adopted. This second order system, through reproducing one more order of the Taylor expansion, will recreate the soundfield within an error bound along the reproduction line to a higher ka value, thus expanding the stereo technique to a higher frequency limit.

4.4 Relation To Higher Order Ambisonics

4.4.1 Transformations Between Soundfield Representations

Upon inspection HOS is similar in nature to HOA. First, they are both soundfield reproduction methods of sorts. HOA in 2D and 3D reproduces the correct soundfield within a circle or sphere respectively, whilst HOS is correct reproduction across a line. Both techniques utilise a mathematical soundfield representation to a given order, then reproduction of said expansion by matching order terms using a loudspeaker array. Increasing the truncation order of the expansion increases its validity with respect to both frequency and distance from the expansion centre. All techniques are derived using similar assumptions (primarily plane wave virtual sources and loudspeakers), and end up as defining loudspeaker gains which are panning functions and thus frequency-independent. Finally, to a first order approximation HOS has been shown to be a subset of HOA, and the soundfield representations are intrinsically linked [82, 122, 123, 138]. Thus it is obvious the techniques are fundamentally linked, particularly to first order.

Consider the system equation from Eqn. 4.14. This states that the target pressure vector, $\mathbf{p}_T$, is the product of the plant matrix, Ψ times the loudspeaker gains vector, $\mathbf{g}$; that is $\mathbf{p}_T = \Psi\mathbf{g}$. This equation holds regardless of the expansion used to express the soundfield, as long as the same representation is used to define all entries for $\mathbf{p}_T$, Ψ and $\mathbf{g}$. Truncation to an order N is assumed.

Let the superscript $(\cdot)'$ indicate truncation to the same order N but using a different soundfield representation, so that $\mathbf{p}'_T = \Psi'\mathbf{g}'$. Assume an order-limited mapping between the two soundfield expansions; that is under truncation to order N the mapping exists for all terms, whilst the set of basis functions used for the two representations both span the same space. The target pressures and plant matrices may be written as

$$\begin{aligned} \mathbf{p}'_T &= \mathbf{A}\mathbf{p}_T \\ \Psi' &= \mathbf{A}\Psi \end{aligned} \tag{4.21}$$

where $\mathbf{A}$ is a matrix that expresses the transformation between the two representations. For the underdetermined case when $L \geq (N+1)$, the pseudoinverse of Ψ is used to define the gains which are a minimum norm solution. Therefore

$$\begin{aligned} \mathbf{g} &= \Psi^{\dagger}\mathbf{p}_T \\ \mathbf{g}' &= \Psi'^{\dagger}\mathbf{p}'_T \\ &= (\mathbf{A}\Psi)^{\dagger}\mathbf{A}\mathbf{p}_T \\ &= \Psi^{\dagger}\mathbf{A}^{-1}\mathbf{A}\mathbf{p}_T \\ &= \Psi^{\dagger}\mathbf{p}_T \\ &= \mathbf{g}. \end{aligned} \tag{4.22}$$

Here the identify $(\mathbf{AB})^{\dagger} = \mathbf{B}^{\dagger}\mathbf{A}^{\dagger}$ has been used, as well as noting that $\mathbf{A}$ is a square matrix and therefore the psueodinverse equals the standard matrix inverse, leading to $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. The above holds for the underdetermined case as the loudspeaker gains are a minimum norm solution, which means the solution $\mathbf{g}$ to $\mathbf{p}_T = \Psi\mathbf{g}$ has zero projection onto the null-space of Ψ. In the overdetermined scenario the solution may have non-trivial elements that map to the null-space of Ψ, in this case

$$\begin{aligned} \mathbf{g} - \tilde{\mathbf{g}} &= \Psi^{\dagger}\mathbf{p}_T \\ \mathbf{g}' - \tilde{\mathbf{g}}' &= \Psi'^{\dagger}\mathbf{p}'_T \end{aligned} \tag{4.23}$$

with $\tilde{\mathbf{g}}$ the component of the solution that lies on the null space of Ψ. The impact of this is that for the overdetermined case the mapping can not be said to hold. Note for the underdetermined scenario the definition of a minimum norm solution is that $\tilde{\mathbf{g}} = \tilde{\mathbf{g}}' = \mathbf{0}$ which removes the issue. Hence for the underdetermined case only both representations will give identical loudspeaker gains, if and only if there is a full mapping between the terms of each expansion type. This would not hold if one term of the first expansion can not be expressed as a linear combination of the terms of the second expansion (the first expansion has a term mapped to the null

space of the second expansion), then the representations are not equivalent and both will lead to differing loudspeaker gains. This result is significant as it shows that two soundfield representations can be considered equivalent in the mode (or order) matching sense, and can both give identical loudspeaker gain definitions if using the minimum norm solution. As of such, the goal is to determine whether such a mapping exists between any given soundfield representations.

4.4.2 2D Ambisonics To Higher Order Stereo Decoder

To consider the mapping between 2D HOA and HOS the Chebyshev polynomials are utilised. The 2D HOA soundfield representation may be regarded as a Fourier series, with the soundfield $p(kr, \hat{\mathbf{r}})$ and $\hat{\mathbf{r}}$ dependent on the azimuthal angle θ [115]

$$p(kr, \theta) = \frac{a_0(kr)}{2} + \sum_{n=1}^{\infty} a_n(kr)\cos(n\theta) + \sum_{n=1}^{\infty} b_n(kr)\sin(n\theta). \tag{4.24}$$

Here $a_0(kr)$, $a_n(kr)$ and $b_n(kr)$ are coefficients found utilising the orthogonality relationships for $\cos(n\theta)$ and $\sin(n\theta)$, which form an orthogonal basis over the unit 1-sphere, $\mathcal{S}^1$ (a unit circle) [121]. The soundfield across the x or y axis may be formulated by setting $\theta = 0, \pi$ or $\theta = \pi/2, 3\pi/2$ respectively. The HOS representation may be expressed using either $\cos^n(\theta)$ or $\sin^n(\theta)$ terms each corresponding to correct reproduction across an orthogonal axis (x and y respectively) as in Eqns. 4.14 and 4.15. Intuitively one might expect the two sets of $\cos(n\theta)$ and $\sin(n\theta)$ terms to span the x, y axis, respectively, based on the HOS results. However, this is not necessarily the case as will now be discussed.

The goal is to find a mapping between the 2D HOA and the HOS representations when considering the soundfield across only the x or the y axis in turn. For this, the Chebyshev polynomials will be used. Notably, the Chebyshev polynomials were used in a similar manner in [129] when mapping soundfield derivatives measured using a differential microphone array to 2D HOA. The Chebyshev polynomials of the first kind, $T_n(x)$, expresses $\cos(n\theta)$ as a polynomial up to order n in terms of $\cos(\theta)$ [121]

$$T_n(\cos\theta) = \cos(n\theta). \tag{4.25}$$

They have the following generating functions [121]

$$\begin{aligned} T_0(\cos\theta) &= 1 \\ T_1(\cos\theta) &= \cos\theta \\ T_n + 1(\cos\theta) &= 2\cos\theta T_n(\cos\theta) - T_{n-1}(\cos\theta) \;\; \forall\, n \;\in\; [2, \infty]. \end{aligned} \tag{4.26}$$

The Chebyshev polynomials provide exactly the mapping which is required to transform between the two soundfield expansions when considering the x axis expansion only, from the 2D HOA B-format representation to the equivalent HOS representation. This shows that a subset of the 2D HOA representation (the $\cos(n\theta)$ terms) spans the equivalent space as the HOS representation (just $\cos^n(\theta)$ terms)

As introduced in the previous section, the 2D mapping matrix $\mathbf{A}^{2D}$ is size $(N+1) \times (N+1)$ and is populated using the Chebyshev polynomial coefficients to give the mapping between HOA to HOS coefficients, having first discarded the sine terms in the HOA representation. Furthermore, expressing a plant matrix and target pressure vector in terms of $\cos(n\theta)$ or $\cos^n(\theta)$ up to the same truncation order N, using the pseudoinverse and assuming the problem is underdetermined will give identical gain definitions. This reinforces the concept that HOS ensures accurate reproduction across a single axis, and generalises previous work linking stereo and Ambisonics from first order to any given order [82, 138].

The entries of the inverse transform (HOS to HOA) given by $(\mathbf{A}^{2D})^{-1}$ are explicitly

$$\begin{aligned} A^{2D,-1}_{n',n} &= t_{n',n} \\ \text{where } T_{n'}(\cos\theta) &= t_{n',0} + t_{n',1}\cos\theta + t_{n',2}\cos^2\theta + \ldots + t_{n',n'}\cos^{n'}\theta \\ &= \sum_{n=0}^{n'} t_{n',n}\cos^n\theta \end{aligned} \tag{4.27}$$

with $t_{n',n}$ the n-th coefficient of $T_{n'}$ whilst due to the nature of the Chebyshev generating functions explicitly stating the HOA to HOS transform entries (for $\mathbf{A}^{2D}$) is non trivial. An example of the order $N = 2$ mapping matrix for both the forward and inverse transform is

$$\mathbf{A}^{2D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix}, \quad (\mathbf{A}^{2D})^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 2 \end{pmatrix}. \tag{4.28}$$

These matrices will be lower triangular, which is a consequence of the mapping being order-limited (the n-th term of one representation is given by terms to order n of the second representation, observable in Eqn. 4.27).

Interestingly the mapping does not exist if the HOS system is expressed in terms of $\sin^n(\theta)$. The Chebyshev polynomials of the second kind, $U_n(\cos\theta)$, expand $\sin(n\theta)$ in an n-th order polynomial as [121]

$$U_n(\cos\theta) = \frac{\sin([n+1]\theta)}{\sin(\theta)} \implies U_{n-1}(\cos\theta)\cdot\sin(\theta) = \sin(n\theta). \tag{4.29}$$

By using the identity $\sin^2(x) = 1 - \cos^2(x)$ to rewrite $U_{n-1}(\cos\theta)$ in terms of an n-th order polynomial in $\sin(x)$ then

$$\sin(n\theta) \equiv \begin{cases} n\text{-th order polynomial in } \sin(\theta) & \text{if } n \text{ odd} \\ (n-1)\text{-th order polynomial in } \sin(\theta) \text{ multiplied by } \cos(\theta) & \text{if } n \text{ even} \end{cases} \tag{4.30}$$

so $\sin(n\theta)$ may only be expanded as a polynomial in $\sin(\theta)$ to the n-th order when n is odd. When n is even, multiplication by an additional $\cos(\theta)$ term is required. This is significant, as the HOS $\sin^n(\theta)$ representation is sufficient to represent the soundfield on the y axis, however the set of $\sin(n\theta)$ from the 2D HOA representation does not. Therefore, a decoder does not exist to map from this subset of 2D HOA to the HOS sine representation, even though in the eyes of HOS the sine representation is as valid as the cosine form. This may be intuitively observed from Eqn. 4.24. Consider evaluation across the x axis, such that $\theta = 0, \pi$, then

$$\begin{aligned} \cos(n0) &= 1 \,\forall\, n \in \mathbb{N}_0 \\ \cos(n\pi) &= -1 \,\forall\, n \in \mathbb{N}_0 \\ \sin(n0) &= 0 \,\forall\, n \in \mathbb{N}_0 \\ \sin(n\pi) &= 0 \,\forall\, n \in \mathbb{N}_0 \\ \implies p(kr, \theta = 0, \pi) &= \frac{a_0(kr)}{2} + \sum_{n=1}^{\infty} a_n(kr)(\pm 1)^n \end{aligned} \tag{4.31}$$

and only the coefficients corresponding to $\cos(n\theta)$ functions are required to represent the soundfield across the x axis. Thus this set of functions span the same space as the HOS cosine representation with the transformation between the two given by Chebyshev polynomials of the first kind. However, when performing analysis across the y axis such that $\theta = \pi/2, 3\pi/2$ then

$$\begin{aligned} \cos\left(\frac{n\pi}{2}\right), \cos\left(\frac{3n\pi}{2}\right) &= \begin{cases} (-1)^{\frac{n}{2}} & \text{if } n \text{ even} \\ 0 & \text{if } n \text{ odd} \end{cases} \quad \forall\, n \in \mathbb{N}_0 \\ \sin\left(\frac{n\pi}{2}\right), \sin\left(\frac{3n\pi}{2}\right) &= \begin{cases} 0 & \text{if } n \text{ even} \\ (\mp 1)^{\frac{n-1}{2}} & \text{if } n \text{ odd} \end{cases} \quad \forall\, n \in \mathbb{N}_0 \\ \implies p\left(kr, \theta = \frac{\pi}{2}, \frac{3\pi}{2}\right) &= \frac{a_0(kr)}{2} + \sum_{n=2,\,\text{n even}}^{\infty} a_n(kr)(-1)^{\frac{n}{2}} + \sum_{n=1,\,\text{n odd}}^{\infty} b_n(kr)(\mp 1)^{\frac{n-1}{2}}. \end{aligned}$$

(4.32)

This result demonstrates that using the 2D HOA representation the soundfield across the y axis requires the $\cos(n\theta)$ coefficients and the $\sin(n\theta)$ coefficients when n is odd and even respectively. No clear mapping then exists between the HOS sine representation which covers the representation of the soundfield across the y axis.

This means in practice, a decoder from 2D HOA to HOS first requires rotation of the 2D HOA soundfield to ensure the x axis aligns with the listener's interaural axis (to pick out the $\cos(n\theta)$ terms), followed by multiplication by the decoding matrix as defined by the Chebyshev polynomials of the first kind. The impact of the existence of this decoder is substantial. It means that all 2D HOA content is able to be rendered over a HOS system. Whilst this does result in discarding some information about the soundfield, the benefit is that whereas 2D HOA requires a minimum of $2N+1$ loudspeakers, the HOS approach requires just $N+1$.

4.4.3 3D Ambsionics To Higher Order Stereo Decoder

A similar decoder from 3D HOA to HOS may also be derived. However, first the spherical harmonic expansion of a plane wave soundfield must be manipulated to reveal the relationship between the two techniques. This will involve deriving the representation of a plane wave across the z axis only by utilising the spherical harmonic expansion, then performing mode matching using this subset of spherical harmonics to define the final decoder.

Consider the pressure due to a plane wave incident with wavevector and wavenumber $\mathbf{k}_i = k\hat{\mathbf{k}}_i$ at a point $\mathbf{r} = r\hat{\mathbf{r}}$ is given by $p(kr, \hat{\mathbf{r}}) = e^{j\mathbf{k}_i \cdot \mathbf{r}}$. Note a 3D coordinate system is now used and the unit vector $\hat{\mathbf{r}}$ denotes the angular dependence through the azimuth and colatitude angles ϕ_i and θ_i. The 3D Jacobi-Anger expansion expresses the plane wave as a summation of spherical harmonics as explained in Eqn. 3.12,

$$p(kr, \hat{\mathbf{r}}) = e^{j\mathbf{k}_i \cdot \mathbf{r}} = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi j^n j_n(kr) Y_n^m(\hat{\mathbf{k}}_i) Y_n^m(\hat{\mathbf{r}})^*. \tag{4.33}$$

with j_n the n-th spherical Bessel function and the direction of arrival of the plane wave being given by $\hat{\mathbf{k}}_i$. The spherical harmonics are a set of functions that form an orthonormal basis over the unit 2-sphere, $\mathcal{S}^2$ (a unit sphere). Hence, any square-integrable well-behaved function on a sphere may be expressed as a weighted linear summation of spherical harmonics. The spherical harmonic Y_n^m, of order n and degree m, may be defined in complex form as [112]

$$Y_n^m(\theta, \phi) = \sqrt{\frac{(2n+1)}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(\cos\theta) e^{jm\phi} \tag{4.34}$$

where $P_n^m(\cos\theta)$ is the associated Legendre polynomial.

Using the spherical harmonic addition theorem the Jacobi-Anger expansion may be expressed purely in terms of Legendre polynomials [115]

$$\begin{aligned} p(kr,\hat{\mathbf{r}}) = e^{j\mathbf{k}_i\cdot\mathbf{r}} &= \sum_{n=0}^{\infty} j^n(2n+1)j_n(kr)P_n(\hat{\mathbf{k}}_i\cdot\hat{\mathbf{r}}) \\ \text{where } P_n(\hat{\mathbf{k}}_i\cdot\hat{\mathbf{r}}) &= \frac{4\pi}{(2n+1)}\sum_{m=-n}^{n} Y_n^m(\hat{\mathbf{k}}_i)^* Y_n^m(\hat{\mathbf{r}}) \text{ with } \hat{\mathbf{k}}_i,\hat{\mathbf{r}}\in\mathcal{S}^2. \end{aligned} \tag{4.35}$$

Consider the product $\hat{\mathbf{k}}_i\cdot\hat{\mathbf{r}} = \cos(\Theta)$, where Θ is the angle between $\hat{\mathbf{k}}_i$ and $\hat{\mathbf{r}}$. This leads to

$$p(kr,\hat{\mathbf{r}}) = e^{jkr\cos(\Theta)} = \sum_{n=0}^{\infty} j^n(2n+1)j_n(kr)P_n(\cos\Theta). \tag{4.36}$$

Next, a slight change to the coordinate system is required. In the literature it is common to align the wavevector of the incident plane wave with the z axis such that $\mathbf{k}_i = k\hat{\mathbf{z}}$, in which case $\mathbf{k}_i\cdot\mathbf{r} = kr\cos(\theta)$ with θ the colatitude. Instead, align $\hat{\mathbf{r}}$ with the z axis such that $\mathbf{r} = r\hat{\mathbf{z}}$ and the dot product $\mathbf{k}_i\cdot\mathbf{r} = kr\cos(\theta_i)$. This effectively fixes the coordinates the plane wave can be evaluated at to positions with $\theta = 0,\pi$ which with $r\in[0,\infty)$ spans the whole z axis as shown in Fig. 4.3. That is the soundfield is now only evaluated across the z axis.

Denote the positive and negative halves of the z axis with subscripts $+,-$. With reference to Fig 4.3 then $\hat{\mathbf{k}}_i\cdot\hat{\mathbf{r}}_+ = \cos(\theta_i)$ and $\hat{\mathbf{k}}_i\cdot\hat{\mathbf{r}}_- = \cos(\pi-\theta_i) = -\cos(\theta_i)$ for the evaluation positions on the positive and negative z axis respectively. Utilising the parity of the Legendre polynomials [115] the plane wave may be expressed as simply

$$\begin{aligned} p_{+,-}(kr,\hat{\mathbf{z}}) &= \sum_{n=0}^{\infty} j^n(2n+1)j_n(kr)P_n(\pm\cos\theta_i) \\ &= \sum_{n=0}^{\infty} (\pm 1)^n j^n(2n+1)j_n(kr)P_n(\cos\theta_i). \\ \text{with } P_n(-x) &= (-1)^n P_n(x). \end{aligned} \tag{4.37}$$

To make use of the spherical Bessel function orthogonality, their argument must be extended to cover the region $(-\infty,\infty)$. Thus define a change in coordinate system

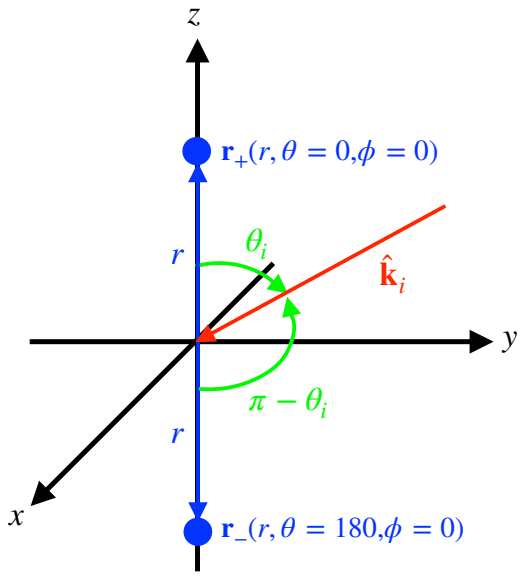


FIGURE 4.3: Geometry for the spherical coordinate system aligning $\hat{\mathbf{r}}$ with the z axis.

$$\begin{aligned} r \in [0,\infty) &\implies r' \in (-\infty,\infty) \\ \theta = 0,\pi &\implies \theta = 0 \\ k > 0 &\implies k > 0 \\ kr \in [0,\infty) &\implies kr' \in (-\infty,\infty) \end{aligned} \tag{4.38}$$

with the soundfield represented as

$$p(kr') = \begin{cases} p_+(kr) = \sum_{n=0}^{\infty} j^n(2n+1)j_n(kr)P_n(\cos\theta_i) & \text{if } kr' \geq 0 \\ p_-(kr) = \sum_{n=0}^{\infty}(-1)^n j^n(2n+1)j_n(kr)P_n(\cos\theta_i) & \text{if } kr' < 0 \end{cases}. \tag{4.39}$$

Note that the expression for $p_-(kr)$ is the same as $p_+(kr)$ except for the additional term $(-1)^n$. In the new coordinate system using r', this factor of $(-1)^n$ may be absorbed back into the spherical Bessel functions of argument kr' using the property $j_n(-x) = (-1)^n j_n(x)$ [117]. Finally,

$$p(kr') = \sum_{n=0}^{\infty} j^n(2n+1)j_n(kr')P_n(\cos\theta_i). \tag{4.40}$$

Crucially, in performing this rotation and fixing the evaluation of the equation to along the z axis, only the zonal spherical harmonics are required. These are the spherical harmonics with $m = 0$ and have no dependence on the azimuthal angle ϕ. It may be observed that these spherical harmonics form a basis for all axisymmetric

functions on a sphere which have no azimuthal dependence. Thus from the full set of spherical harmonics only the following are utilised:

$$\begin{aligned} Y_n^0(\theta,\phi) &= \sqrt{\frac{(2n+1)}{4\pi}\frac{(n-0)!}{(n+0)!}}P_n^0(\cos\theta)e^{j\cdot 0\cdot\phi} \\ &= \sqrt{\frac{(2n+1)}{4\pi}}P_n(\cos\theta). \end{aligned} \tag{4.41}$$

Hence, instead of considering $(N+1)^2$ spherical harmonics to the N-th order, now only $(N+1)$ play a role. Therefore, the 3D HOA soundfield representation using spherical harmonics has been manipulated using a rotation to consider a subset of spherical harmonics that represent the soundfield along the z axis only.

Target Soundfield

In a similar manner to Chapter 3, a set of mode matching equations will now be defined but instead using the expansion in Eqn. 4.40. The target soundfield, $p_T(kr')$ is that of a plane wave and given by

$$p_T(kr') = e^{jkr'\cos(\theta_T)} = \sum_{n=0}^{\infty} j^n(2n+1)j_n(kr')P_n(\cos\theta_T). \tag{4.42}$$

Reproduced Soundfield

For a reproduction array of L equidistant loudspeakers, that act as plane waves with the ℓ-th loudspeaker making an angle θ_ℓ with the z axis and being driven by a gain g_ℓ, the reproduced soundfield, $p_R(kr')$, is

$$p_R(kr') = \sum_{\ell=1}^{L} g_\ell e^{jkr'\cos(\theta_\ell)} = \sum_{\ell=1}^{L} g_\ell \sum_{n=0}^{\infty} j^n(2n+1)j_n(kr')P_n(\cos\theta_\ell). \tag{4.43}$$

Mode Matching

As before, the aim is to find the loudspeaker gains which lead to accurate reproduction of the target soundfield. Begin by equating $p_T(kr') = p_R(kr')$ which leads to

$$\sum_{n=0}^{\infty} j^n(2n+1)j_n(kr')P_n(\cos\theta_T) = \sum_{\ell=1}^{L} g_\ell \sum_{n=0}^{\infty} j^n(2n+1)j_n(kr')P_n(\cos\theta_\ell) \tag{4.44}$$

however there is not an colatitude or azimuthal angle to integrate over and thus no corresponding angular dependent function for which an orthogonality condition can be exploited. Instead, make use of the orthogonality of the spherical Bessel functions over the region $(-\infty, \infty)$, corresponding to the fact that the region being considered is an infinite line (the z axis). The spherical Bessel functions form an orthogonal basis over this region $(-\infty, \infty)$ [116], as presented in Section 3.1.1. The orthogonality relation, which is derived in Appendix C, is

$$\int_{-\infty}^{\infty} j_n(x) j_{n'}(x) dx = \frac{\pi}{(2n+1)} \delta_{nn'} \quad \forall\, n \in \mathbb{N}_0. \tag{4.45}$$

Therefore, multiply both sides of Eqn. 4.44 by a dummy variable $j_{n'}(kr')$ and integrate over $[-\infty, \infty]$

$$\begin{aligned} &\int_{-\infty}^{\infty} \sum_{n=0}^{\infty} j^n (2n+1) P_n(\cos\theta_T) j_n(kr') j_{n'}(kr') dkr' \\ = &\int_{-\infty}^{\infty} \sum_{\ell=1}^{L} g_\ell \sum_{n=0}^{\infty} j^n (2n+1) P_n(\cos\theta_\ell) j_n(kr') j_{n'}(kr') dkr' \end{aligned} \tag{4.46}$$

and using the spherical Bessel function orthogonality condition and removing common terms this simplifies to

$$P_n(\cos\theta_T) = \sum_{\ell=1}^{L} g_\ell P_n(\cos\theta_\ell). \tag{4.47}$$

This is an interesting form of mode matching equation, dependent on mode matching the Legendre polynomials. Due to this specific rotation to align the evaluation along the z axis, this actually equates to correct reproduction along the z axis only. This is significant and bares striking resemblance to HOS.

Decoder Definition

Armed with the new mode matching approach in Eqn. 4.47, all that is needed to decode from the 3D Ambisonics representation to HOS is a mapping between the two soundfield representations. This is in fact very simple, as the n-th order Legendre polynomial is exactly a polynomial in terms of $\cos(\theta)$ to the n-th order by definition. Therefore, the coefficients of the Legendre polynomials fill the entries of the mapping matrix $\mathbf{A}^{3D}$ to decode from the relevant subset of the 3D Ambisonics representation (the space spanned by the spherical harmonics with $m = 0$) to the cosine HOS representation. Furthermore as proven earlier, because such a mapping exists the gain definitions from mode matching the Legendre polynomials in this

manner will be exactly the same as those from the HOS approach (when using the pseudoinverse and assuming $L \geq N + 1$).

The entries of the inverse transform (HOS to HOA) given by $(\mathbf{A}^{3D})^{-1}$ are explicitly

$$\begin{aligned} A^{3D,-1}_{n',n} &= l_{n',n} \\ \text{where } P_{n'}(\cos\theta) &= l_{n',0} + l_{n',1}\cos\theta + l_{n',2}\cos^2\theta + \ldots + l_{n',n'}\cos^{n'}\theta \\ &= \sum_{n=0}^{n'} l_{n',n}\cos^n\theta \end{aligned} \tag{4.48}$$

with $l_{n',n}$ the n-th coefficient of $P_{n'}$. An example of such of the decoder matrix up to order $N = 2$ is

$$\mathbf{A}^{3D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} \end{pmatrix}, \quad (\mathbf{A}^{3D})^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\frac{1}{2} & 0 & \frac{3}{2} \end{pmatrix}. \tag{4.49}$$

As with the 2D case, these matrices will be lower triangular. The rotation is key for the definition of the decoder. The soundfield must be rotated so that the HOS expansion axis is along the z axis, or equivalently that θ is the angle from both the z and interaural axis. This is the only rotation that will pick out the set of $(N + 1)$ spherical harmonics with $m = 0$ that can represent the soundfield along one axis only. This may be viewed as actually rotating the soundfield so that the listener's interaural axis lies along the z axis.

Therefore, a 3D Ambisonics to cosine HOS decoder exists in a similar manner to the 2D Ambisonics decoder. First, the Ambisonics soundfield must be rotated such that the interaural axis aligns with the z axis. Then a subset of the 3D Ambisonic B-format signals must be multiplied by a matrix $\mathbf{A}^{3D}$ whose entries are defined by the Legendre polynomials. As with the 2D Ambisonics decoder, the 3D Ambisonics decoder means all 3D Ambisonics content can be rendered over a HOS system, using only $(N + 1)$ loudspeakers as opposed to $(N + 1)^2$.

4.5 Formulation Of 3D Higher Order Stereophony

So far the HOS approach has been defined using a 2D coordinate system. However, the existence of the 3D HOA to HOS decoder suggests the technique is also applicable for virtual source positions in 3D. In Section 4.2.5 the instability condition demonstrated how, by considering the soundfield reproduced across a single line only, a loudspeaker in front or behind the listener are viewed as identical by the HOS system. This is the scenario when $\sin(\theta_i) = \sin(\theta_j)$ which is satisfied when $\theta_i = \pi - \theta_j$, and is the 2D equivalent of the cone of confusion. Whilst this creates a limitation on the reproduction loudspeaker positions, it can be taken advantage

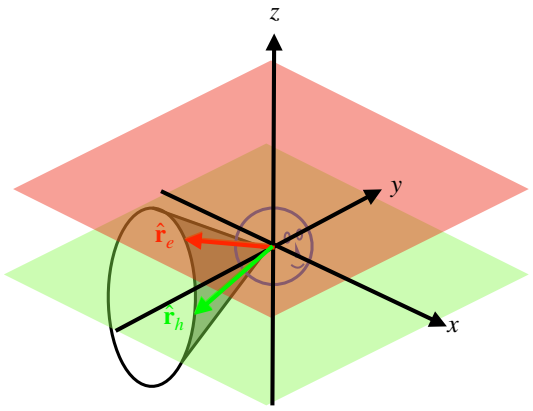


FIGURE 4.4: Illustration of mapping an elevated position ($\hat{\mathbf{r}}_e$ with θ_e, ϕ_e in red), through the cone of confusion to one with no elevation ($\hat{\mathbf{r}}_h$ with $\theta_h = \pi/2, \phi_h$ in green). The planes indicate positions with the same elevation.

of when considering the virtual target source. That is a virtual source behind the listener can be equally represented as a virtual source in the frontal region, as both positions lead to an identical soundfield across the analysis axis, and therefore identical loudspeaker gains.

Now consider the 3D scenario where the virtual source is elevated. The cone of confusion about a given axis is defined as all positions which have the same angle measured from the interaural axis [83]. Therefore, in a similar manner to the 2D case, a source with elevation can be mapped to a source position in the frontal horizontal plane (with no elevation) through the cone of confusion, as both positions will lead to the same soundfield across the evaluation axis. This holds when the free field and plane wave assumptions made in the HOS derivation are satisfied. In this case, consider using a 3D coordinate system with evaluation across the y axis, defined by $\phi_y = \pi/2, \theta_y = \pi/2$. Let $\hat{\mathbf{r}}_e, \hat{\mathbf{r}}_h$ be the desired elevated source position and the equivalent horizontal only position mapped through the cone of confusion respectively, which is illustrated in Fig. 4.4. Note that $\theta_h = \pi/2$ to ensure $\hat{\mathbf{r}}_h$ is on the horizontal plane. We will now derive the equivalent horizontal only source position, with ϕ_h to be determined and $\theta_h = \pi/2$, that maps from $\hat{\mathbf{r}}_e$ to $\hat{\mathbf{r}}_h$ using the cone of confusion. Begin with the scalar product and the great circle distance that states between two positions on a circle

$$\hat{\mathbf{r}}_1 \cdot \hat{\mathbf{r}}_2 = \cos(\Theta) = \cos(\theta_1)\cos(\theta_2) + \sin(\theta_1)\sin(\theta_2)\cos(\phi_1 - \phi_2). \tag{4.50}$$

Θ is defined as the angle between $\hat{\mathbf{r}}_e$ and $\hat{\mathbf{r}}_h$. The cone of confusion requires the angle of the two source positions from the y axis to be equal, therefore

$$\begin{aligned} \hat{\mathbf{r}}_h \cdot \hat{\mathbf{y}} &= \hat{\mathbf{r}}_e \cdot \hat{\mathbf{y}} \\ \cos(\Theta_h) &= \cos(\Theta_e) \\ \cos\left(\phi_h - \frac{\pi}{2}\right) &= \sin(\theta_e)\cos\left(\phi_e - \frac{\pi}{2}\right) \\ \implies \phi_h &= \arccos\left[\sin(\theta_e)\sin(\phi_e)\right] + \frac{\pi}{2}. \end{aligned} \tag{4.51}$$

This relationship maps any elevated source position to the equivalent horizontal only position that leads to the same soundfield across the y axis. Therefore, the resulting HOS loudspeaker gains will be identical for both positions. In this sense, HOS will reproduce any elevated source position through an equivalent horizontal only position, using horizontal only loudspeakers. However, elevation specific cues such as pinna notches will not be reproduced as the assumption holds only when the cone of confusion is valid, which is true for low frequencies, below approximately 4000 Hz [83].

4.6 Experimental Validation

To verify the new proposed HOS technique and the analytical results derived so far, experimental measurements were performed in an anechoic environment to provide data for numerical simulations demonstrating the features of HOS. First, the experimental procedure will be presented. Then numerical results using measurements from a linear microphone array and an Eigenmike EM32 fourth order HOA microphone array will be presented.

The aim of the measurements was to collect a database of transfer functions from a reference source to a microphone array, with the source at a fixed radial distance but measuring for different angular positions in the horizontal plane. From these measurements, any given arrangement of loudspeakers in the horizontal plane can then be simulated using a forward problem as per $\mathbf{p}^{measured} = \Psi^{measured}\mathbf{g}$ where the measured transfer functions fill the entries of the plant matrix, $\Psi^{measured}$. Here the loudspeaker gains, $\mathbf{g}$, are specified as per the desired order HOS or HOA approach. This allows for the evaluation of any given horizontal loudspeaker arrangement and reproduction technique by evaluating the reproduced soundfield $\mathbf{p}$. Here measurements are made in the horizontal plane only and therefore a 2D HOA approach will be considered.

Four systems were compared and are detailed in Fig. 4.5 and Table 4.1. Three HOS systems with increasing truncation orders of $N = 1, 2, 12$ were compared to investigate if increasing the order did yield increased accuracy in the reproduced soundfield at higher frequencies/larger distances from the reproduction point. The first was a classic first order stereo pair (HOS O1) with loudspeakers at $\pm 30^\circ$. The second

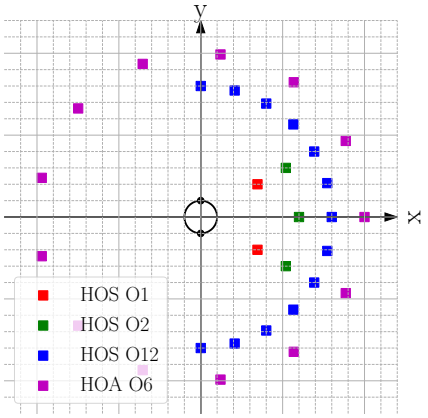


FIGURE 4.5: Loudspeaker layouts for the three HOS and one 2D HOA system. The plot indicates angular arrangements only, as all loudspeaker for all systems were positioned at the same radial distance.

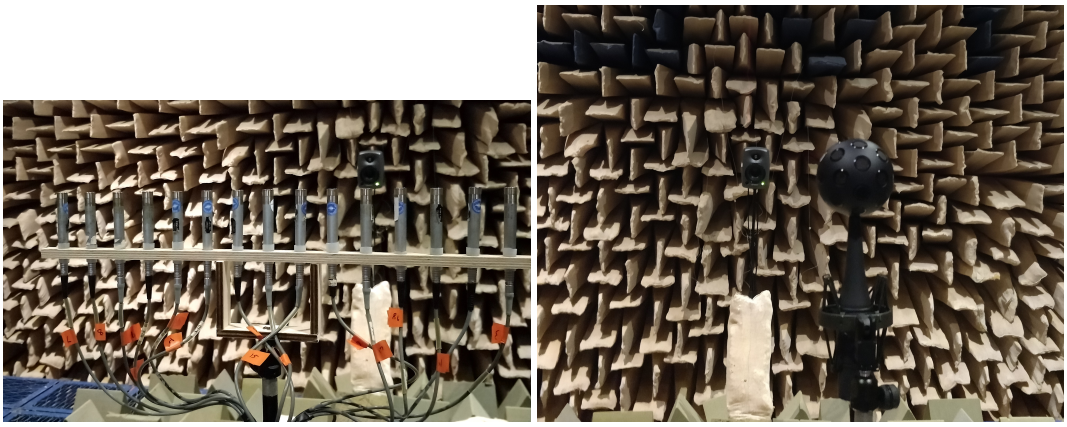
Approach	Truncation Order	Number of Loudspeakers
HOS	1	2
HOS	2	3
HOS	12	13
HOA	6	13

TABLE 4.1: System details for the HOS and HOA loudspeaker arrays under comparison.

order extension to this system (HOS O2) with loudspeakers at $0, \pm 30,^{\circ}$ was included to consider the effect of adding a single loudspeaker and increasing the truncation order, but not increasing the overall loudspeaker array span. A twelfth order system (HOS O12) with thirteen loudspeakers spaced equally across a semicircle in front of the listener was then chosen to compare to a reference sixth order 2D HOA rig (HOA O6), which also requires thirteen loudspeakers but now spaced equally across a circle surrounding the listener. These systems were chosen to compare the performance between HOS and HOA as when using the same number of loudspeakers HOS can achieve a higher truncation order than HOA, as well as considering that the HOS approach requires loudspeakers in front of the listener only.

4.6.1 Experimental Setup

All measurements were performed in the large anechoic chamber at the Institute of Sound and Vibration Research (ISVR), University of Southampton, to ensure freefield conditions. The experimental apparatus was the same for all measurements except for changing the microphone array. Each microphone array was mounted in turn on a controllable turntable. A single Genelec 8020C loudspeaker was used as a reference sound source, positioned 3 meters from the microphone array to best approximate a plane wave source, by ensuring the microphone array was in the far field of the loudspeaker. Exponential sine sweeps were utilised to measure the transfer



(A) Picture of the linear microphone array.

(B) Picture of the Eigenmike EM32 microphone array.

FIGURE 4.6: Experimental setup of the two different microphone arrays.

functions from the loudspeaker to the microphone array [139]. Following each measurement, the turntable was rotated by 1 degree and the measurement was repeated for the new direction. Therefore, the loudspeaker was measured at 360 horizontal angular positions to a 1 degree resolution around the microphone array. All measurements were performed at a sample rate of 48, 000 Hz.

Two different microphone arrays were utilised. The first was a linear array of B&K type 4189 omnidirectional microphones spaced with 0.037 m separation between each microphone and is shown in Fig. 4.6(A). This microphone spacing results in a spatial aliasing frequency of approximately 4600 Hz corresponding to the point at which the microphone spacing equals a half wavelength. The linear array was used to sample the soundfield across a line, to investigate the claim that HOS results in correct reproduction across a given axis. Each microphone was powered by a custom-built set of high-quality preamplifiers which were connected to a Ferrofish A16 analog-digital converter. The Ferrofish A16 was connected via Multichannel Audio Digital Interface (MADI) to a RME Madiface Pro audio interface, which also drove the measurement loudspeaker through its analog output. The microphone array was aligned such that the top of the capsules were approximately in line with the acoustic centre of the loudspeaker.

The second microphone array utilised was an Eigenmike EM32 fourth order HOA microphone, shown in Fig. 4.6(B). The Eigenmike is a 32 capsule array that can be used to sample the spherical harmonic coefficients of a soundfield [140, 141]. This array was measured to consider the contribution of a HOS system to the spherical harmonic modes of the reproduced soundfield, to verify whether HOS does accurately reproduce the $m = 0$ modes in the rotated reference coordinate system only as claimed through the 3D HOA to HOS decoder derivation. The Eigenmike's proprietary FireWire audio interface was used for output to the loudspeaker as well as input of the microphone signals.

4.6.2 Calibration and Post-Processing

The linear microphone array and preamplifier underwent a calibration procedure to account for any gain offsets between microphones or preamplifier channels, as well as to assess the consistency in their frequency responses. First, a standard sound level calibrator outputting a 1000 Hz sine tone at 94 dB was fitted over each microphone in turn, however always using the same preamplifier channel. The level of each microphone was recorded and a calibration offset was calculated to ensure all microphones registered the same level for this fixed sound source. The calibration offsets were relative to the central microphone with ID 8 and are shown in Table 4.2. This type of relative calibration was sufficient as the aim was to ensure all microphones were operating identically, so calibration of the whole system to measure true SPL was not required. The largest calibration value was $+1.3$ dB and the average offset was 0.9 dB, indicating small offsets across the microphone set. This calibration value was applied across all frequencies. Next, an identical measurement was performed however now fixing the microphone capsule and instead changing the preamplifier channel in turn. No variation between preamplifier channels to the precision of the measurement procedure was noticeable thus no additional calibration was required.

Finally, to compare any variations in frequency response across each microphone capsule, each was used in turn to measure the transfer function of the Genelec loudspeaker from the same position. As this relied on manual positioning and alignment of each microphone in turn, this approach is only valid at lower frequencies where small misalignments in microphone position have a negligible affect on the transfer function. The results having undergone the generic gain calibration detailed in Table 4.2 are shown in Fig. 4.7, where the magnitude response using each of the 15 microphones are shown with the ID 8 reference microphone in bold. It is clear there is very little variation in the measurements up till approximately $2000 - 3000$ Hz, where small misalignments in the positioning of the microphones will begin to result in large variations of the measured transfer functions. However, it is clear that all of the microphone responses are very consistent, thus they are well matched and ideal for array type measurements.

No calibration procedure was performed for the Eigenmike, as it came with a native microphone capsule calibration applied within its own software. However, the raw output of the Eigenmike resulted in measurements from the loudspeaker to each of the microphone capsules. Thus the accompanying Eigenmike software was used to convert these measurements to B-format signals, obtaining the impulse responses not from the loudspeaker to the microphone capsules but from the loudspeaker to the B-format channels. This was performed up to truncation order $N = 4$.

Post-processing on all measurements was performed to reduce the impact of any experimental errors introduced due to the equipment. The first issue encountered

Microphone ID	Gain Calibration (dB)
1	0.6
2	-0.2
3	0.5
4	-0.5
5	-0.2
6	1.3
7	-0.9
8	0.0
9	0.7
10	0.1
11	0.5
12	0.4
13	-0.3
14	-0.3
15	-0.4

TABLE 4.2: Microphone calibration offsets relative to the central microphone (ID 8) for a 1000 Hz sine tone.

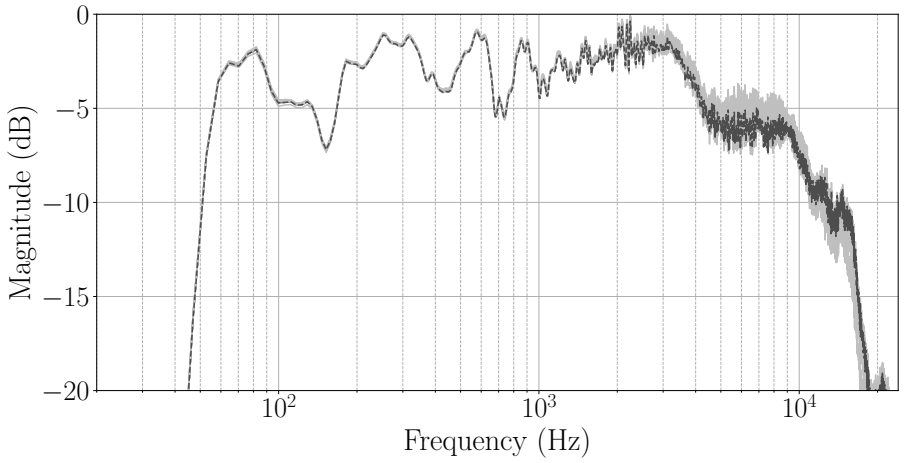


FIGURE 4.7: Freefield measurement of a Genelec 8020C loudspeaker using an identical measurement setup, except changing microphones across the 15 used for the final linear microphone array. The bold dotted line is the microphone with ID 8, which was used as the reference.

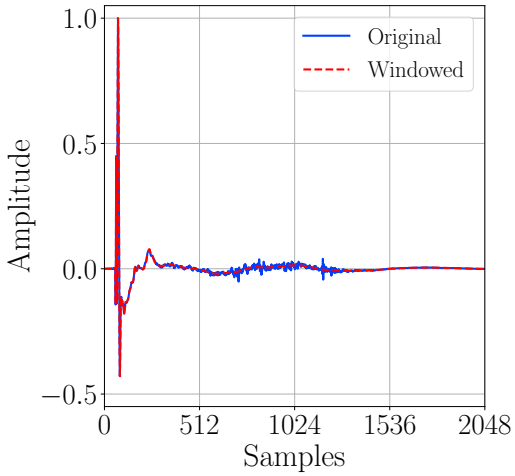
Cutoff Frequency (Hz)	Window Length (ms)	Window Length (samples)
0	85.3	4096
150	62.5	3000
250	25.0	1200
500	7.3	350
700	5.3	256

TABLE 4.3: Set frequencies at which the window lengths were specified. In-between these frequencies the window lengths were calculated by interpolating between the two adjacent cutoff frequencies.

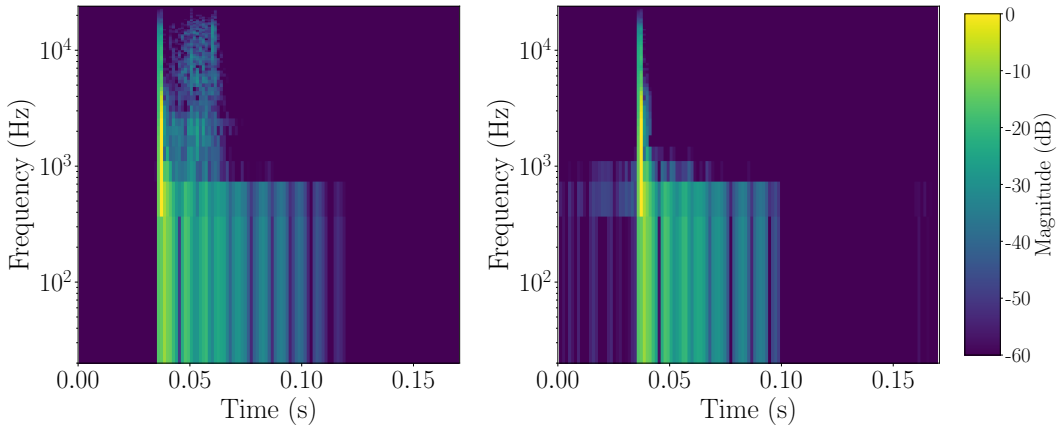
was the presence of a metal grid floor in the anechoic chamber, which was covered with absorbing wedges. However, some reflections due to the floor as well as the equipment remained present in the measured responses. Frequency-dependent windowing was therefore employed [142, 143], which was used in favour of standard windowing to ensure a short window could be applied at high frequencies whilst a longer window could be utilised at low frequencies to avoid prematurely cutting the response of the loudspeaker. The frequency-dependent windowing used a Tukey window. Five window lengths with cutoff frequencies shown in Table 4.3 were tuned manually to ensure the windows operated as desired. However, to ensure no discontinuities through the abrupt changing of window lengths the technique was adapted from the literature and instead a smoothly varying window size between the frequencies in Table 4.3 was defined. Thus between these values the window length was calculated using linear interpolation to ensure a smoothly decreasing window length with increasing frequency.

An example of an impulse response of the Genelec loudspeaker measured using one of the omnidirectional microphones, then processed using the frequency-dependent windowing is shown in Fig. 4.8(A). This demonstrates how the main structure of the impulse response is clearly maintained after windowing however later reflections have been removed. Spectrograms of these two impulse responses are also shown in Fig. 4.8(B) and 4.8(C) where it is apparent the reflections were primarily an issue at high frequencies and are removed by the process. The longer, time-varying low frequency response of the loudspeaker is maintained through the process, however a small amount of pre-ringing artefacts are introduced which may be viewed in the spectrogram.

Following the windowing procedure, time alignment was applied but only to the Eigenmike dataset across all angular positions. The goal of the time alignment was to compensate for any small misalignments due to wobble about the axis of rotation from having rotated the Eigenmike using the turntable. This was only a noticeable issue with the Eigenmike due to the construction of its shock mount, where a maximum deviation of 4 samples was noticed due to this effect. Therefore, onset-based time alignment as detailed in [144] was used, which is more commonly applied to remove the interaural time difference from measured HRTFs. The time of arrival for



(A) Example measured impulse response before and after frequency-dependent windowing.



(B) Spectrogram of the impulse response before windowing.

(C) Spectrogram of the impulse response after windowing.

FIGURE 4.8: Comparison of a measured impulse response before and after the frequency-dependent windowing.

each angular measurement position was calculated using the first B-format channel only, corresponding to an omnidirectional microphone response. It was assumed that any delay apparent in this channel would also be identical in the other B-format channels of the same measurement position, as it is due to a physical offset of the microphone. The time of arrival was defined as when the impulse response passed a threshold of -20 dB relative to the main peak of that given impulse response, on the 10 times upsampled and lowpass filtered (8th order butterworth, cutoff frequency at 3000 Hz) impulse response of the first B-format channel. Next, the time of arrivals were converted into compensation delays relative to the first measurement position where the loudspeaker was positioned at 0 degrees. This means the absolute time of arrival from the loudspeaker to the microphone array was not removed from the impulse responses, rather any relative delay across the dataset compared to the first 0 degree measurement position was equalised. The delays were removed using the fractional delay implementation in the SUpDEq toolbox [144].

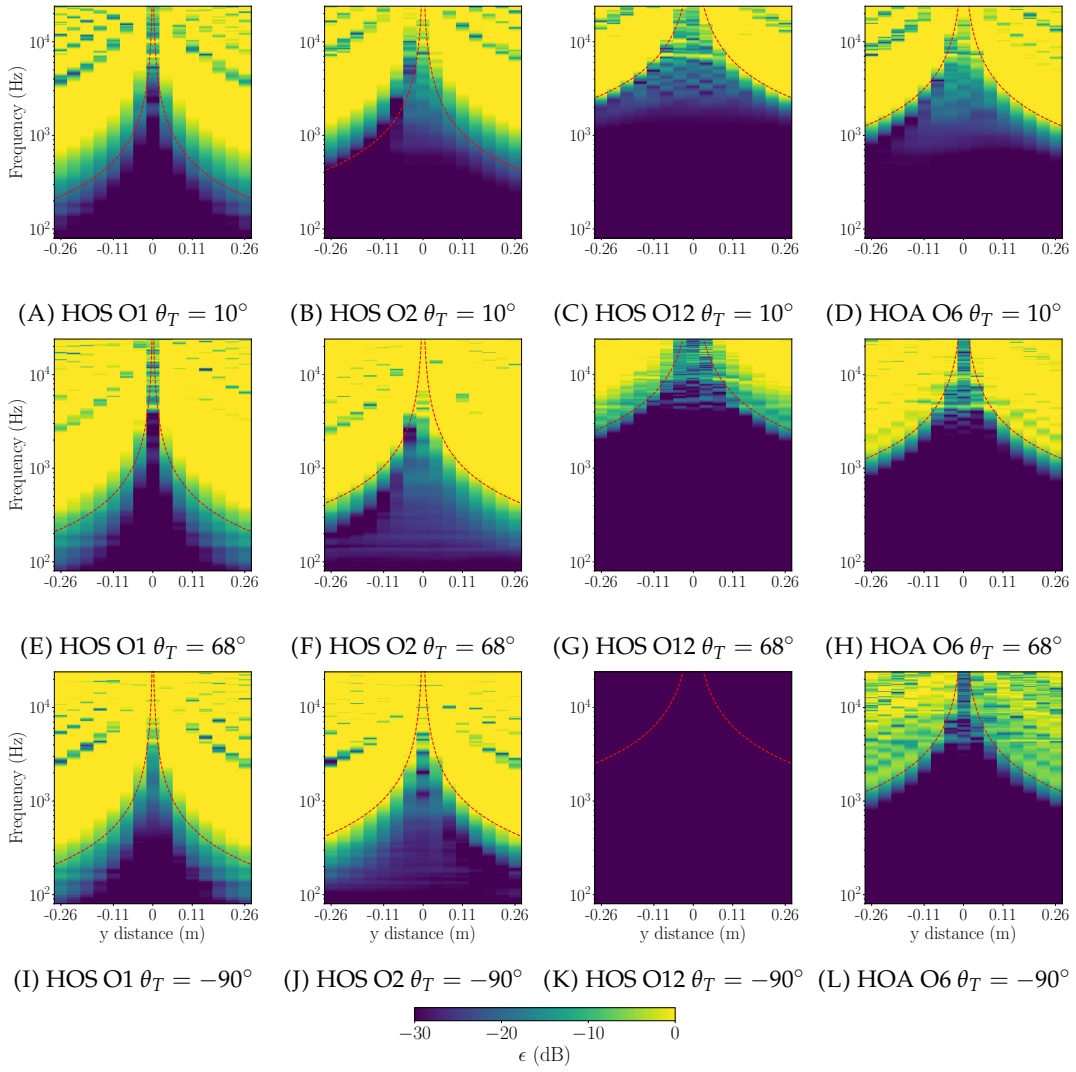
4.6.3 Linear Array Results

First, results from using the linear microphone array data will be considered. The reproduced soundfield at each microphone position, $\mathbf{p}_R^{measured}$, was simulated via a forward problem such that $\mathbf{p}_R = \Psi^{measured}\mathbf{g}$. Here, $\mathbf{p}_R^{measured}$ is the vector of pressures at the microphone positions, thus the soundfield is being analysed across the reproduction axis which in this case is the y axis. The target soundfield, $\mathbf{p}_T^{measured}$, was defined as the measurement of the loudspeaker at the required virtual source position. The normalised complex error, ϵ, is defined as the difference between the reproduced and target soundfields, normalised by the target pressure such that for each microphone position

$$\epsilon = \frac{|p_T^{measured} - p_R^{measured}|^2}{|p_T^{measured}|^2}. \tag{4.52}$$

This error metric takes into account both magnitude and phase differences between the reproduced and target field. The error is considered using a decibel scale, where a smaller value indicates less error therefore the soundfield reproduced by the system matches that of the target.

Fig. 4.9 shows the complex error across the array length (x axis on the plots) and as a function of frequency for each of the reproduction systems. Three virtual source positions are considered individually, chosen to be $\theta_T = 10°, 68°, -90°$. The red dotted lines indicate the $N = kr$ limit for each of the systems [48], under which it is expected that all systems should perform well with little error in this region. These results confirm that increasing the order of the HOS approach leads to more accurate reproduction with respect to the kr quantity, as little error is observed within the $N = kr$ limit whilst outside of it maximal error occurs. Therefore utilising a higher order system leads to more accurate reproduction at higher frequencies and across a



(A) HOS O1 $\theta_T = 10°$ (B) HOS O2 $\theta_T = 10°$ (C) HOS O12 $\theta_T = 10°$ (D) HOA O6 $\theta_T = 10°$

(E) HOS O1 $\theta_T = 68°$ (F) HOS O2 $\theta_T = 68°$ (G) HOS O12 $\theta_T = 68°$ (H) HOA O6 $\theta_T = 68°$

(I) HOS O1 $\theta_T = -90°$ (J) HOS O2 $\theta_T = -90°$ (K) HOS O12 $\theta_T = -90°$ (L) HOA O6 $\theta_T = -90°$

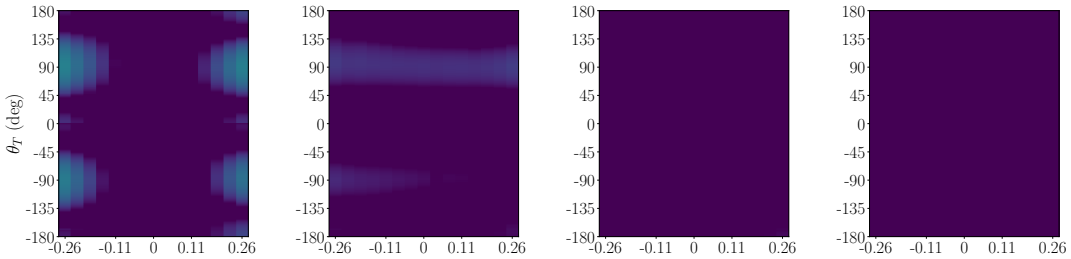
FIGURE 4.9: Complex error for three HOS and one 2D HOA system, across frequency and distance along the y axis for a fixed virtual source position θ_T. The red dotted lines indicate the $N = kr$ limit for that given order N.

larger distance across the y axis. In general, HOS appears to follow this $N = kr$ rule of thumb originally derived for HOA. Little variation is seen when changing the position of the virtual source. There is however a small increase in error within the $N = kr$ area for both HOS O1 and O2 for both $\theta_T = 68°, -90°$, when reproducing a source out of span of the loudspeaker array. The HOS O12 system has zero error for $\theta_T = -90°$, as the system has a loudspeaker positioned here and the gain definitions lead to just turning this loudspeaker on, therefore the virtual source becomes real and there must be zero error. Finally, if the number of loudspeakers is fixed the HOS approach is advantageous to the HOA technique. This is because a higher order of reproduction can be achieved leading to a larger region of validity across the reproduction axis, whilst also only requiring loudspeakers in front of the listener making for a more accessible loudspeaker array. Finally, even within the $N = kr$ bounds there is some level of error at high frequencies, which corresponds to the spatial aliasing limit due to this microphone array spacing of approximately 4600 Hz.

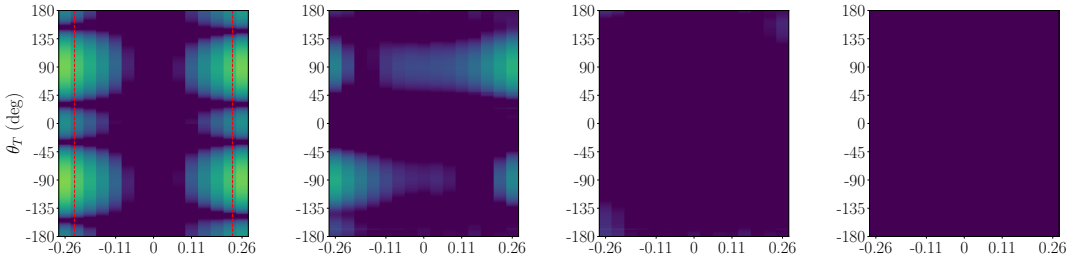
Fig. 4.10 shows the complex error for all the systems across the array length, however now as a function of all virtual source angular positions. Individual frequencies are shown in turn from $125 - 4000$ Hz remaining under the spatial aliasing limit of the linear microphone array. Again, the red dotted lines indicate the $N = kr$ region. Once more it is apparent that increasing the order of the system leads to an increased area of accurate reproduction across the y axis. As all systems use the minimum number of loudspeakers required, both the HOS and HOA gain definitions activate a single loudspeaker when the virtual source is positioned at that given loudspeaker. Therefore, lines of zero error are apparent throughout the results when θ_T is at a loudspeaker position. This also reveals how the HOS technique takes advantage of the instability condition explained in Section 4.2.5, as due to the cone of confusion the soundfield due a virtual source along the analysis axis is equal for θ_T and $\theta'_T = 180° - \theta_T$. This can be viewed as a mirroring operation about the analysis axis. Therefore at loudspeaker positions the soundfield is also correct for the mirrored position on the cone of confusion. For example with HOS O1 the loudspeakers are positioned at $\pm 30°$, thus correct reproduction is observed when $\theta_T = \pm 30°$ and $\pm 150°$. Finally, comparing HOS O12 and HOA O6 it is clear that as the frequency increases, the soundfield is reproduced correctly over a larger distance on the evaluation axis for the HOS technique due to it working to a higher truncation order.

4.6.4 Eigenmike Measurements

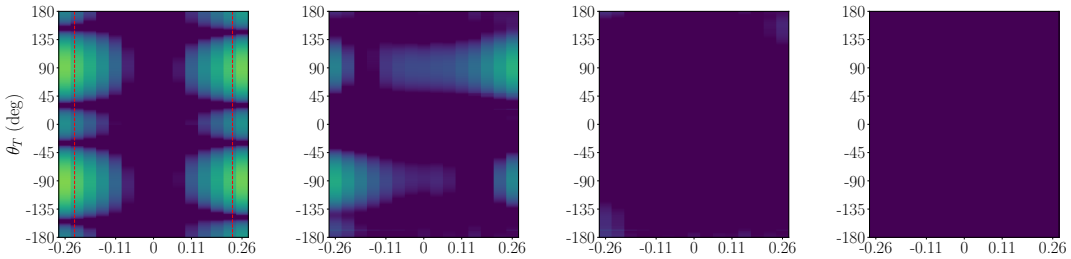
For the Eigenmike results a similar approach was used. A forward problem was performed except now the plant matrix was populated with the loudspeaker to B-format transfer functions, and the vector $\mathbf{p}_R^{measured}$ is now the reproduced spherical harmonic coefficients of order $n \in [0, 4], m = 0$ only. Only the $m = 0$ spherical



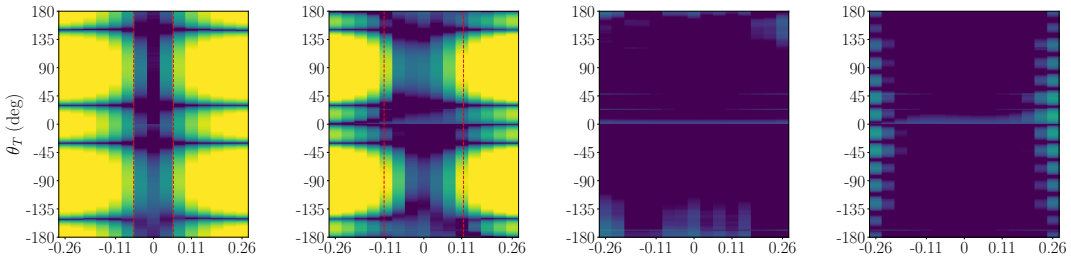
(A) HOS O1, 125 Hz (B) HOS O2, 125 Hz (C) HOS O12, 125 Hz (D) HOA O6, 125 Hz



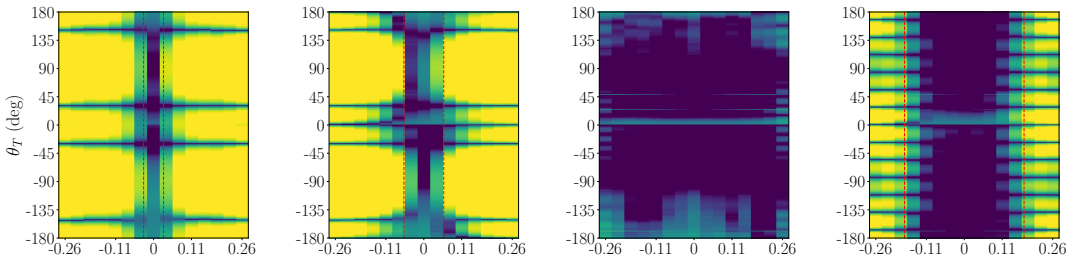
(E) HOS O1, 250 Hz (F) HOS O2, 250 Hz (G) HOS O12, 250 Hz (H) HOA O6, 250 Hz



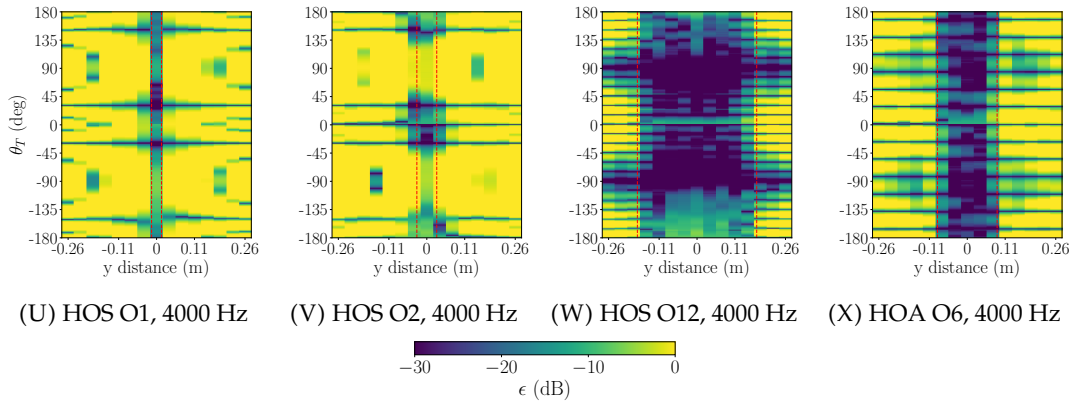
(I) HOS O1, 500 Hz (J) HOS O2, 500 Hz (K) HOS O12, 500 Hz (L) HOA O6, 500 Hz



(M) HOS O1, 1000 Hz (N) HOS O2, 1000 Hz (O) HOS O12, 1000 Hz (P) HOA O6, 1000 Hz



(Q) HOS O1, 2000 Hz (R) HOS O2, 2000 Hz (S) HOS O12, 2000 Hz (T) HOA O6, 2000 Hz



(U) HOS O1, 4000 Hz (V) HOS O2, 4000 Hz (W) HOS O12, 4000 Hz (X) HOA O6, 4000 Hz

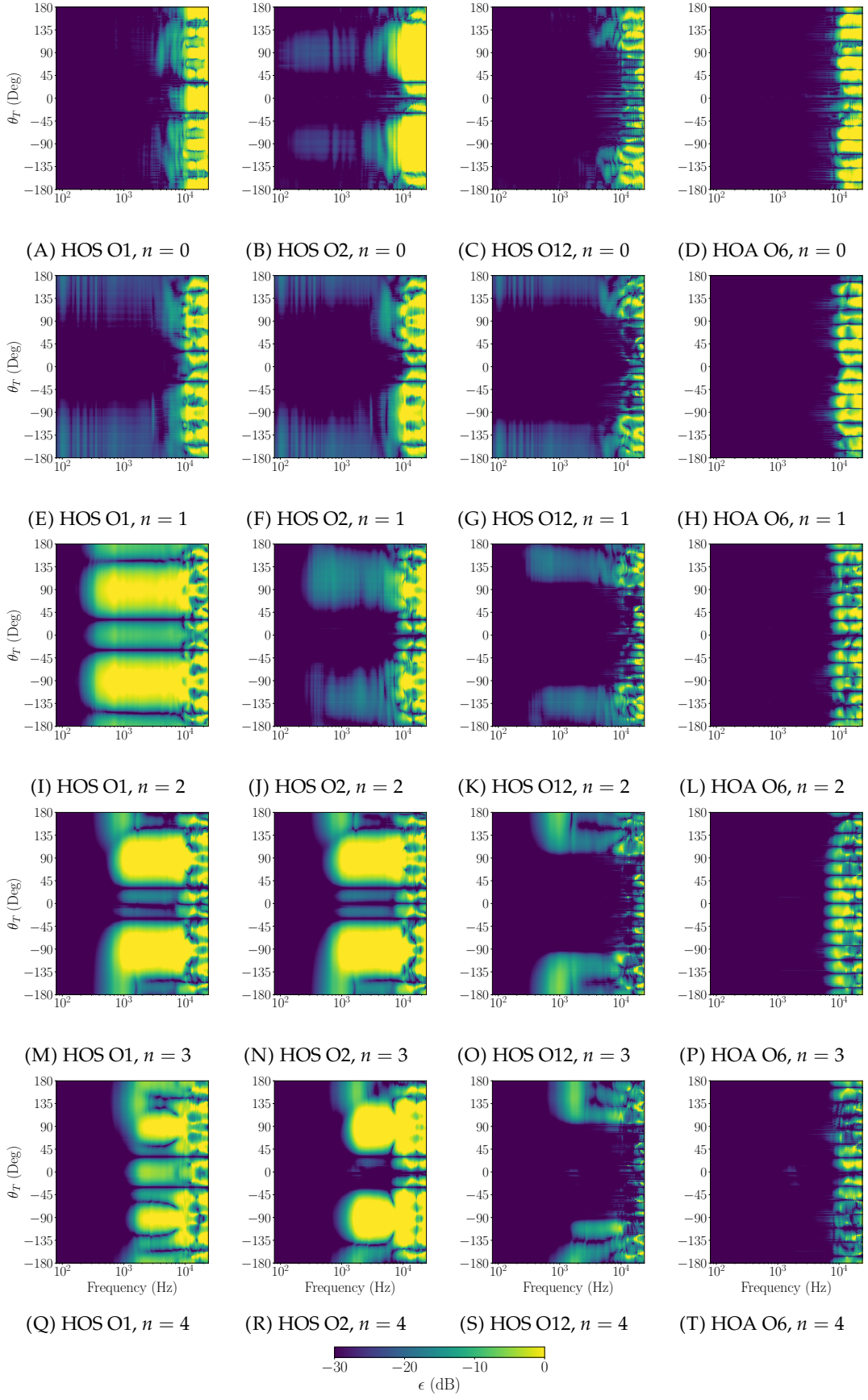
FIGURE 4.10: Complex error for three HOS and one 2D HOA system, across all virtual source positions and distance along the y axis for fixed frequencies. The red dotted lines indicate the $N = kr$ limit for that given order N.

harmonics are considered as following the rotation to align the z axis with the reproduction axis, only the $m = 0$ subset is required to represent the soundfield along this axis as detailed in the HOA to HOS decoder. It is worth recalling that a soundfield reproduced using HOS will be incorrect when deviating from the line defined by the $m = 0$ spherical harmonics however, unlike with HOA. Following the forward problem, the complex error was again calculated between the reproduced and target soundfields. However, the first spherical harmonic coefficient corresponding to $n, m = 0, 0$ (or an omnidirectional microphone response) was chosen as the normalisation for all channels. Therefore

$$\epsilon = \frac{|p_T^{measured} - p_R^{measured}|^2}{|W_T^{measured}|^2} \tag{4.53}$$

with $W_T^{measured}$ the $n, m = 0, 0$ channel.

Fig. 4.11 shows the complex error for the reproduced spherical harmonic coefficients corresponding to each B-format channel, as a function of frequency and virtual source angular position. For all systems and B-format channels the Eigenmike introduces spatial aliasing above 6000 Hz, therefore there is considerable error in this region regardless of the technique. It is expected that beyond the truncation order of each of the systems, the spherical harmonic coefficient should not be correctly reproduced. The results demonstrate that the HOS system does indeed accurately reproduce the $m = 0$ channels up to the truncation order. This experimentally verifies the link established between HOS and HOA through the decoder definition, and that controlling the $m = 0$ channels only corresponds to accurate soundfield reproduction along a single axis. Whilst the HOA O6 technique exhibits no error at all for all virtual source positions, the HOS approaches all encounter issues when the virtual source is positioned behind the listener where there are no loudspeakers positioned. Here, the error ranges from approximately -25 to -15 dB. This suggests



(A) HOS O1, $n = 0$ (B) HOS O2, $n = 0$ (C) HOS O12, $n = 0$ (D) HOA O6, $n = 0$

(E) HOS O1, $n = 1$ (F) HOS O2, $n = 1$ (G) HOS O12, $n = 1$ (H) HOA O6, $n = 1$

(I) HOS O1, $n = 2$ (J) HOS O2, $n = 2$ (K) HOS O12, $n = 2$ (L) HOA O6, $n = 2$

(M) HOS O1, $n = 3$ (N) HOS O2, $n = 3$ (O) HOS O12, $n = 3$ (P) HOA O6, $n = 3$

(Q) HOS O1, $n = 4$ (R) HOS O2, $n = 4$ (S) HOS O12, $n = 4$ (T) HOA O6, $n = 4$

FIGURE 4.11: Complex error for three HOS and one 2D HOA system, for the reproduced degree $m = 0$ spherical harmonic coefficients as a function of frequency and virtual source position. Each row shows a different order spherical harmonic.

that whilst rear virtual sources can be achieved with HOS and just frontal loudspeakers, they may not be reproduced as robustly compared to using the HOA approach. However, HOA requires a fully surrounding loudspeaker array for optimal performance. Interestingly, at low frequencies for spherical harmonic coefficients above the truncation order the HOS approach reproduces the coefficient correctly up to approximately $800 - 1000$ Hz. This can be observed with HOS O1 and O2, for channels $n, m = 3, 0$ and $n, m = 4, 0$.

4.7 Chapter Review

This chapter has introduced the theoretical foundations for a new soundfield reproduction technique titled Higher Order Stereophony (HOS). HOS is founded in the Taylor expansion of the soundfield due to an incident plane wave across one axis only. This expansion represents the soundfield as an infinite summation of the soundfields derivatives evaluated about an expansion point. To reproduce a virtual sound source all that is required is the correct binaural signals, therefore the assumption is that accurate reproduction of the target soundfield along the interaural axis only may be sufficient to reproduce these binaural signals. For now, no HRTF was included in the analysis however this scenario will be discussed later. HOS is thus order matching with each term defined by the derivative of the soundfield at the central reproduction point. The resulting loudspeaker gains are shown to be simple panning functions, assuming plane wave virtual sources and loudspeakers. The stereo sine law is shown to be a first order HOS system. This motivates the name of the technique, where HOS may be recognised as a generalisation of this classic audio reproduction approach to a general order and any number of loudspeakers.

HOS shares many similarities to Higher Order Ambisonics (HOA). Utilising the 2D and 3D HOA soundfield representation, decoders from both approaches to HOS were derived. Both decoders first required a rotation to align the interaural axis across the x or z axis in 2D and 3D respectively. The 2D HOA to HOS decoder utilised the Chebyshev polynomials, whilst the 3D HOA to HOS decoder used a subset of spherical harmonics with $m = 0$. Importantly, N-th order HOS requires only $(N + 1)$ channels/reproduction loudspeakers, whilst 2D and 3D HOA requires $(2N + 1)$ and $(N + 1)^2$ respectively. Ideally HOA requires loudspeakers equally distributed over a circle or sphere, however HOS can use loudspeakers in front of the listener only across a semicircle.

Experimental validation of the theoretical results was also presented. Two different microphone arrays were utilised, a linear microphone array to consider the reproduced soundfield across a line and a spherical microphone array to consider the reproduced spherical harmonic coefficients of the soundfield. The linear microphone array results confirmed that HOS correctly reproduces the soundfield across a line, appearing to follow the classic $N = kr$ rule of thumb. Increasing the order of the

system leads to increased accuracy of the soundfield at higher frequencies/larger distances away from the expansion centre point. Furthermore, a twelfth order HOS system and a sixth order 2D HOA system were compared as both require a minimum of thirteen loudspeakers. The HOS system was shown to be advantageous as higher order reproduction could be achieved using both the same number of loudspeakers and loudspeakers positioned only in front of the listener, which led to reproduction to a higher frequency/spatial limit. The spherical microphone array results also confirmed that HOS correctly reproduces the $m = 0$ spherical harmonic channels only, and that the subset of $m = 0$ spherical harmonics can be used to represent a soundfield across the z axis. However, for rear virtual sources the HOA approach appeared to be more robust.

Chapter 5

Dynamic Higher Order Stereophony

So far, the HOS technique has been presented as accurate reproduction of the sound-field along a single line (or axis) only. The core principle behind HOS is that this reproduction line coincides with the listener's interaural axis. This means the listener must remain in a fixed head position, which is a severely limiting factor for the usefulness of the approach.

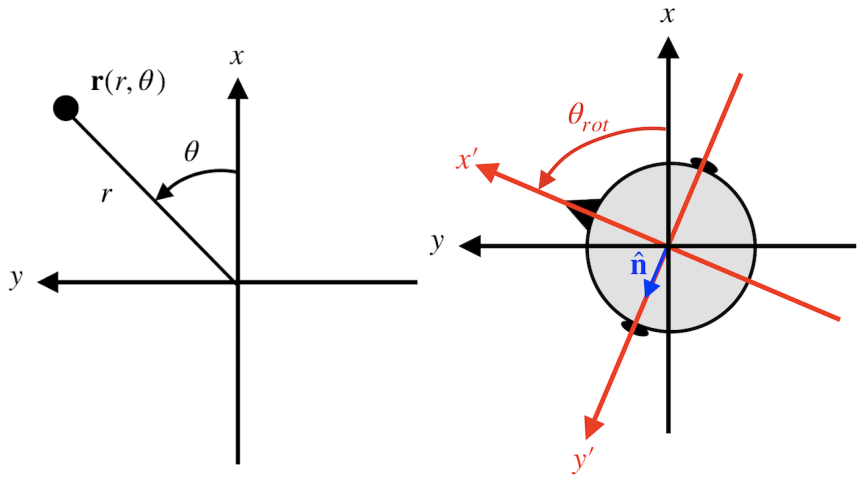
One important consequence of fixing the head orientation is that the listener can not attain dynamic localisation cues which are important in resolving front-back confusions. Dynamic cues may be introduced by a moving sound source or by movements of the head such that a change in the ITD and ILD is dynamically introduced [99, 100]. The alteration in the binaural cues during these head rotations reveals the position of the sound source. This is particularly relevant as the HOS technique takes advantage of front-back confusions, for example by rendering rear virtual sources on the equivalent cone of confusion in the frontal position of the listener as the sound-field due to both source positions across the reproduction axis is equivalent.

This chapter will introduce a dynamic extension of HOS, that utilises listener tracking and allows for dynamic adaption of the loudspeaker gains to adjust for listener head movements. In doing so, the system is extended to dynamically rotate the reproduction axis, so that it always aligns with the interaural axis regardless of listener movements. Following this, the head-tracked sine law is derived using the HOS framework as well as special cases leading to other classic stereo techniques. The HOA to HOS decoders are also extended to allow for listener head rotations. Finally, a listening test comparing dynamic HOS to 2D HOA is presented.

5.1 Dynamic Higher Order Stereophony Order Matching

5.1.1 Expansion Along A Generalised Axis

Using a 2D coordinate system as in Fig. 5.1(A), consider a listener with their head centred at the origin according to Fig. 5.1(B). The vector $\hat{\mathbf{n}}$ points from the head centre



(A) 2D coordinate system.

(B) Example head orientation allowing the listener to rotate their head, thus defining both the fixed and listener's frame of reference.

FIGURE 5.1: Coordinate system and geometry layout.

to the left ear defining the interaural axis, with the assumption that the listener's ears are diametrically opposed across the head. $\hat{\mathbf{n}}$ is defined by the head rotation angle θ_{rot}. A second frame of reference, $\mathbf{x}'$ and $\mathbf{y}'$ is defined as the listeners frame of reference such that $\hat{\mathbf{x}}'$ always points straight in front of the listener and $\hat{\mathbf{y}}' = \hat{\mathbf{n}}$, that is the interaural axis is always along the $\mathbf{y}'$ axis as demonstrated in Fig. 5.1(B).

Let the function $p(\mathbf{r})$ be an infinitely differentiable function at a point $\mathbf{r}_0$ where $\mathbf{r} = [x, y]^T$. The multi-variable Taylor expansion of this soundfield is [121]

$$p(\mathbf{r}) = \sum_{n=0}^{\infty} \frac{[a\hat{\mathbf{n}} \cdot \nabla]^n}{n!} p(\mathbf{r}_0) \tag{5.1}$$

where the step from the expansion point $\mathbf{r} - \mathbf{r}_0 = a\hat{\mathbf{n}}$. Previously, HOS was defined by evaluating this expression assuming the soundfield is an incident plane wave. Thus $p(\mathbf{r}) = e^{j\mathbf{k}_i \cdot \mathbf{r}}$ with $\mathbf{k}_i = k[\cos(\theta_i), \sin(\theta_i)]^T$, θ_i the incident angle of the plane wave, k the wavenumber and j the imaginary unit. Assuming that the listener's head aligns along the y axis such that $\hat{\mathbf{n}} = \hat{\mathbf{y}}$ leads to the HOS expansion

$$p(a) = \sum_{n=0}^{\infty} \frac{[jka \sin(\theta_i)]^n}{n!}. \tag{5.2}$$

where the plane wave pressure at the head centre is unitary. Truncation of this infinite summation is made to an order N, and the n-th order term is given by the n-th order derivative of the soundfield, which for a plane wave includes the sine of the incident angle to the power of n. However, this simple representation is only given when the expansion is made across the y axis. Performing the same procedure using $\hat{\mathbf{n}} = \hat{\mathbf{x}}$ results in a similar representation except the sine terms become cosines.

$$p(a) = \sum_{n=0}^{\infty} \frac{[jka\cos(\theta_i)]^n}{n!}. \tag{5.3}$$

For any $\hat{\mathbf{n}}$ direction between these x and y axes, the evaluation is more complex.

This issue is due to the evaluation of the product

$$[a\hat{\mathbf{n}} \cdot \nabla]^n = a^n \left[n_x \frac{\partial}{\partial x} + n_y \frac{\partial}{\partial y} \right]^n \tag{5.4}$$

which considers the soundfield in any generalised direction $\hat{\mathbf{n}}$ however results in many cross-derivative products to evaluate. To account for rotations of the listener's head, it might first be obvious to change the definition of $\hat{\mathbf{n}}$ such that the n-th order term for the multi-variable Taylor expansion will be dependant on $[\hat{\mathbf{n}} \cdot \nabla]^n$. However, as $\hat{\mathbf{n}}$ may not be not aligned with an axis a simplified representation can not be achieved. When $\hat{\mathbf{n}}$ is aligned along one Cartesian axis then only the single variable Taylor expansion is required.

Instead, head rotations may be compensated for easily by considering the two separate frames of reference, a fixed frame of reference and the listener's frame of reference. This is demonstrated in Fig. 5.1(B). Measuring all necessary angles in the listener's frame of reference will therefore compensate for the head rotation, θ_{rot}. In practice, this is easily implemented by converting between the two frames of reference for any given angle with $\theta' = \theta - \theta_{rot}$. In doing so, the dynamic HOS representation is always such that $\hat{\mathbf{n}} = \hat{\mathbf{y}}'$ and the expansion

$$\begin{aligned} p(a) &= \sum_{n=0}^{\infty} \frac{[jka\sin(\theta_i')]^n}{n!} \\ &= \sum_{n=0}^{\infty} \frac{[jka\sin(\theta_i - \theta_{rot})]^n}{n!} \end{aligned} \tag{5.5}$$

may always be used. This means, HOS can be adapted for listener head rotations by using a head-tracker and compensating all angles in the fixed frame of reference by θ_{rot}.

An interesting point arises when considering head rotation compensation. As noted previously, expansion along the x axis results in cosine terms in the expansion instead of sine terms. Choice of the x or y axis to perform the expansion across (or indeed any arbitrary axis) is equally valid and corresponds to reproduction along two orthogonal axes in a given reference frame. However, these two representations using $\sin^n(x)$ or $\cos^n(x)$ may be seen to be equal by setting $\theta_{rot} = -90°$, such that $\hat{\mathbf{n}} = \hat{\mathbf{x}}$ and any given angle $\theta' = \theta + 90°$. Using the identity $\sin(x + 90°) = \cos(x)$

this shows the two representations may be equally used, as long as the angles are defined properly. Therefore, it is valid to use either sine or cosine terms in the definition of the plant matrix.

So far, only listener head rotations have been considered. To compensate for translations of the listener's head, a delay may be applied to each of the loudspeaker gains to keep them acoustically equidistant. In this sense, all loudspeakers in the array can be kept at a constant radius away from the listener regardless of the listener's position. As this is just a simple delay per loudspeaker, which may be calculated directly from the output of a head-tracker, the delays may be formulated and implemented independently to the HOS gain definitions to maintain the frequency-independence in the HOS gain calculations. Physically, a translation of the listener's head will move the point about which the expansion is performed, which is always kept as the listener's head centre.

With the dynamic HOS soundfield representation defined that allows for listener movements, this representation will now be used as previously to define a set of loudspeaker panning functions. However, these will now be adaptive panning functions that depend on the listener head orientation, reproducing the soundfield accurately across the interaural axis regardless of the listener head movements.

5.1.2 Target Soundfield

The target soundfield is that of the incident plane wave which the loudspeaker array is attempting to reproduce. The target soundfield, $p_T(ka)$, is given by the dynamic HOS expansion as in Eqn. 5.5 such that

$$p_T(ka) = \sum_{n=0}^{\infty} \frac{[jka\sin(\theta_T - \theta_{rot})]^n}{n!} \tag{5.6}$$

where it is assumed the listener is rotating their head freely as defined by the head rotation angle θ_{rot} and the incident target plane wave arrives at an angle defined by θ_T.

5.1.3 Reproduced Soundfield

The reproduced soundfield is given by the overall contributions of the loudspeaker array. Assume an array of L equidistant loudspeakers which propagate as plane wave sources. The ℓ-th loudspeaker is situated at θ_ℓ and driven by a gain g_ℓ. Here, the gain may contain a delay term to ensure the loudspeakers are acoustically radially equidistant if this is required. In this manner, listener translations can also be accounted for as discussed previously. Thus the reproduced soundfield is the summation of all contributions from each individual loudspeaker:

$$p_R(ka) = \sum_{\ell=1}^{L} g_\ell \sum_{n=0}^{\infty} \frac{[jka\sin(\theta_\ell - \theta_{rot})]^n}{n!} \tag{5.7}$$

5.1.4 Order Matching

For exact soundfield reproduction the following must be satisfied, $p_T(ka) = p_R(ka)$. In practise this requires an infinite array of loudspeakers, however truncation to a finite order/finite loudspeaker array will be considered later. Therefore,

$$\begin{gathered} p_T(ka) = p_R(ka) \\ \sum_{n=0}^{\infty} \frac{[jka\sin(\theta_T - \theta_{rot})]^n}{n!} = \sum_{\ell=1}^{L} g_\ell \sum_{n=0}^{\infty} \frac{[jka\sin(\theta_\ell - \theta_{rot})]^n}{n!}. \end{gathered} \tag{5.8}$$

The order matching principle is now applied, where the n-th order terms from the expansion of the target and reproduced soundfields are equated. Whilst traditionally an orthogonality principle is applied to achieve this matching, manually setting this still results in the correct overall summation and thus representation of the soundfield. However, it may not be the only possible solution. It is therefore required that

$$\frac{[jka\sin(\theta_T - \theta_{rot})]^n}{n!} = \sum_{\ell=1}^{L} g_\ell \frac{[jka\sin(\theta_\ell - \theta_{rot})]^n}{n!} \;\; \forall\, n \;\in\; \mathbb{N}_0. \tag{5.9}$$

Finally removing all common terms in the n-th order definition leaves

$$\sin^n(\theta_T - \theta_{rot}) = \sum_{\ell=1}^{L} g_\ell \sin^n(\theta_\ell - \theta_{rot}) \;\; \forall\, n \;\in\; \mathbb{N}_0 \tag{5.10}$$

which is the n-th dynamic HOS order matching equation.

5.1.5 Loudspeaker Gain Definitions

Finally, to define the loudspeaker gains, formulate the dynamic HOS order matching equations for all orders $n \in [0, N]$ as a set of linear equations. Here, truncation of the expansion to a finite order N is performed, with the assumption that $L \geq N + 1$ to ensure an exact solution to the problem can be calculated. Let $\mathbf{p}_T$ be a length $(N+1)$ vector of target signals, Ψ be an $(N+1) \times L$ plant matrix and $\mathbf{g}$ be a length L vector of loudspeaker gains. The n-th entry of $\mathbf{p}_T$ is the n-th order term of the target signal, given by $\sin^n(\theta_T - \theta_{rot})$. The n-th row and ℓ-th column entry of Ψ is the n-th order contribution due to the ℓ-th loudspeaker, which is $\sin^n(\theta_\ell - \theta_{rot})$. The ℓ-th entry of $\mathbf{g}$ is the gain for the ℓ-th loudspeaker, g_ℓ.

The loudspeaker gains are found by solving the following inverse problem:

$$\mathbf{p}_T = \Psi \mathbf{g} \implies \mathbf{g} = \Psi^{\dagger} \mathbf{p}_T \tag{5.11}$$

where the superscript $(\cdot)^{\dagger}$ indicates the Moore-Penrose pseudoinverse as discussed in Section 3.1.4. By performing this inversion, the loudspeaker panning functions may be calculated for any given head orientation and virtual source position. Furthermore, the loudspeaker gains have no frequency-dependence and therefore form simple panning functions. Importantly, the loudspeaker gains are dependant on the listener head position thus they are dynamic panning functions. Practically, loudspeaker gains can be calculated offline then applied using a look-up table, or can be calculated quickly in real-time using a simple pseudoinverse of the plant matrix and multiplication with the target vector.

5.2 The Instability Condition

HOS takes advantage of the cone of confusion, however this can also cause issues when defining loudspeaker arrays. Consider two loudspeakers with indexes i and j situated at angles θ_i, θ_j respectively. As HOS only requires the contribution of the source across the reproduction axis ($\cos(\theta)$ or $\sin(\theta)$ for the x or y axis respectively), there exists in 2D two separate loudspeaker positions that result in the same soundfield along this axis. If expanded to 3D, these positions become a cone of confusion. Importantly, the soundfield from all of these positions is equal along the reproduction line only, not if evaluated elsewhere. Mathematically this can be seen when the sine or cosine of both loudspeaker angles is equal. When this occurs, the HOS system loses a degree of freedom as the two loudspeakers can not contribute in a unique manner, or equivalently two columns of the plant matrix are identical. Practically this means the number of loudspeakers available to the system is reduced by one, and if the number of unique contributing loudspeakers is less than $N + 1$ then an exact solution can not be achieved and the loudspeaker gain definitions diverge. The instability condition thus occurs for every permutation of loudspeaker pairs within the array and is defined (using the HOS sine representation) as

$$\text{Instability when } \sin(\theta_i - \theta_{rot}) = \sin(\theta_j - \theta_{rot}) \; \forall \; i, j \; \in [1, L]. \tag{5.12}$$

This issue, previously presented for standard HOS, is easily avoidable if the listener's head is fixed with respect to rotation, as the HOS loudspeaker system may be set up such that all loudspeakers can contribute to the problem uniquely. However, issues arise when allowing the listener to rotate their head freely in dynamic HOS,

that is $\theta_{rot} \in [0°, 360°]$. For every pair of loudspeakers there exists a head position that gives rise to the instability, defined as

$$\theta_{rot}^{instability} = \arctan\left(\frac{\sin(\theta_j) - \sin(\theta_i)}{\cos(\theta_j) - \cos(\theta_i)}\right). \tag{5.13}$$

Note that this condition is valid for all permutations of every pair of loudspeakers used by the system, which for a large system can result in many angles where instabilities can occur.

The impact of this issue is that to allow for full 360° listener head rotations a minimum of $2N+1$ loudspeakers are required, not $N+1$. This is because when considering symmetry about an axis, there are two given positions on the unit circle that give the same value when taking the sine (or equivalently the cosine) of those positions. Hence, if N-th order requires $N+1$ loudspeakers, a maximum of $(N/2)$ of these loudspeakers can have symmetric counterparts at any one time. This means the largest possible reduction in the effective size of the loudspeaker array is reduction to size $(N/2)+1$ loudspeakers. To ensure that there is always $N+1$ loudspeakers available to the system (to achieve an exact solution) there consequently must be $L = 2N+1$ loudspeakers if full 360° head rotations are allowed.

This number is the same as for 2D HOA - note that in Section 4.4 a decoder from 2D HOA to HOS was presented that reproduced a subset of the HOA information. This is because, to allow for any head orientation, the system must be able to reproduce along a line in any given direction. The figure is the same as 2D HOA because this is exactly what 2D HOA does - accurate reproduction across the region inside a circle (all possible reproduction lines at the same time). An important difference between the two approaches is that 2D HOA ideally requires these $(2N+1)$ loudspeakers distributed evenly across the circle. However, this restriction does not exist for HOS which can handle more irregular loudspeaker distributions.

To overcome the instability condition careful design of the system and loudspeaker array is required. Tikhonov regularisation can be employed in the pseudoinverse to limit the effects of ill-conditioning that occurs when approaching an instability as described in Section 4.2.5. This allows for the minimal number of $L = N+1$ loudspeakers to still be used, but introduces an error into the solution whilst at an instability position one of the order terms will also not be correctly reproduced. Alternatively, the number of loudspeakers used can be $L > N+1$, therefore allowing for up to $L-N-1$ loudspeakers to become unstable whilst ensuring an exact solution can still be achieved. Furthermore, if a smaller fixed range of head rotations is only required, then the loudspeaker layout can be optimised for that head rotation span to minimise the number of instabilities by using Eqn. 5.13. This might occur naturally, for example, if the listener is watching content on a display screen it is

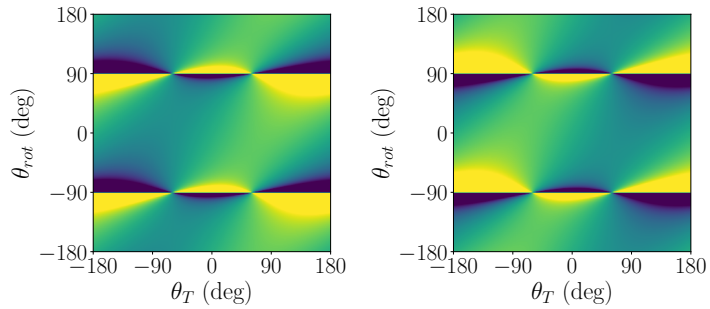
unlikely they will desire to rotate and face away from the screen, thus defining a smaller range of head rotations.

To demonstrate the issue of the instability condition, consider the simplest case of a first order system with two loudspeakers situated at $\theta_1 = -\theta, \theta_2 = -\theta$ - a symmetric left right (LR) pair. As per Eqn. 5.13 there should be two instabilities which occur when $\theta_{rot} = \pm 90^\circ$. The loudspeaker gains for this system across all virtual source incident angles and head rotations, and with $\theta = 60^\circ$, are shown in Fig. 5.2(A) and 5.2(B). It may be clearly seen that at $\theta_{rot} = \pm 90^\circ$ the loudspeaker gains tend towards infinity, thus becoming impracticable to implement. This issue may be avoided by adding a third loudspeaker. Thus, consider a left centre right (LCR) system that simply adds a third loudspeaker at $\theta_3 = 0^\circ$. Using this system but only to first order means the problem is underdetermined as $L > N + 1$, however avoids the instability issues as at $\theta_{rot} = \pm 90^\circ$ there are still two uniquely contributing loudspeakers therefore an exact solution can still be found. These loudspeaker gains are presented in Fig. 5.2(C), 5.2(D) and 5.2(E). Here it is clear that whilst there is still a large increase in the amplitude of the loudspeaker gains around $\theta_{rot} = \pm 90$, the actual instability where the gains diverge to infinity is avoided by adding the third loudspeaker.

5.3 Formulation Of 3D Higher Order Stereophony

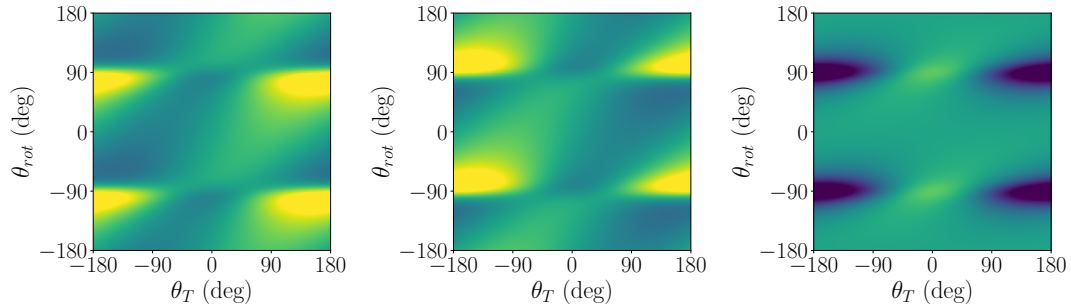
So far the derivation has considered a 2D coordinate system and yaw only head rotations. However, the approach may be extended to 3D by extending the vector notation and the 3D multi-variable Taylor expansion. In this case $\mathbf{r} = [x, y, z]^T$, $p(\mathbf{r}) = e^{j\mathbf{k}_i \cdot \mathbf{r}}$ with $\mathbf{k}_i = k[\cos(\phi_i)\sin(\theta_i), \sin(\phi_i)\sin(\theta_i) + \cos(\theta_i)]^T$. This means elevated sources and loudspeakers may be included in the approach. Due to the assumptions made in the soundfield representation (plane wave propagation, free field conditions), the results derived in Section 4.5 still hold. That is for any possible source/loudspeaker position, a family of alternate positions on a cone of confusion exist that create the same soundfield across the expansion axis only. This means any one of these cone of confusion positions can be alternatively used to create the same soundfield. For example, it may be more practical to use loudspeakers in the horizontal plane only as opposed to elevated positions, which can be mapped to using the cone of confusion.

With the 3D representation, expansion across a single axis ($\hat{\mathbf{n}} = \hat{\mathbf{x}}, \hat{\mathbf{y}}$ or $\hat{\mathbf{z}}$) leads to simplified expressions of the soundfield as in Eqns. 5.2 and 5.3, as well as simple gain definitions where the head rotation compensation is explicit (as will be demonstrated in the following section). When considering generalised 3D head rotations however, the mathematics is not as tractable and the resulting gain definitions are not as simple as in the 2D case. Despite this, the solutions remain panning gains which are implemented in an identical manner, but most easily formulated using vector notation.



(A) First order stereo LR pair, left loudspeaker.

(B) First order stereo LR pair, right loudspeaker.



(C) First order stereo LCR trio, left loudspeaker.

(D) First order stereo LCR trio, right loudspeaker.

(E) First order stereo LCR trio, centre loudspeaker.

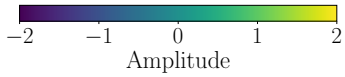


FIGURE 5.2: Loudspeaker gains for two different first order stereo systems. The first system uses a left right (LR) symmetric pair and exhibits the instability condition, whilst the second system uses a left centre right (LCR) trio and does not encounter any instabilities.

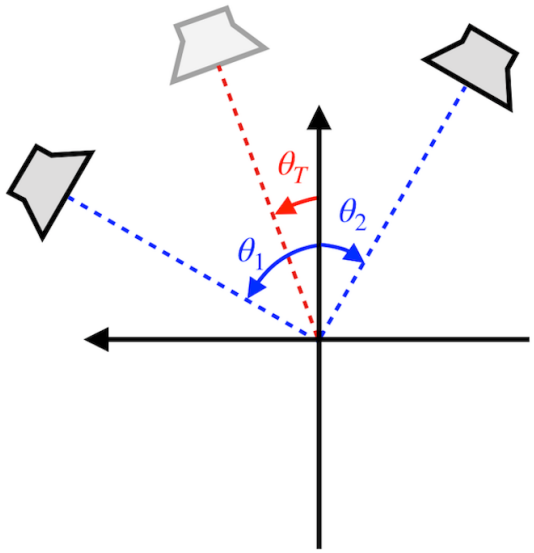


FIGURE 5.3: A generalised first order stereo loudspeaker arrangement.

5.4 Example Higher Order Stereophony Loudspeaker Systems

This section will introduce a generalised first order dynamic HOS system, which results in the definition of classic stereo approaches under certain conditions. The minimum number of loudspeakers required is $L = N + 1 = 2$. Consider two loudspeakers positioned at arbitrary angles θ_1 and θ_2 as in Fig. 5.3, with the virtual source positioned at θ_T. The listener is allowed to rotate their head by θ_{rot}, therefore let all angles be compensated for head rotation as discussed previously by setting $\theta' = \theta - \theta_{rot}$. In this case the target pressure vector and plant matrix are

$$\begin{aligned} \mathbf{P}_T &= \begin{bmatrix} 1 \\ \sin(\theta'_T) \end{bmatrix} \\ \Psi &= \begin{bmatrix} 1 & 1 \\ \sin(\theta'_1) & \sin(\theta'_2) \end{bmatrix} \\ \Psi^+ &= \frac{1}{\sin(\theta'_2) - \sin(\theta'_1)} \begin{bmatrix} \sin(\theta'_2) & -1 \\ -\sin(\theta'_1) & 1 \end{bmatrix} \end{aligned} \tag{5.14}$$

that leads to the loudspeaker gain definitions

$$\begin{aligned} \mathbf{g} &= \frac{1}{\sin(\theta'_2) - \sin(\theta'_1)} \begin{bmatrix} \sin(\theta'_2) - \sin(\theta'_T) \\ -\sin(\theta'_1) + \sin(\theta'_T) \end{bmatrix} \\ &= \frac{1}{\sin(\theta_2 - \theta_{rot}) - \sin(\theta_1 - \theta_{rot})} \begin{bmatrix} \sin(\theta_2 - \theta_{rot}) - \sin(\theta_T - \theta_{rot}) \\ -\sin(\theta_1 - \theta_{rot}) + \sin(\theta_T - \theta_{rot}) \end{bmatrix}. \end{aligned} \tag{5.15}$$

This is the head-tracked stereo sine law for generalised loudspeaker geometries. The head-tracked stereo sine law has been previously derived through the work of Compensated Amplitude Panning (CAP) [12, 13, 14]. Furthermore, the low-frequency approximation of the CTC technique for two loudspeakers leads to the same gain definitions [81, 82, 138].

From this system, classic stereo techniques can be derived under particular conditions. First, consider the scenario when the listener is assumed to face forward ($\theta_{rot} = 0$), this gives the stereo sine law for generalised loudspeaker geometries

$$\mathbf{g} = \frac{1}{\sin(\theta_2) - \sin(\theta_1)} \begin{bmatrix} \sin(\theta_2) - \sin(\theta_T) \\ -\sin(\theta_1) + \sin(\theta_T) \end{bmatrix}. \tag{5.16}$$

Next let the loudspeaker angles be symmetric, such that $\theta_1 = \theta = -\theta_2$, as with a traditional stereo loudspeaker arrangement. For arbitrary head rotations the loudspeaker gains are

$$\begin{aligned} \mathbf{g} &= \frac{1}{\sin(-\theta - \theta_{rot}) - \sin(\theta - \theta_{rot})} \begin{bmatrix} \sin(-\theta - \theta_{rot}) - \sin(\theta_T - \theta_{rot}) \\ -\sin(\theta - \theta_{rot}) + \sin(\theta_T - \theta_{rot}) \end{bmatrix} \\ &= \frac{1}{2} \begin{bmatrix} 1 + \frac{\tan(\theta_{rot})}{\tan(\theta)} + \frac{\sin(\theta_T)}{\sin(\theta)} - \frac{\cos(\theta_T)\tan(\theta_{rot})}{\sin(\theta)} \\ 1 - \frac{\tan(\theta_{rot})}{\tan(\theta)} - \frac{\sin(\theta_T)}{\sin(\theta)} + \frac{\cos(\theta_T)\tan(\theta_{rot})}{\sin(\theta)} \end{bmatrix} \end{aligned} \tag{5.17}$$

where the trigonometric identities $\sin(-x) = -\sin(x)$ and $\sin(A \pm B) = \sin(A)\cos(B) \pm \cos(A)\sin(B)$ have been used. This is the head-tracked stereo sine law [12]. Setting the head rotation to be $\theta_{rot} = 0$ so that the listener is assumed to be facing forward gives the traditional stereo sine law as defined in [7, 8]

$$\mathbf{g} = \frac{1}{2} \begin{bmatrix} 1 + \frac{\sin(\theta_T)}{\sin(\theta)} \\ 1 - \frac{\sin(\theta_T)}{\sin(\theta)} \end{bmatrix}. \tag{5.18}$$

Finally, consider the case when the stereo symmetric loudspeaker angles are used, except now the listener always faces the virtual source such that $\theta_{rot} = \theta_T$. This gives the stereo tangent law as given in [21, 145]

$$\mathbf{g} = \frac{1}{2} \begin{bmatrix} 1 + \frac{\tan(\theta_T)}{\tan(\theta)} \\ 1 - \frac{\tan(\theta_T)}{\tan(\theta)} \end{bmatrix}. \tag{5.19}$$

Hence, using the Taylor expansion of a plane wave soundfield, all of the most established stereo panning techniques have been derived under certain assumptions.

This covers generalised loudspeaker geometries, classic stereo loudspeaker arrangements, generalised listener head rotations and also specific listener head arrangements such as facing forward or facing the virtual source. This shows that these key existing stereo methods are actually first order Taylor approximations of reproducing the actual plane wave target soundfield across a line, with either assumptions about the listener's head orientation or assumed head positions such that the reproduction line coincides with the interaural axis. Thus stereo is actually a soundfield reproduction technique to the first order and, as shown in this work, through the Taylor expansion may be expanded to higher orders for more accurate reproduction of the soundfield. As previously all common stereo techniques have been defined as low frequency methods, HOS therefore both generalises and expands the stereo theory to any given order, for any given loudspeaker array and reproduction across any frequency or spatial range (as restricted by the loudspeaker array and truncation order).

5.5 Relation To Higher Order Ambisonics

Previously, decoders to convert from both 2D and 3D HOA to HOS have been defined in Section 4.4. Here the decoders are reviewed using the same steps as in Section 4.4 and then extended to include head rotations. First, consider reproduction to the N-th order expressed using the HOS representation where the vector of reproduced HOS order terms, $\mathbf{p}_T$, is given by the product of the plant matrix Ψ multiplied by the loudspeaker gains $\mathbf{g}$, all represented using the HOS representation, such that $\mathbf{p}_T = \Psi\mathbf{g}$. Now consider the analogous synthesis problem using N-th order HOA denoted by $\mathbf{p}'_T = \Psi'\mathbf{g}'$. The two representations may be transitioned between by a matrix $\mathbf{A}$ defined for 2D HOA by the Chebyshev polynomials and for 3D HOA by the Legendre polynomials. In this case

$$\begin{aligned} \mathbf{p}'_T &= \mathbf{A}\mathbf{p}_T \\ \Psi' &= \mathbf{A}\Psi \end{aligned} \tag{5.20}$$

As in Section 4.4, the reproduction problems are related by

$$\begin{aligned} \mathbf{g} &= \Psi^{\dagger}\mathbf{p}_T \\ \mathbf{g}' &= \Psi'^{\dagger}\mathbf{p}'_T \\ &= (\mathbf{A}\Psi)^{\dagger}\mathbf{A}\mathbf{p}_T \\ &= \Psi^{\dagger}\mathbf{A}^{-1}\mathbf{A}\mathbf{p}_T \\ &= \Psi^{\dagger}\mathbf{p}_T \\ &= \mathbf{g} \end{aligned} \tag{5.21}$$

using the identify $(\mathbf{AB})^\dagger = \mathbf{B}^\dagger\mathbf{A}^\dagger$ and noting $\mathbf{A}$ is a square matrix. The above is true for the underdetermined case where the loudspeaker gains are a minimum norm solution, which means the solution $\mathbf{g}$ to $\mathbf{p}_T = \Psi\mathbf{g}$ has zero projection onto the null-space of Ψ. In this case, the HOA and HOS representations will lead to the same loudspeaker gains.

In the overdetermined scenario the solution may have non-trivial elements that map to the null-space of Ψ, in this case

$$\begin{aligned} \mathbf{g} - \tilde{\mathbf{g}} &= \Psi^\dagger \mathbf{p}_T \\ \mathbf{g}' - \tilde{\mathbf{g}}' &= \Psi'^\dagger \mathbf{p}'_T \end{aligned} \tag{5.22}$$

with $\tilde{\mathbf{g}}$ the component of the solution that lies on the null space of Ψ. The impact of this is that for the overdetermined case the mapping can not be said to hold.

It is important to note that the transformation through the matrix $\mathbf{A}$ selects a subset of the HOA representation. This means the full soundfield across a circle (2D HOA) or sphere (3D HOA) is not reproduced accurately. Instead, the subset of HOA that considers the soundfield along a single line only is extracted. This reinforces that HOS is indeed soundfield reproduction along a single axis and may be represented using a number of mathematical representations, including the Taylor expansion of a plane wave soundfield, the Chebyshev polynomials and the Legendre polynomials.

Furthermore, before these decoders may be applied a rotation of the soundfield is required to ensure a subset of the original HOA representation can be used to represent the soundfield across the interaural axis only. In 2D, this corresponds to ensuring the cosine terms of the circular harmonics may be isolated through rotating the desired reproduction axis to align with the x axis. In 3D the rotation is so that the desired reproduction line lies along the z axis, to isolate the spherical harmonics with degree $m = 0$ and no azimuthal dependence.

From this, it is clear that the rotation is important for the decoders to be correctly defined. However, the previous formulation of the decoders and HOS utilised a static rotation defined by the assumption that the listener's head does not rotate. Therefore, a simple extension of these decoders to work for dynamic HOS is to make this rotation dynamic, adapting to the head orientation and ensuring the correct line representation of the soundfield is extracted from the full HOA representation to just reproduce across the interaural axis only.

5.6 Comparative Listening Test

To investigate the viability of dynamic HOS and to compare it to the existing standard HOA techniques, a listening test was designed and run to compare the approaches. The MUltiple Stimulus with Hidden Reference and Anchor (MUSHRA)

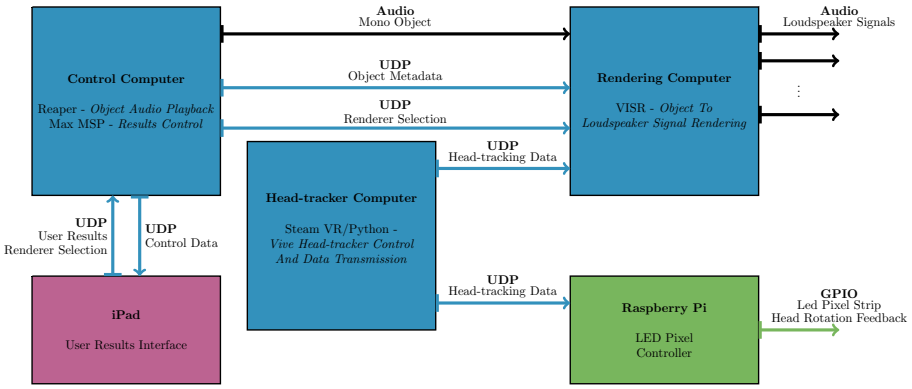
approach was chosen, as this test allows for the comparison of multiple options within each page with relation to a reference source, as well as a hidden reference and hidden anchor to grade the listener's ability to discern changes in the stimuli [146]. The objective of the listening test was to compare dynamic HOS to 2D HOA, to investigate if any perceived difference in the localisation of a virtual source was apparent between the two techniques, including when matching the truncation order or matching the number of loudspeakers available.

5.6.1 Experimental Setup

The listening test was run in the large anechoic chamber at the ISVR to ensure freefield conditions of which both HOS and HOA assume in their derivation. The experimental setup delivered real-time loudspeaker rendering of a single static virtual source about the listener in the horizontal plane. The listener was therefore surrounded by a number of loudspeakers, of which different subsets were activated for different renderers using either HOS or 2D HOA. Real loudspeakers were also placed at the chosen target virtual source positions so the listener could compare the virtual source to a real reference. The listener was seated within a box of acoustically transparent black material, which was also lit from the inside to ensure the loudspeakers could not be seen. Photographs of inside and outside the setup are shown in Fig. 5.4(B) and 5.4(C).

A diagram of the experimental equipment is shown in Fig. 5.4(A). A control computer managed playback of mono audio objects through a Reaper Digital Audio Workstation (DAW) session. The Reaper session hosted the Versatile Interactive Scene Renderer (VISR) object-based audio production plugins to encode the appropriate positional object metadata [147]. Both the audio and object metadata was then sent via MADI and User Datagram Protocol (UDP) to a rendering computer, which ran an instance of the VISR framework which allows for custom audio rendering algorithms to be run utilising python. Here, VISR was used to run a number of different HOS and HOA approaches simultaneously, all which received the audio mono object and outputted the appropriate loudspeaker signals. All renderers were contained in a master switching controller, allowing the listener to switch between each of the rendering approaches as they wished. Triggering a renderer switch led to a 5 ms fade out and 5 ms fade in of the audio using a cosine ramp, as required by the MUSHRA standard [146]. When appropriate, the rendering approaches were dynamically adapted based on the head orientation of the listener. The audio output of the VISR engine was sent via MADI to an RME M32 AD which drove the loudspeakers.

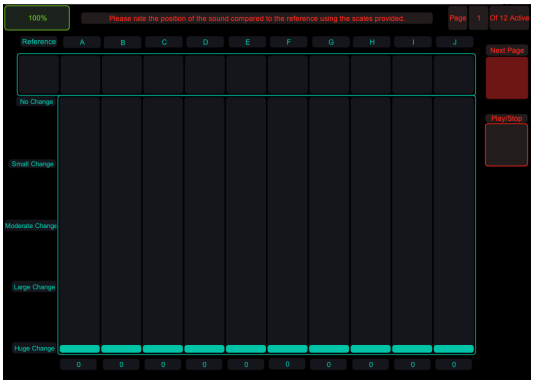
The reproduction loudspeaker array consisted of Genelec 8020C loudspeakers. For the reference source positions, Genelec 8020D loudspeakers had to be used to ensure enough loudspeakers were available for the reproduction array. The 8020D's are an updated model of the 8020C's, and little difference is seen in their frequency



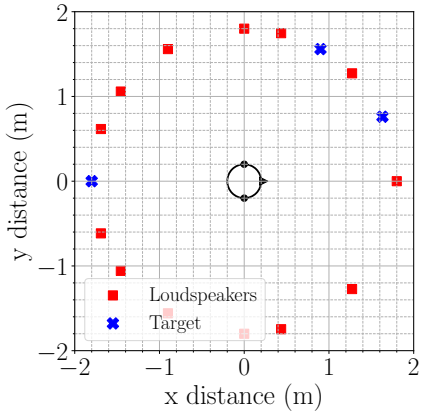
(A) Diagram of the listening test apparatus and signal flows for audio and metadata.

(B) Picture of the listening test setup.

(C) Picture of the view of the subject with the head rotation LED feedback active.



(D) User interface for controlling and rating the different renderer options.



(E) Geometry of the listener, loudspeaker and target source positions.

FIGURE 5.4: Experimental setup for the listening test.

response. However, to ensure there was no level difference due to both the different loudspeaker models for the reference source positions, as well as level differences due to the different type of loudspeaker renderers, a normalisation process was run to ensure consistent volume across all source positions and renderer types to negate any listener bias due to volume differences. Here an omnidirectional microphone was set up in the middle of the loudspeaker array and noise was played through each of the reference target loudspeakers, then each of the renderers reproducing a virtual source at the target positions in turn. The Root Mean Square (RMS) level due to each of these scenarios was recorded, and a gain calibration for each renderer was calculated to ensure they matched the level of the reference sources.

A HTC Vive tracker was chosen for the head-tracking solution, which has found common use in loudspeaker rendering where listener tracking is required and outputs to a high accuracy listener position and orientation data. A dedicated head-tracker computer was used to run Steam VR and python to control the head-tracker and output via UDP the head-tracking metadata to the rendering computer. This was only required for the HOS approaches, which dynamically adapt the loudspeaker gains depending on listener head orientation.

The instability condition can lead to issues for HOS loudspeaker arrays at certain listener head orientations. Therefore, to avoid these issues the head rotation range of the listener was limited. A strip of LED pixels was used to give the listener feedback on their current head orientation as shown in Fig. 5.4(C). A block of blue pixels indicated the two limits the listener could rotate their head within, whilst the rest of the strip flashed red when the listener exceeded these limits. The LED pixel strip was controlled via a Raspberry Pi, which received head-tracking data via UDP from the head-tracker computer.

The test was controlled by the subject using an iPad running TouchOSC, of which a screenshot of the user interface is shown in Fig. 5.4(D). Here the audio could be played/stopped as required, the active renderer could be switched and each renderer could be rated using a slider. Finally, once all renderers were assessed, the next page of the test could be triggered. The iPad communicated through UDP with the control computer where a Max MSP patch acted as a hub for the control metadata, including saving the subject's renderer ratings and sending the requested audio renderer selection to the rendering computer. At all stages the audio processing was performed at a sample rate of 48 kHz and a block size of 512 samples.

5.6.2 Experimental Design

The MUSHRA approach was chosen for the listening test as it allows for simultaneous comparison of a large number of stimuli with respect to a reference sound source. Furthermore, inclusion of a hidden reference and anchor allows rating of the reliability and quality of each subject's results.

Renderer	Order	Num. Loudspeakers	Loudspeaker Positions (°)
HOS	1	2	± 45
HOS	1	2	± 90
HOS	2	3	$0, \pm 90$
HOS	3	4	$\pm 45, \pm 90$
HOS	4	5	$0, \pm 45 \pm 90$
HOS	5	6	$\pm 45, \pm 76, \pm 90$
HOA 2D	2	5	$0, \pm 45, \pm 75, \pm 144$
HOA 2D	4	9	$0, \pm 45, \pm 76, \pm 120, \pm 160$

TABLE 5.1: Spatial audio techniques and their configuration used in the listening test.

The renderers chosen for comparison are given in Table 5.1. These approaches were chosen to investigate the following research questions:

- Does increasing the truncation order for HOS increase the accuracy of the perceived virtual source position?
- Can HOS reproduce virtual images out-of-span of the loudspeaker array or behind the listener, using loudspeakers in front only?
- Is there any perceived difference between HOS and HOA renderers of the same order, but different number of loudspeakers?
- For a fixed number of loudspeakers, is there any perceived improvement when using a higher order HOS approach compared to HOA?

Therefore, in an effort to answer these questions, HOS systems of increasing order from 1 to 5 (HOS O1-5) were used. Furthermore, two HOA systems of order 2 and 4 (HOA O2 and O4) were employed to compare to HOS at the same truncation order. Finally, to consider the scenario when a fixed number of loudspeakers is allowed, HOS O4 and HOA O2 both using the minimum requirement of 5 loudspeakers were included for comparison.

The loudspeaker positions for each system are given in Table 5.1 and the final loudspeaker array used by all systems is shown in Fig. 5.4(E). All loudspeakers were positioned at head height and equidistantly, 1.8 m away from the listener. This distance was maximised as the space allowed so that the sources approximated plane waves for as low a frequency limit as possible - this was because both HOS and HOA make a plane wave assumption in their derivation. All renderers used a small amount of Tikhonov regularisation ($\beta = 0.01$) in the plant matrix inversion to define the loudspeaker signals, to help stabilise the solutions in the presence of experimental errors (for example small errors in loudspeaker positions).

For HOS, at each order $(N + 1)$ loudspeakers were required and chosen by sampling a semicircle in front of the listener (maximum position $\pm 90°$). To consider whether the loudspeaker span affected the performance, two HOS O1 systems with

varying spans of $\pm 45^\circ$ and $\pm 90^\circ$ (HOS O1-45 and HOS O1-90 respectively) were included. Thus for the two matching order systems, a virtual source position larger than 45° meant an out-of-span source could be compared to an in-span source. To investigate whether increasing order does increase virtual source position accuracy, an increasing number of loudspeakers are required for each higher order system. Therefore, to isolate whether a difference in performance was just due to using more loudspeakers or to increasing the order of the reproduction, HOS O1-45, O3 and O5 were designed to add loudspeakers outside of $\pm 45^\circ$ only as order increased. Therefore, a virtual source positioned within $\pm 45^\circ$ could assess this hypothesis without being biased by a new loudspeaker position being added closer to the virtual source position.

The loudspeaker positions for the HOA systems were chosen by equally sampling a circle around the listener. The minimum number of loudspeakers, $(2N + 1)$, was used. However, some loudspeaker positions had to be shared across renderers due to a limit in the number of available loudspeakers as well as their physical size restricting how close any two loudspeakers could be positioned. Therefore, for HOA O2 two positions should have been $\pm 72^\circ$ and instead were $\pm 76^\circ$. Furthermore, for HOA O4 loudspeakers should have been positioned at $\pm 40^\circ, \pm 80^\circ$ and instead were at $\pm 45^\circ, \pm 76^\circ$. These small shifts in position should theoretically have minimal impact of the performance of the approaches, as they are very close to optimal sampling positions.

The highest order system used was order 5, as per the $N = kr$ rule this corresponds to a frequency limit of approximately 3500 Hz above which aliasing will occur for the highest order system, assuming a standard head radius of $r = 0.08$ metres. Furthermore, at high frequencies an intensity panning approach is often favoured over mode matching techniques [42]. Therefore, a lowpass filter was applied to all audio for all approaches with a cut-off at 4000 Hz. The reference source was chosen as real loudspeakers at the same position as the virtual sound sources, thus the participants compared a real source source to a virtual one. The anchor was designed to be both spatially and tonally impeded. Therefore, the anchor used an equal sampling of loudspeakers positioned around the listener, here corresponding to the HOA O4 array, however with all loudspeakers active with equal gain to create a large ambiguity in the position of the sound source. The anchor was also low-passed with a filter as defined in [146], however the cut-off frequency was set to 1000 Hz. This was lower than the standards require as all audio had already been lowpassed, therefore the anchor was redefined to ensure it was sufficiently altered from the other signals.

Three different source stimuli were tested at three different positions. Therefore, there were 9 pages in the test. The order of these pages was randomised for all subjects and on average the test took one hour including a break in the middle. The stimuli were looped anechoic recordings of a male speech sample, a short rock drum beat and a flamenco acoustic guitar recording from [148]. Drums have been shown

to be a particularly appropriate critical signal for listening tests, combining both broadband content and sharp transients [149]. The source positions used may be viewed in comparison to the reproduction loudspeaker positions in Fig. 5.4(E) and were

$$\theta_1 = 25^\circ, \qquad \theta_2 = 60^\circ, \qquad \theta_3 = 180^\circ. \tag{5.23}$$

θ_1 meant the effect of increasing HOS order without adding loudspeakers close to the virtual source could be assessed for HOS O1-45, O3 and O5 as explained previously. The source at θ_2 meant in-span and out-of-span sources could be compared for HOS O1-45 and O1-90 while the rear position of θ_3 was chosen to see whether any HOS system could create the illusion of a rear virtual source whilst using frontal loudspeakers only.

The participants were asked to rate the position of the virtual source created by each of the renderers, in comparison to the real reference sound source. The participants were encouraged to consider the following positional properties of the virtual source compared to the reference - absolute position, apparent source width and stability with head rotations. The rating was on a scale from $0 - 100$ with labels as shown in Fig. 5.4(D).

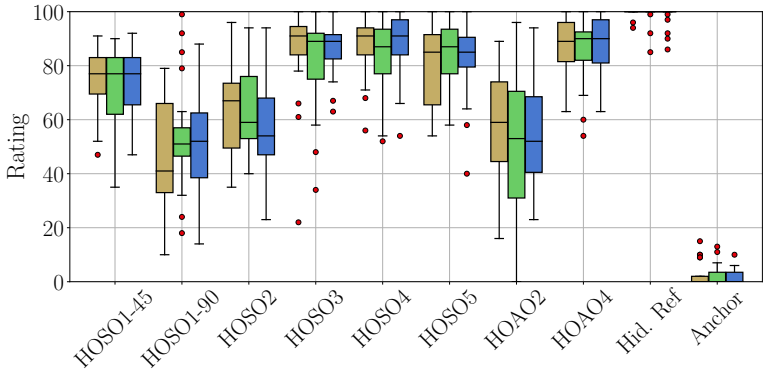
Head rotations are key to ensuring dynamic HOS performs properly, therefore the participants were strongly encouraged to rotate their head through-out the test and every time they selected a new renderer. To minimise the effect of the instability condition for these given HOS loudspeaker arrays, the participants head rotation was limited to within a $\pm 25^\circ$ span with the LEDs used to give the subject feedback when they exceeded this value.

As required by the MUSHRA standard, a training phase was first run before the test to introduce the participants to the test interface, the range of variation in sounds they would encounter and how they might translate this onto the provided scale. This training phase used a representative set of the renderers, including HOS O1-45, HOS O5, HOA O2, HOA O4, the reference and the anchor. The training phase was three pages, using the three source stimuli at the three positions.

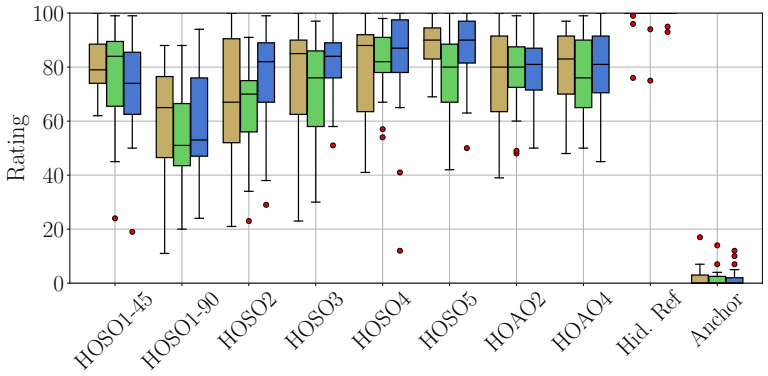
5.6.3 Results

24 subjects took part in the listening test, of which 21 were between the ages of 22-50 and 3 were over 50. 18 of the subjects were male and 6 female. All subjects had self-reported normal hearing. A single participant's results were discarded as they did not identify the hidden reference correctly enough times as per the MUSHRA standard.

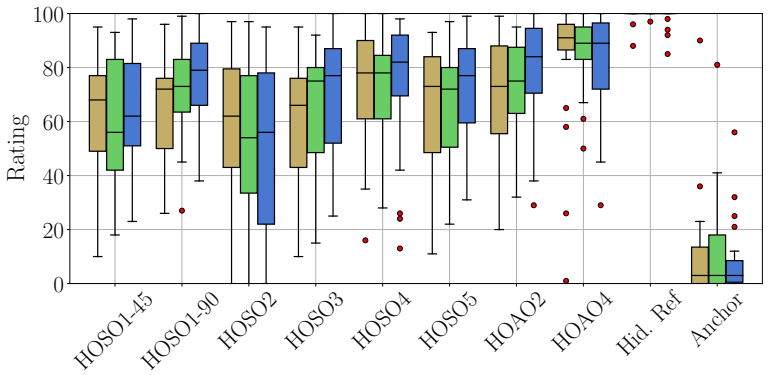
The raw data from the subject's responses is shown as boxplots in Fig. 5.5. These figures show the subjects ratings of all of the different rendering approaches, for the three different source signals and three virtual source positions. It is clear that



(A) Results for a source at $\theta_1 = 25°$.



(B) Results for a source at $\theta_2 = 60°$.



(C) Results for a source at $\theta_3 = 180°$.

FIGURE 5.5: Raw results from the listening test in the form of boxplots considering changing the renderer type and signal type for each of the three virtual source positions. The box represents the upper and lower quartiles, with the intersecting line within the box the median value. The whiskers represent the limits to identify outliers which are represented as red circles. The brown, green and blue boxes represent the speech, drums and guitar signal types respectively.

regardless of the source position or source stimuli, the hidden reference and anchor were both consistently identified and scored appropriately. One interesting deviation is for the anchor at the rear source position of 180°, which was scored higher than expected by some subjects. This is likely due to the confusing nature of the soundfield created by the anchor approach, with some participants reporting a larger concentration of energy arriving from behind them. Despite this anomaly, the impact on the overall experimental design is minimal as the overall anchor rating remained very low as required.

Overall, considering the results there appears to be a trend that increasing the truncation order of the HOS system leads to a higher rating. The HOS O1-45 system however does not follow this pattern, and scores consistently higher than might otherwise be expected. This is observable most clearly for $\theta_1 = 25°$ and $\theta_2 = 60°$. However, for the rear source $\theta_3 = 180°$, the ratings across all participants are more variable, signified by the large boxplots indicating a large spread in the ratings for many of the techniques. In general, there are few observable trends when considering the results for different types of source stimuli.

Repeated measures ANalysis Of VAriance (ANOVA) was utilised for statistical analysis of the results. First, a Kolmogorov-Smirnoc normality test was applied to the residuals of the dataset. The test rejected the null hypothesis for 50 out of the 72 tests at a significance level of $p = 0.05$. Whilst the majority of conditions passed the normality test, 22 did not which indicates some degree of non-normality. However, approaches such as ANOVA are fairly insensitive to such violations [146].

Within-subject effects were tested for every combination of the three independent variables (renderer approach, source stimuli and source position). All but two of these combinations (source stimuli, source position) violated sphericity therefore for all combinations except these two the Huynh-Feldt corrected values are reported. The results of the three-way repeated measures ANOVA considered to a 5% significance level are detailed in Table 5.2.

The only combinations found to not be significant were the technique $\times$ stimuli and the stimuli $\times$ position combinations, importantly suggesting that all techniques were rated similarly regardless of the type of playback signal. All main effects were found to be significant, for technique $[F(6.619, 101.62) = 35.449, p < 0.001]$, source stimuli $[F(2, 44) = 4.158, p = 0.022]$ and source position $[F(2, 44) = 12.364, p < 0.001]$. First order interactions that were significant included technique $\times$ position $[F(8.432, 185.512) = 16.125, p < 0.001]$ which suggests that the different renders performed variably depending on where the virtual source was located, which is particularly noticeable in the boxplots when comparing $\theta_1 = 25°$ to $\theta_3 = 180°$.

Mean ratings of the scores for each technique are given in Fig. 5.6. The overall means in Fig. 5.6(A) demonstrate that increasing the order of the HOS approach leads to a higher overall rating. Two exceptions appear however, the first is that the HOS O1-45

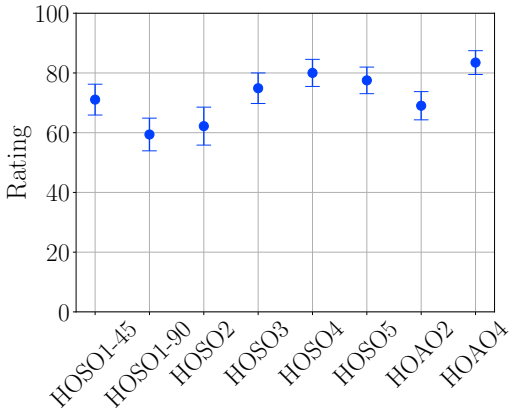
Effect	df	df Error	F	p
Technique *	6.619	101.62	35.449	< 0.001
Stimuli	2	44	4.158	0.022
Position	2	44	12.364	< 0.001
Technique $\times$ Stimuli *	12.489	274.761	0.985	0.465
Technique $\times$ Position *	8.432	185.512	16.125	< 0.001
Stimuli $\times$ Position *	3.166	69.658	1.55	0.207
Technique $\times$ Stimuli $\times$ Position *	20.064	441.409	1.644	0.04

TABLE 5.2: Within-subject effects from the three-way repeated measures ANOVA. An asterisk indicates a condition with the Huynh-Feldt correction applied.

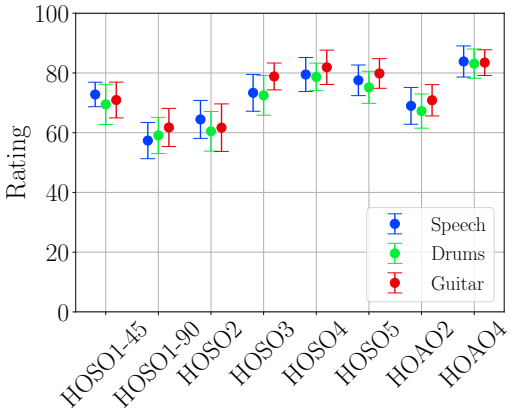
performs stronger than expected and is on par with the O5 version of the technique. Secondly, a slight reduction in score is seen when increasing from HOS O4 to O5. As expected, the HOA O4 scored higher than HOA O2, however was also rated the highest of all the renderers, performing slightly better than the HOS techniques of a similar order although the scores remain within confidence intervals of each other. Furthermore, when comparing between HOA O2 and HOS O4 which both use the same number of loudspeakers, the HOS approach scores higher suggesting there is a benefit to using the new technique.

Very similar trends are seen when considering the means of each technique with respect to varying the source stimuli, as in Fig. 5.6(B), as here it is apparent the scores for each technique are in agreement across all types of source stimuli. This is not the case when considering the means across technique and virtual source positions in Fig. 5.6(C). Here, it is apparent that for every renderer assessed there are considerable variations in their scores depending on the source position, including variability in the scores indicated by the increased size of the confidence intervals for certain technique-position combinations. Notably, comparing the $\theta_1 = 25^\circ$ position for HOS O1-45 and O3 where loudspeakers were added outside of $\pm 45^\circ$ to increase the order, an increase in the score is observed showing that increasing the order does indeed increase the accuracy of the source localisation. Furthermore, HOS O1-45 scored significantly higher than HOS O1-90 except for the rear position. Overall, the HOS renderers perform poorest for the rear position at $\theta_3 = 180^\circ$ although a larger range is seen due to the size of the confidence intervals at these positions. Regardless, both increasing the order aids the performance and a significant rating is still observed, showing that even with loudspeakers just in front of the listener a rear virtual source could be achieved using HOS.

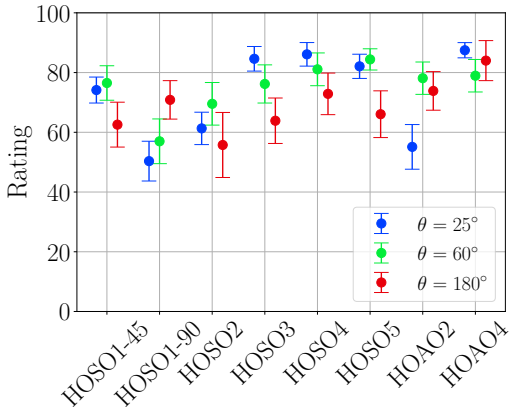
Finally, post-hoc paired comparisons with a Bonferroni adjustment were also performed. Across the technique comparisons the majority of the tests showed a significant difference for each given combination of the renderers. Although notably comparing each of HOS O3/O4 to HOS O5 resulted in no significant difference. A comparison of the HOA and HOS renderers of matching order (for order 2 and 4) also showed no significant difference and likewise with a comparison of HOS O1-45



(A) Overall mean ratings.



(B) Mean ratings for each source stimuli (speech, drums and guitar)



(C) Mean ratings for each of the three source positions ($\theta = 25°, 60°, 180°$).

FIGURE 5.6: Mean ratings for each of the rendering approaches, with 95% confidence intervals indicated by the whiskers.

to HOA O2. The post-hoc tests for source stimuli were all insignificant except for the drums and guitar comparison, whilst for the source position all comparisons were significant verifying the trends seen in Fig. 5.5 and 5.6(C) that different positions lead to significantly different ratings from the participants.

5.6.4 Discussion

The listening test was designed to answer a number of questions comparing different versions of HOS alongside a direct comparison to HOA. These covered whether increasing the truncation order of HOS increased the accuracy of the virtual source position, including in-span versus out-of-span and rear positions, as well as comparing HOS to HOA when the truncation order or the number of loudspeakers is fixed.

From the results it is clear that there is a trend between increasing the order of the truncation and the perceived localisation of the virtual source compared to the real reference. This verifies that HOS is indeed the higher order extension of classical stereo, in a similar manner to HOA. This trend remains across different source material and source positions, and is particularly proven by considering the $\theta_1 = 25^\circ$ position for HOS O1-45 and O3. Here when increasing the order between these two systems it was ensured that no new loudspeakers were added closer to the virtual source position, with the higher order system still scoring higher. However, a drop in performance is observed when moving from O4 to O5 which suggests a perceptual limit beyond which, for this scenario, a higher order is not advantageous. Furthermore, a first order system with a small span, HOS O1-45, performed considerably better than expected except for rear virtual source positions, in some cases outperforming higher order systems. This suggests that adding the dynamic version of the technique to readily available classic stereo arrangements could be advantageous, and in the results of the test the performance was not statistically different to the HOA O2 system.

Moreover the data demonstrates that when using the dynamic version of HOS with loudspeakers situated only in front of the listener, convincing virtual sources can also be placed in the rear, albeit with some reduction in accuracy compared to frontal sources. This is a significant result as loudspeaker arrays that fully enclose the listener as required for optimal performance of HOA can be impractical. Nevertheless HOS comes at the cost of requiring dynamic rendering and a head-tracker implementation.

Comparing HOS and HOA implementations at matching order (O2 and O4 for both techniques) showed that there was no statistical difference between the two rendering approaches. Notably, this means dynamic HOS presents an alternative to HOA requiring a smaller number of loudspeakers arranged only in front of the listener, however rendering virtual sources to a similar level of accuracy. Furthermore in

the scenario where a subject may have a fixed number of loudspeakers available, for example only 5 loudspeakers, a HOS O4 implementation would perform better than the HOA O2 alternative with regards to virtual source localisation, as well as requiring frontal loudspeakers only.

Therefore, in general HOS has been shown to be a higher order extension of classical stereo, in a similar manner to trends in HOA. In comparison to HOA, less loudspeakers are needed in potentially more practical arrangements, however HOS requires a dynamic head-tracked rendering implementation.

5.7 Chapter Review

This chapter has introduced a dynamic head-tracked extension of HOS, allowing a listener to rotate their head freely resulting in adaptive loudspeaker panning to maintain consistent virtual source imaging with a minimum of only $(N+1)$ loudspeakers for N-th order reproduction. The dynamic extension of the technique is formulated mathematically by rotating the definition of the loudspeaker and virtual source positions by the head rotation angle, which means the loudspeaker gains remain as panning functions like with the original HOS implementation. Furthermore, the approach was demonstrated to be a higher order extension of classic stereo techniques by deriving and showing that the generalised head-tracked sine law, the stereo sine law and the stereo tangent law are all forms of first order HOS systems with specific head orientation assumptions.

An issue called the instability condition arises from the dynamic version of HOS, which occurs when two loudspeakers fall on the same cone of confusion. This is a particular issue when the listener is allowed to rotate their head a full 360°, and leads to requiring for N-th order ideally $(2N+1)$ as opposed to only $(N+1)$ loudspeakers. A simple formula for evaluating at what head rotations a loudspeaker array will become unstable is presented. Techniques to overcome the instability condition include limiting the range of the listener's head rotation, adding more loudspeakers and also regularisation techniques.

HOS has been demonstrated to be intrinsically linked to both 2D and 3D HOA through the existence of decoders between the techniques. The expansion of these decoders to account for dynamic listener head rotations was presented. This extension consists of a dynamic rotation which accounts for the listener orientation and then ensures the interaural axis of the listener is always aligned with a specified axis (depending on which version of HOA is used).

Finally, a listening test was designed to compare a number of HOS and 2D HOA systems subjectively. The MUSHRA listening test was performed in anechoic conditions where participants were asked to compare the localisation of a virtual source in comparison to a real sound source in the desired position. The results demonstrate

that increasing the order of HOS generally results in higher accuracy of the perceived virtual source location. HOS and HOA systems of matching order were shown to be generally indistinguishable, which is advantageous as HOS requires both less loudspeakers and positioning of the loudspeakers in front of the listener only. However, whilst HOS could reproduce rear virtual source positions with just loudspeakers in the front, generally HOA did perform better in these regions. Furthermore, a simple first order HOS implementation using a classic stereo loudspeaker pair performed beyond expectations suggesting that dynamic HOS could bring significant advantages if implemented on this readily available and standard loudspeaker setup.

Chapter 6

Binaural Rendering Using Higher Order Stereophony

In the previous chapters, HOS has been considered as a soundfield reproduction technique along a line with no inclusion of the acoustical effects of a HRTF. The assumption has been made that aligning the reproduction axis with the listener's interaural axis will lead to sufficiently accurate reproduction of the desired binaural signals. This has been shown to be an acceptable assumption at low frequencies [82, 138]. However for high frequencies, when the inclusion of the HRTF becomes more significant (for example due to head shadowing), the validity of HOS requires further investigation. Furthermore, the target virtual source has so far been restricted to that of a single plane wave object. In reality, the goal in spatial audio is often to reproduce a more complex soundfield.

In this chapter, the Plane Wave Decomposition (PWD) is used to expand the HOA technique from a single plane wave target source to a generalised soundfield, defining the B-format representation. Next, the classic HOA mode matching equations are derived and HOA approaches for binaural rendering revised for the case of a generic HRTF and the special case of a rigid sphere HRTF. Following this, similar results are derived when using HOS as the rendering approach. This establishes an efficient binaural rendering technique using HOS, reinforces its link to HOA and the spherical harmonic representation and also extends the technique to reproduction of general soundfields. HOS is shown to fully recreate the rigid sphere HRTF using only $(N+1)$ channels and the approach is verified for generic HRTF rendering. Finally, binaural renderers using HOS and HOA of varying orders are compared using numerical simulations of reproduced HRTFs as well as a headphone listening test.

6.1 Higher Order Ambisonics

6.1.1 Soundfield Representation

For HOA using loudspeaker arrays (including virtual loudspeaker array rendering) gain definitions are generally derived considering reproduction of a single virtual

plane wave sound source, which results in a set of panning functions. Extensions such as directly encoding nearfield sources have been investigated [53] whilst the representation also allows the encoding of microphone array signals [150]. Considering the foundational representation of just plane wave sources, the technique may be generalised through the concept of linearity to include any soundfield that satisfies the homogeneous Helmholtz equation and may be expressed as a summation of plane waves, titled the Plane Wave Decomposition (PWD) [150, 134]. Here, the PWD will be used to formulate the HOA B-format representation and then later with the HOS technique.

A soundfield, $p(\mathbf{r}, \omega)$, at position $\mathbf{r}$ and angular frequency ω with wavenumber k that satisfies the 3D homogeneous Helmholtz equation may be represented through the PWD [118]. Here, the soundfield is decomposed into a linear summation of plane waves impinging from all possible directions $\hat{\mathbf{x}}$ under free field conditions. Therefore the PWD is defined as [134, 151]

$$p(\mathbf{r}, \omega) = \int_{\Omega} q(\hat{\mathbf{x}}, \omega) e^{jk\mathbf{r}\cdot\hat{\mathbf{x}}} d\Omega(\hat{\mathbf{x}}). \tag{6.1}$$

Here $q(\hat{\mathbf{x}}, \omega)$ is the *plane wave density*, a function that gives the amplitude and phase of each plane wave from the direction $\hat{\mathbf{x}}$, where $\hat{\mathbf{x}}$ scans across the entire $\mathcal{S}^2$ unit sphere. Integration over $\mathcal{S}^2$ is given by $\int_{\Omega} d\Omega = \int_0^{2\pi} \int_0^{\pi} \sin\theta d\theta d\phi$. For the PWD free field conditions are assumed thus the propagation of each plane wave is simply $e^{jk\mathbf{r}\cdot\hat{\mathbf{x}}}$. The PWD is also known as the Hergoltz Wave Function (HWF) in the literature [118][1]. A similar function here titled the Generalised Plane Wave Decomposition (GPWD) expresses the same scenario but for non free field conditions. Here the *kernel*, $H(\mathbf{r}, \hat{\mathbf{x}})$, is introduced which gives the propagation characteristics of each plane wave from each direction $\hat{\mathbf{x}}$, hence the kernel is dependent on the acoustical conditions of the reproduction medium. This allows for the consideration of more complex acoustic scenarios, for example the presence of a scattering object such as a listener's head. The GPWD is defined as

$$p(\mathbf{r}, \omega) = \int_{\Omega} q(\hat{\mathbf{x}}, \omega) H(\mathbf{r}, \hat{\mathbf{x}}, \omega) d\Omega(\hat{\mathbf{x}}). \tag{6.2}$$

The GPWD encompasses all solutions for the PWD as is seen by setting the kernel $H(\mathbf{r}, \hat{\mathbf{x}}) = e^{jk\mathbf{r}\cdot\hat{\mathbf{x}}}$ thus assuming free field conditions.

Continuing with the standard free field PWD and with the inclusion of the spherical harmonic expansion of a plane wave, Eqn 3.12, then

[1]Note that in the HWF literature, the density is also sometimes called the kernel. In this work we follow the notation of generalised integral representations with the density and kernel separate variables.

$$p(\mathbf{r},\omega) = \int_{\Omega} q(\hat{\mathbf{x}},\omega) \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi j^n j_n(kr) Y_n^m(\hat{\mathbf{r}}) Y_n^m(\hat{\mathbf{x}})^* d\Omega(\hat{\mathbf{x}}) \tag{6.3}$$

where $j_n(kr)$ is the n-th order spherical Bessel function. Part of this equation may be recognised as a $\mathcal{SHT}$, thus define

$$q_n^m(\omega) = \int_{\Omega} q(\hat{\mathbf{x}},\omega) Y_n^m(\hat{\mathbf{x}})^* d\Omega(\hat{\mathbf{x}}) \tag{6.4}$$

which are the *B-format* signals. This set of $q_n^m(\omega)$'s are the spherical harmonic coefficients of the PWD, which in turn represent the given soundfield. This representation is advantageous due to its ease in representing generalised soundfields. Furthermore, the decomposition onto the spherical harmonic basis means manipulations of the soundfield may be performed, for example rotations, warping or focusing [1, 152, 153]. Using the B-format definition then

$$p(\mathbf{r},\omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi j^n q_n^m(\omega) j_n(kr) Y_n^m(\hat{\mathbf{r}}). \tag{6.5}$$

Multiply both sides by a dummy variable $Y_{n'}^{m'}(\hat{\mathbf{r}})^*$ and integrate over $\mathcal{S}^2$

$$\int_{\Omega} p(\mathbf{r},\omega) Y_{n'}^{m'}(\hat{\mathbf{r}})^* d\Omega(\hat{\mathbf{r}}) = \int_{\Omega} \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi j^n q_n^m(\omega) j_n(kr) Y_n^m(\hat{\mathbf{r}}) Y_{n'}^{m'}(\hat{\mathbf{r}})^* d\Omega(\hat{\mathbf{r}}). \tag{6.6}$$

Exploiting the orthogonality of the spherical harmonics on the right hand side (Eqn. 3.4), whilst recognising the left hand side as a $\mathcal{SHT}$, the spherical harmonic coefficients of the reproduced soundfield may be expressed as

$$p_n^m(\omega) = 4\pi j^n j_n(kr) q_n^m(\omega) \tag{6.7}$$

where $p_n^m(\omega)$ is the n-th order and m-th degree spherical harmonic coefficient of the soundfield. The soundfield will be fully recreated by

$$p(\mathbf{r},\omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} p_n^m(\omega) Y_n^m(\hat{\mathbf{r}}). \tag{6.8}$$

This shows there is a one-to-one correspondence between the spherical harmonic coefficients of the reproduced soundfield and the spherical harmonic coefficients of the plane wave density (the B-format signals).

The classical HOA approach is to recreate the desired soundfield by correctly reproducing the set of q_n^m B-format signals. This is normally performed by decoding to a loudspeaker array (which may be a physical set of loudspeakers or for binaural reproduction a virtual array to be synthesised using HRTFs), as reviewed in Chapters 2 and 3. Whilst Chapter 3 derived the HOA mode matching approach for a single virtual plane wave source, decoding of B-format signals will now be derived. Consider a spherical distribution of L radially equidistant loudspeakers acting as plane waves with the ℓ-th loudspeaker positioned at $\mathbf{r}_\ell$ and driven by a signal g_ℓ. Eqn. 6.8 defines the target soundfield the loudspeaker array aims to reproduce, $p_T(\mathbf{r}, \omega)$. The reproduced soundfield, $p_R(\mathbf{r}, \omega)$, is that created by the loudspeaker array and thus a summation of L plane waves. Therefore

$$\begin{aligned} p_T(\mathbf{r}, \omega) &= \sum_{n=0}^{\infty} \sum_{m=-n}^{n} p_n^m(\omega) Y_n^m(\hat{\mathbf{r}}) \\ p_R(\mathbf{r}, \omega) &= \sum_{\ell=1}^{L} g_\ell \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi j^n j_n(kr) Y_n^m(\hat{\mathbf{r}}_\ell)^* Y_n^m(\hat{\mathbf{r}}) \end{aligned} \tag{6.9}$$

where the 3D Jacobi-Anger expansion is used to expresses each plane wave as summation of spherical harmonics as per Eqn. 3.12. Equating the target and reproduced soundfields $p_T(\mathbf{r}, \omega) = p_R(\mathbf{r}, \omega)$, multiplying by a dummy variable $Y_{n'}^{m'}(\hat{\mathbf{r}})$, exploiting the orthogonality of the spherical harmonics and removing common terms leads to the classic set of HOA mode matching equations

$$q_n^m(\omega) = \sum_{\ell=1}^{L} g_\ell Y_n^m(\hat{\mathbf{r}}_\ell)^* \;\; \forall \; n \; \in \; \mathbb{N}_0 \text{ and } m \; \in \; [-n, n]. \tag{6.10}$$

Solutions for the loudspeaker gains that reproduce the desired soundfield are found by formulating Eqn. 6.10 as a set of linear equations for all indices n and m then inverting the plant matrix, formed by sampling the spherical harmonics at the loudspeaker positions [43, 48]. Equivalently, synthesis of the reproduced soundfield can be performed in the spherical harmonic domain as seen by combining Eqn. 6.10 and Eqn. 6.8.

For practical implementation truncation of the expansion to a finite order N is performed, which requires $(N+1)^2$ spherical harmonic channels/coefficients and a minimum of $L = (N+1)^2$ loudspeakers for accurate order truncated reproduction. Thus in free field conditions correct reproduction of the B-format signals up to order N fully recreates the desired soundfield to the same order N.

6.1.2 Generalised HRTF Rendering

This representation of the soundfield will now be used to consider rendering with the inclusion of a HRTF centred at the reproduction point. To include the scattering from the HRTF the GPWD is utilised and for simplicity analysis is performed at the position of the left and right ears only (subscripts l, r). Thus the kernel is defined as the spherical harmonic representation of the HRTF

$$H(\mathbf{r}_{l,r}, \hat{\mathbf{x}}, \omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} h_n^m(\mathbf{r}_{l,r}, \omega) Y_n^m(\hat{\mathbf{x}})^* \tag{6.11}$$

where $h_n^m(\mathbf{r}_{l,r}, \omega)$ is the order n and degree m spherical harmonic coefficient of the HRTF [106]. In this case the GPWD evaluated at the left and right ear positions becomes

$$\begin{aligned} p(\mathbf{r}_{l,r}, \omega) &= \int_{\Omega} q(\hat{\mathbf{x}}, \omega) \sum_{n=0}^{\infty} \sum_{m=-n}^{n} h_n^m(\mathbf{r}_{l,r}, \omega) Y_n^m(\hat{\mathbf{x}})^* d\Omega(\hat{\mathbf{x}}) \\ &= \sum_{n=0}^{\infty} \sum_{m=-n}^{n} h_n^m(\mathbf{r}_{l,r}, \omega) q_n^m(\omega) \end{aligned} \tag{6.12}$$

defining the B-format signals as before through the recognition of a $\mathcal{SHT}$. The binaural signals can be rendered by multiplication of the B-format and HRTF coefficients in the spherical harmonic domain. It is notable that rendering using a loudspeaker array is equivalent to rendering by multiplication in the spherical harmonic domain if the sampling points of the HRTF equal those of the loudspeakers [57]. This is apparent through the inverse of Eqn. 6.11

$$h_n^m(\mathbf{r}_{l,r}, \omega) = \int_{\Omega} H(\mathbf{r}_{l,r}, \hat{\mathbf{x}}, \omega) Y_n^m(\hat{\mathbf{x}})^* d\Omega(\hat{\mathbf{x}}) \tag{6.13}$$

which states that sampling the HRTF by the set of positions $\Omega(\hat{\mathbf{x}})$ leads to the definition of the HRTFs spherical harmonic coefficients. This is equivalent to rendering using a loudspeaker array with a loudspeaker at each position $\hat{\mathbf{x}}$. This fact holds when considering the discrete scenario as is the case for practical rendering, as long as the set of (now finite and discrete) positions $\Omega(\hat{\mathbf{x}})$ is the same for both approaches. In this scenario the HRTF spherical harmonic coefficients are now estimated using the finite set of sampled HRTF positions (each $\hat{\mathbf{x}}$).

It is notable that the B-format signals are identical to those defined using free field conditions, therefore the HOA loudspeaker gains derived using the mode matching approach are valid whether assuming free field or the effects of the HRTF.

6.1.3 Rigid Sphere HRTF Rendering

In the special case where the HRTF is that of the analytical model of a rigid sphere for plane wave sources, the definition of the HRTFs spherical harmonic coefficients is explicit. This HRTF models the human head as a symmetric rigid sphere, with the ears situated at two diametrically opposed positions on either side of the head [95, 133, 154]. Despite its simplicity it remains a commonly used HRTF model, for example in HRTF personalisation and parameterisation [155, 156], as well as crosstalk cancellation [157]. No model of the pinna is taken into account, therefore the rigid sphere is useful up to approximately 4000 Hz, above which the pinna effects become particularly prominent. The advantage of the rigid sphere HRTF is that in a wide frequency range it is a good approximation of an actual HRTF, whilst remaining an analytical model which may be exploited in the mathematics.

A full derivation of the rigid sphere formula is presented in Appendix E and in [95, 115]. The kernel due to an incident plane wave from $\hat{\mathbf{x}}$, scattered by a sphere of radius a and measured at position $\mathbf{r}$ is defined as

$$H(\mathbf{r}, \hat{\mathbf{x}}, \omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi j^n R_n(kr) Y_n^m(\hat{\mathbf{r}}) Y_n^m(\hat{\mathbf{x}})^* \tag{6.14}$$

where the radial filters are

$$R_n(kr) = \begin{cases} j_n(kr) - \frac{j_n'(ka)}{h_n^{(2)\prime}(kr)} h_n^{(2)}(kr) & \text{if } r > a \\ -\frac{j}{(ka)^2 h_n^{(2)\prime}(ka)} & \text{if } r = a \\ \text{Undefined} & \text{if } r < a \end{cases} \tag{6.15}$$

with $h_n^{(2)}(kr)$ the spherical Hankel function of the second kind and the superscript $'$ denoting differentiation. Utilising these definitions the GPWD becomes

$$\begin{aligned} p(\mathbf{r}, \omega) &= \int_{\Omega} q(\hat{\mathbf{x}}, \omega) \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi j^n R_n(kr) Y_n^m(\hat{\mathbf{r}}) Y_n^m(\hat{\mathbf{x}})^* d\Omega(\hat{\mathbf{x}}) \\ &= \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi j^n R_n(kr) Y_n^m(\hat{\mathbf{r}}) q_n^m(\omega). \end{aligned} \tag{6.16}$$

As before multiplying by a dummy variable $Y_{n'}^{m'}(\hat{\mathbf{r}})^*$, integrating over the $\mathcal{S}^2$ unit sphere and exploiting the orthogonality of the spherical harmonics leads to

$$p_n^m(\omega) = 4\pi j^n R_n(kr) q_n^m(\omega) \tag{6.17}$$

and evaluating at the ear positions ($\mathbf{r}_{l,r} = a\hat{\mathbf{r}}_{l,r}$)

$$\begin{aligned} p_n^m(\omega) &= -\frac{4\pi(j)^{n+1}}{(ka)^2 h_n^{(2)\prime}(ka)} q_n^m(\omega) \\ \implies h_n^m(\mathbf{r}_{l,r}, \omega) &= 4\pi j^n R_n(ka) Y_n^m(\hat{\mathbf{r}}_{l,r}). \end{aligned} \tag{6.18}$$

This shows that again there is a one-to-one relationship between the spherical harmonic coefficients of the soundfield and the B-format signals. However now the coefficients of that correspondence are more complicated as they include the scattering effects due to the rigid sphere. Furthermore, this mapping is explicit. As before the reproduced soundfield can still be synthesised by multiplication in the spherical domain for each order n and degree m. Note that the pressure anywhere in the surrounding soundfield is synthesised using the coefficients in Eqn. 6.17. The pressure at the ear positions on the rigid sphere surface is given by using the coefficients in Eqn. 6.18.

6.2 Higher Order Stereophony

6.2.1 Rigid Sphere HRTF Rendering

So far it has been shown that the HOA rendering approach can be viewed as multiplication in the spherical harmonic domain with the spherical harmonic coefficients of the soundfield and HRTF. Furthermore, under the special case of the rigid sphere, analytical expressions for the HRTF exist. Now, the HOS technique will be utilised for direct binaural reproduction taking advantage of the same soundfield representation.

Previously in Chapters 4 and 5, HOS has been derived by use of the Taylor expansion of a single plane wave soundfield resulting in order matching equations using functions of $\sin^n$ or $\cos^n$. This approach results in accurate soundfield reproduction across a line only, unlike with HOA which results in accurate reproduction across a sphere or circle (in 3D and 2D respectively). It was demonstrated that reproduction using a basis of spherical harmonics with $m = 0$ only to truncation order N was identical to using a basis of $\cos^n$ functions to the same order truncation, if the number of loudspeakers used was enough to ensure exact reproduction (a minimum of $(N+1)$ loudspeakers). This established a mapping between HOA to HOS, and also showed that reproduction using the set of $m = 0$ spherical harmonics reproduces the soundfield correctly across the z axis only. The fundamental assumption of HOS is that by aligning the system so that the reproduction axis is along the interaural axis, the correct binaural signals will be reproduced. However, this has yet to be shown true when a HRTF is included in the acoustics of the problem.

Now, the HOS approach will be shown to fully reproduce a soundfield represented using the PWD with the inclusion of a rigid sphere HRTF. This will form a highly efficient HOA to HOS to binaural decoder, and will also reinforce how HOS may be represented using a spherical harmonic representation.

Begin with the GPWD using the rigid sphere as the kernel. Any general position in the soundfield, $\mathbf{r}$, may be expressed as

$$p(\mathbf{r},\omega) = \int_{\Omega} q(\hat{\mathbf{x}},\omega) \sum_{n=0}^{\infty} \sum_{m=-n}^{n} 4\pi j^n R_n(kr) Y_n^m(\hat{\mathbf{r}}) Y_n^m(\hat{\mathbf{x}})^* d\Omega(\hat{\mathbf{x}}). \tag{6.19}$$

Next perform the HOS rotation such that the interaural axis aligns with the z axis. This corresponds to fixing the evaluation vector at $\mathbf{r} = \pm r\hat{\mathbf{z}}$ which is the axis the ears are assumed to lie along. Crucially, when evaluating the field along the z axis only spherical harmonics with $m = 0$ are non-zero. With this simplification

$$\begin{aligned} p(\mathbf{r} = r\hat{\mathbf{z}},\omega) &= \int_{\Omega} q(\hat{\mathbf{x}},\omega) \sum_{n=0}^{\infty} 4\pi j^n R_n(kr) Y_n^0(\hat{\mathbf{z}}) Y_n^0(\hat{\mathbf{x}})^* d\Omega(\hat{\mathbf{x}}) \\ &= \sum_{n=0}^{\infty} 4\pi j^n R_n(kr) Y_n^0(\hat{\mathbf{z}}) q_n^0(\omega). \end{aligned} \tag{6.20}$$

This key result is explained by the behaviour of the associated Legendre polynomials in the spherical harmonics definition. When evaluating along the z axis, the associated Legendre polynomials, $P_n^m(\cos\theta)$, are evaluated as

$$P_n^m(\cos\theta) = \begin{cases} 1 & \text{if } m = 0, \theta = 0 \\ -1 & \text{if } m = 0, \theta = \pi \\ 0 & \text{if } m \neq 0, \theta = 0, \pi \end{cases} \tag{6.21}$$

with θ the colatitude angle. Note the reduction to just the $m = 0$ spherical harmonics is equivalent to utilising the spherical harmonic addition theorem [112],

$$P_n(\hat{\mathbf{z}} \cdot \hat{\mathbf{x}}) = \frac{4\pi}{2n+1} \sum_{m=-n}^{n} Y_n^m(\hat{\mathbf{z}})^* Y_n^m(\hat{\mathbf{x}}) \tag{6.22}$$

leading to the equivalent representation of Eqn. 6.20 using just the Legendre polynomials

$$
\begin{aligned}
p(\mathbf{r} = \pm r\hat{\mathbf{z}}, \omega) &= \int_{\Omega} q(\hat{\mathbf{x}}, \omega) \sum_{n=0}^{\infty} j^n (2n+1) R_n(kr) P_n(\hat{\mathbf{z}} \cdot \hat{\mathbf{x}}) d\Omega(\hat{\mathbf{x}}) \\
&= \int_{\Omega} q(\hat{\mathbf{x}}, \omega) \sum_{n=0}^{\infty} j^n (2n+1) R_n(kr) P_n(\cos\theta) d\Omega(\hat{\mathbf{x}}).
\end{aligned} \tag{6.23}
$$

The set of $m = 0$ spherical harmonics and the Legendre polynomials may be switched between freely. The dot product is given by $\hat{\mathbf{z}} \cdot \hat{\mathbf{x}} = \cos\theta$, with θ the incident angle of the plane wave measured from the z axis (again the colatitude angle). This shows that when the evaluation points of the soundfield are along the z axis, representation and reproduction only requires the set of $(N+1)$ spherical harmonics with $m = 0$. We note that due to the special axisymmetric nature of the rigid sphere HRTF, seen in the definition of the rigid sphere HRTF kernel, after this rotation there is no dependence on the azimuthal position of the incoming source and therefore the binaural signals may be fully reproduced using the $m = 0$ spherical harmonics only. In turn this means the soundfield is correctly reproduced anywhere along the z axis with $r \geq a$ (as the field is undefined inside of the scattering rigid sphere), using only this small subset of spherical harmonic channels. Most importantly these positions includes the ear coordinates of the HRTF, and therefore the desired binaural signals are reproduced. This is the exact approach of the HOS technique and shows full binaural rendering for any soundfield represented using the PWD can be reproduced using HOS for the special case of the rigid sphere HRTF.

Note that it is common in the literature not to fix the evaluation point but instead the incident direction of the plane wave, $\hat{\mathbf{k}} = \hat{\mathbf{z}}$, to take advantage of symmetry in a different manner. In this special case the mathematics reveals that using the $m = 0$ spherical harmonic coefficients of the rigid sphere HRTF fully reproduce this soundfield with plane waves directed along the z axis only, but now for any generalised orientation of the interaural axis. This result is due to this special incident soundfield being axisymmetric about the z axis. Here however, by fixing the evaluation point to the z axis, $\hat{\mathbf{r}} = \hat{\mathbf{z}}$, and allowing any generalised plane wave direction of incidence means the soundfield will be correct along the z axis only when representing using just the $m = 0$ spherical harmonics, however for any other evaluation points the soundfield will not be correct. This may be intuitively understood as a cone of confusion. With this specific HRTF and orientation a plane wave incident from a given θ lies on a cone of confusion for all possible azimuthal angles ϕ. All positions on this cone of confusion lead to identical binaural signals, due to the axisymmetric nature of the rigid sphere. This leads to interesting rendering possibilities, as sources with elevation (relative to the HRTFs rotated frame of reference) can be rendered using a horizontal only source position.

Two key concepts have been demonstrated. First, it has been shown that the HOS technique of reproducing the subset of channels with $m = 0$ only is valid both for

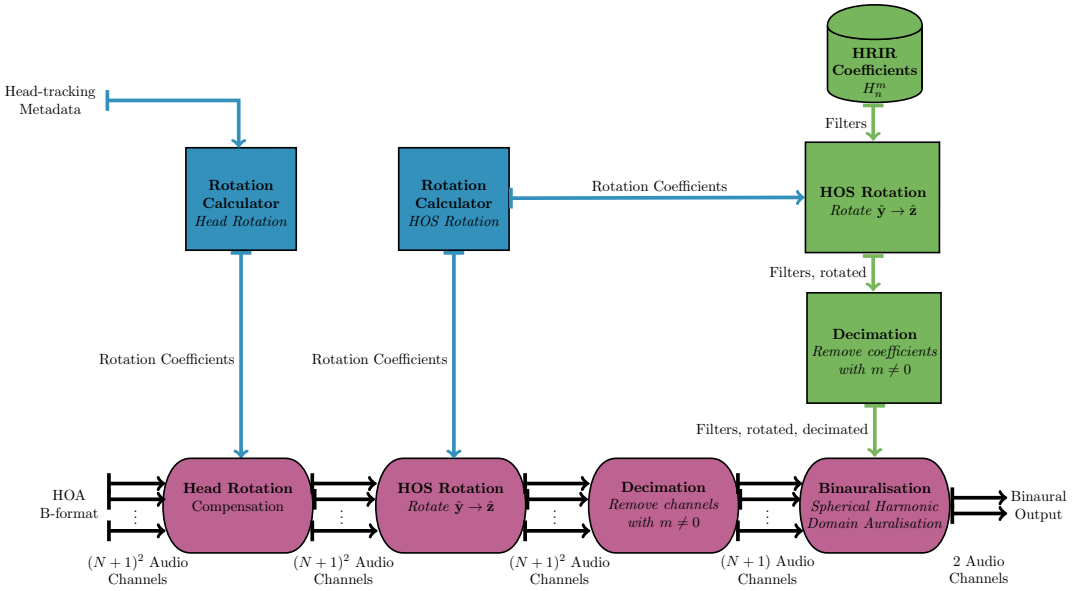


FIGURE 6.1: Signal flow to render HOA B-format to binaural using HOS in the spherical harmonic domain. The binauralisation step may also be performed using virtual loudspeaker rendering.

the case of free field and when including a rigid sphere HRTF (if the specific HOS rotation has been performed). As with the previous definition of HOS, this means for N-th order reproduction only $(N+1)$ channels are required as opposed to $(N+1)^2$ with HOA - a substantial reduction particularly as the truncation order increases. The second key point is that the HOS technique, whereas previously derived utilising a Taylor expansion of a plane wave soundfield, may also be derived using a spherical harmonic decomposition of a soundfield. This important representation is popular due to its use in HOA and therefore compatibility with existing HOA techniques and content is ensured.

6.2.2 Ambisonic to Stereo Binaural Decoder

The above forms the basis for a highly efficient 3D HOA to HOS to binaural decoder, of which the signal flow is demonstrated in Fig. 6.1. First, to account for listener head rotations, the soundfield must be rotated in the opposite direction to the listener's rotation based on the output of a head-tracker (as is the case for any adaptive binaural rendering of HOA material). Next, to align the interaural axis with the z axis the soundfield must be rotated again (notably in real-time these two rotations may be combined into one operation to improve computational efficiency). A decimation operation is next performed, where the number of audio channels is reduced from $(N+1)^2$ to $(N+1)$, keeping only the B-format channels with $m = 0$. Finally, the soundfield is binauralised, which in this case is done using direct convolution/multiplication of the remaining B-format channels with the relevant spherical harmonic HRIR/HRTF coefficients, which have also undergone a similar rotation and decimation operation. The HRIR coefficients may be processed once off-line,

as these coefficients do not change during rendering. However, the operations on the soundfield representation must be performed continuously. Alternatively, the soundfield may be binauralised using a virtual loudspeaker array approach, using HOS mode matching to define the loudspeaker gains and sampled HRIRs as opposed to HRIR spherical harmonic coefficients.

6.2.3 Generalised HRTF Rendering

So far, it has been demonstrated that HOS is a valid reproduction technique for free field conditions and when including a rigid sphere HRTF. Furthermore, an efficient binaural renderer using HOS to convert 3D HOA to binaural for a rigid sphere HRTF has been presented. However, a valid question is whether HOS reproduces with sufficient accuracy a target soundfield to get the correct binaural signals when using a more realistic HRTF. The rigid sphere HRTF exhibits the phenomena of only requiring the $m = 0$ channels due to its completely axisymmetric nature. Therefore for a more lifelike HRTF, channels with $m \neq 0$ will likely be important for reproduction. However, it is known that the rigid sphere is a very good model of a HRTF within a frequency limit (up to approximately 4000 Hz), thus it may be expected that the HOS approach still works with a lifelike HRTF to an error bound, within this frequency limit.

To assess this concept, the energy distribution across the spherical harmonic coefficients of a HRTF before and after the specified HOS rotation gives an indication of how the HOS approach will perform. The energy of a spherical harmonic coefficient indicates the importance of its contribution to rendering the HRTF. Before the rotation, the energy is expected to be spread across all different spherical harmonic coefficients for both the rigid sphere and the general HRTF. Following the rotation, the energy for the rigid sphere HRTF is predicted to be only in the $m = 0$ coefficient channels. Interestingly, this specific rotated reference system for a HRTF has utilised in [158] in the context of optimal phase unwrapping for HRTFs over a sphere. Here it was observed that this head orientation (chosen to improve the phase unwrapping operation) also led to a reordering of the energy into coefficients close to $m = 0$.

The question here is to what extent is this energy focused into the $m = 0$ coefficients when considering the non-symmetric, general HRTF? If this focusing is strong enough, in practise HOS could work sufficiently well reproducing the desired binaural signals to within an error bound using just $m = 0$ channels. In the region the rigid sphere approximates a general HRTF (up to approximately $4,000$ Hz), the HOS approach is predicted to be acceptable.

To consider the energy in each spherical harmonic coefficient the rigid sphere HRTF and a Neumann KU100 HRTF were considered. For the KU100 the far-field 2702 point spherical Lebedev grid measurements of the Neumann KU100 by TH Köln was used [159]. For both HRTFs, the truncation order was set to $N = 40$ corresponding

to a spatial aliasing frequency as per the standard $N = kr$ rule of $\approx$ 25000 Hz [48]. This truncation order was chosen as this spatial aliasing frequency is high enough to not create issues in the recreation of full-range audio signals, however resulted in a considerable number of $(40+1)^2 = 1681$ spherical harmonic channels under consideration.

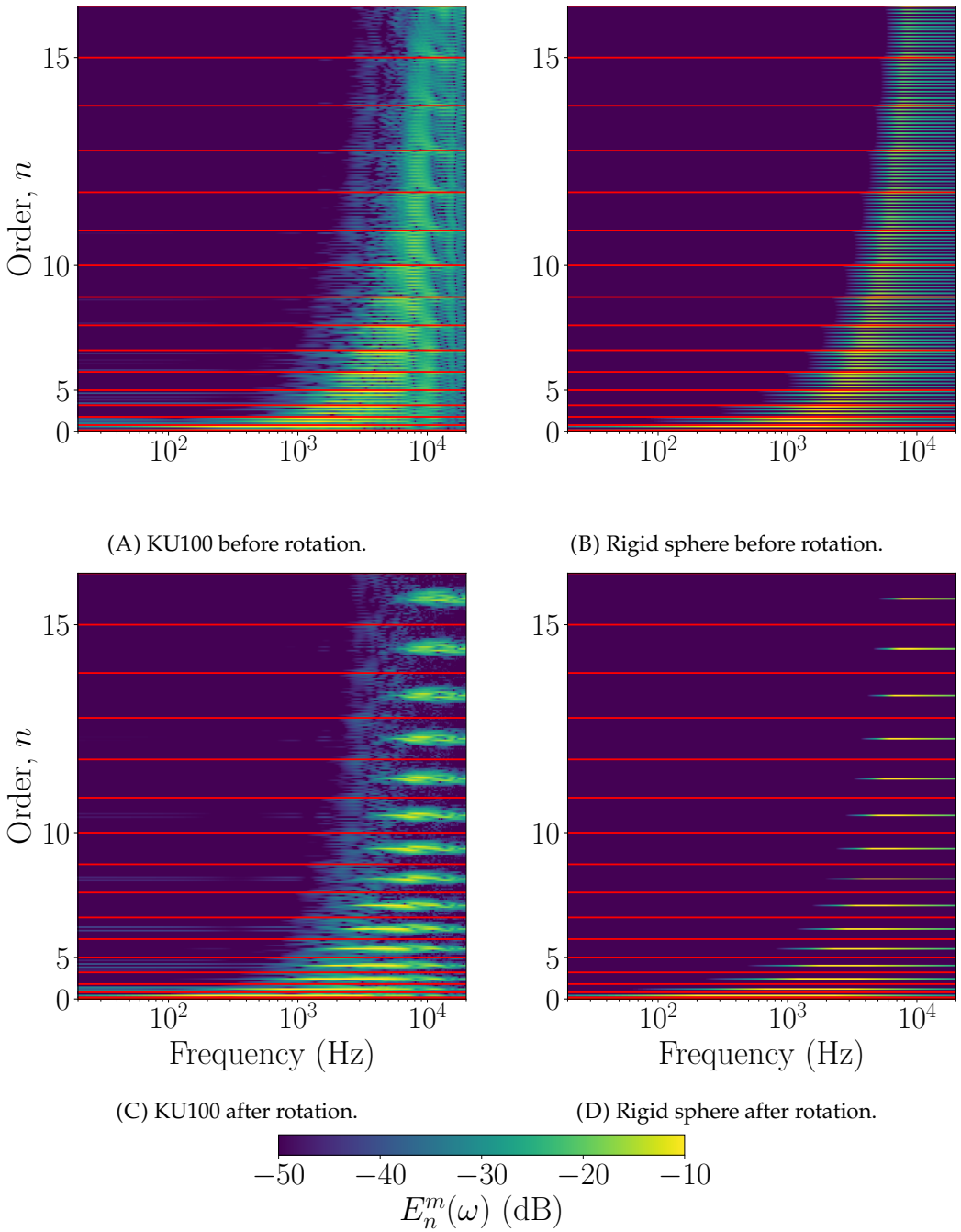
The spherical harmonic transform of the KU100 HRTF was performed to find the coefficients of its spherical harmonic decomposition using a pseudoinverse approach [150]. The rigid sphere spherical harmonic coefficients were calculated using its analytical expression with a head radius of $a = 0.10$ m; this head radius was found to give a similar fit to the characteristics of a Neumann KU100. Next, the normalised modal energy of each spherical harmonic coefficient as a function of frequency, $E_n^m(\mathbf{r}_{l,r}, \omega)$, was calculated by

$$E_n^m(\mathbf{r}_{l,r}, \omega) = \frac{|h_n^m(\mathbf{r}_{l,r}, \omega)|^2}{\sum_{n'=0}^{40} \sum_{m'=-n'}^{n'} |h_{n'}^{m'}(\mathbf{r}_{l,r}, \omega)|^2} \tag{6.24}$$

where the denominator is a normalisation corresponding to the energy of the entire HRTF.

The energetic distribution of coefficients up to $N = 15$ are shown in Fig. 6.2 for the left ear of each HRTF. Here the normalised energy across all frequencies is shown. The spherical harmonic coefficients are ordered using Ambisonic Channel Numbering (ACN) [1], $acn = n^2 + n + m$, such that for each given order $m = -n$ is shown first, and $m = n$ shown last with $m = 0$ in the middle. As $m \in [-n, n]$ there are $(2n+1)$ spherical harmonic coefficients per each order. Before the rotation it is clear for both HRTFs the energy is spread across all degrees of each order. Furthermore, as per the $N = kr$ rule at low frequencies only lower orders are required and contain any significant energy, whilst as the frequency increases higher orders become significant.

However, following the HOS rotation so that the interaural axes for both HRTFs align along the z axis, the energy is shifted amongst the spherical harmonic coefficients considerably. For the rigid sphere, it is clear that now only the $m = 0$ spherical harmonics contain any energy, where on the figure the $m = 0$ spherical harmonics are positioned exactly halfway between each red order lines. This demonstrates numerically that only the $m = 0$ are required when using a rigid sphere HRTF, and therefore the HOS approach will fully reproduce the HRTF. Interestingly, with the KU100 there is also a strong reordering of the energy in the HRTF to focus around the $m = 0$ channels. Whilst coefficients outside of $m = 0$ still contain significant energy, it appears that for each given order the extreme degree channels where $m \to -n, n$ are less important. Furthermore, this effect is observed across all orders. Thus, by applying this rotation the energy of the KU100 HRTF has been reordered to focus



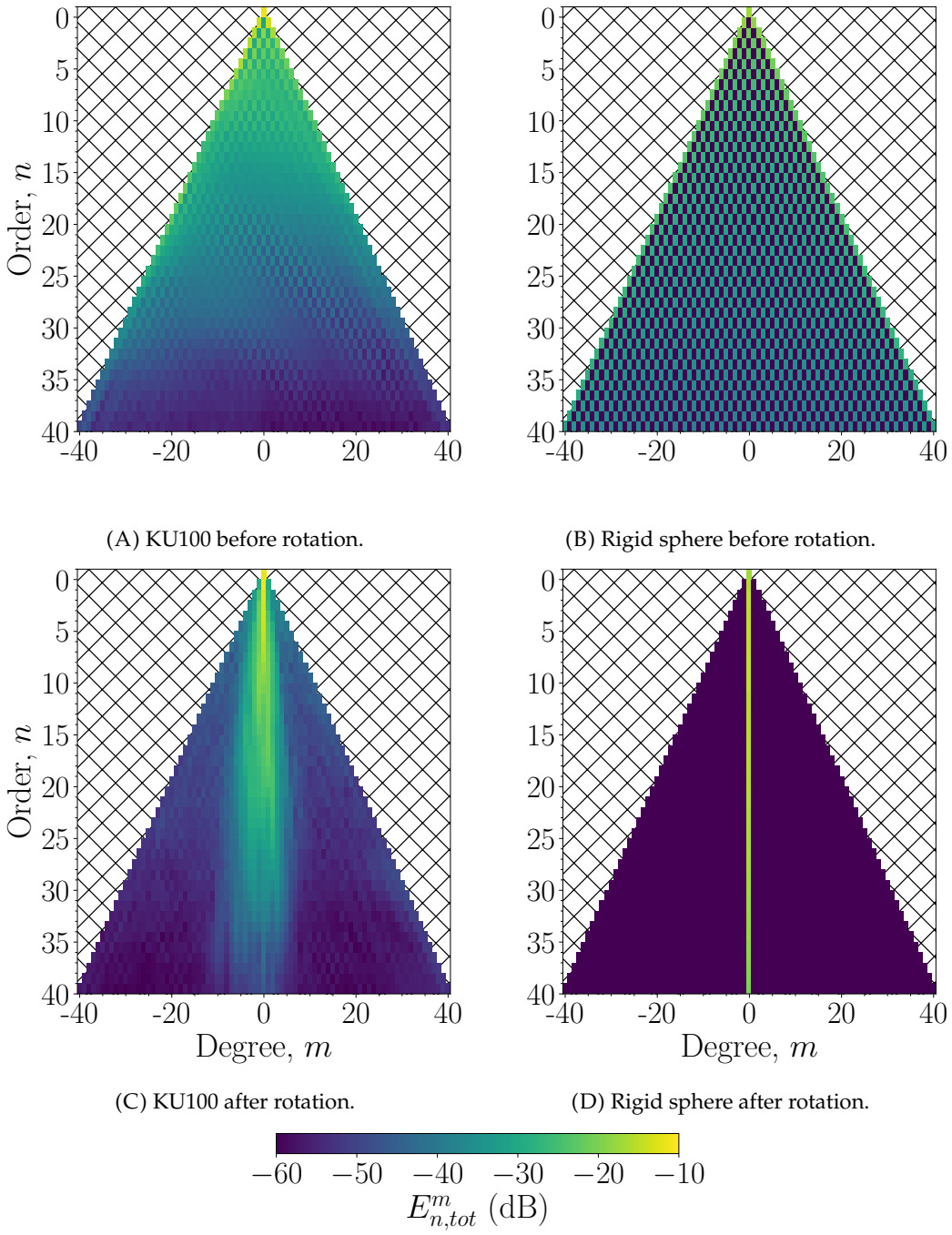
(A) KU100 before rotation.

(B) Rigid sphere before rotation.

(C) KU100 after rotation.

(D) Rigid sphere after rotation.

FIGURE 6.2: Left ear HRTF spherical harmonic coefficients across frequency up to order 15, before and after the energy reordering rotation. The coefficients are ordered in increasing order n and degree $m \in [-n, n]$ and the red lines indicate the start of each new order.



(A) KU100 before rotation.

(B) Rigid sphere before rotation.

(C) KU100 after rotation.

(D) Rigid sphere after rotation.

FIGURE 6.3: Left ear HRTF spherical harmonic coefficients across all frequencies up to order 40, before and after the energy reordering rotation. The coefficients are ordered in sets of order n with degree $m \in [-n, n]$.

around spherical harmonic coefficients with m close to 0. In this sense, a more efficient representation of the HRTF has been developed, such that less spherical harmonic coefficients may be required to represent the HRTF.

Next define the normalised total energy across all frequencies for each given spherical harmonic coefficient, $E^m_{n,tot}(\mathbf{r}_{l,r}, \omega)$, as

$$E^m_{n,tot}(\mathbf{r}_{l,r}, \omega) = \frac{\int |h^m_n(\mathbf{r}_{l,r}, \omega)|^2 d\omega}{\int \sum_{n'=0}^{40} \sum_{m'=-n'}^{n'} |h^{m'}_{n'}(\mathbf{r}_{l,r}, \omega)|^2 d\omega}. \tag{6.25}$$

Fig. 6.3 shows the normalised total energy for the left ear of the rigid sphere and KU100 before and after the rotation. It is clear that after the rotation the energy for both representations is focused about the $m = 0$ channels, with the rigid sphere fully defined using just this subset of coefficients. For the KU100 it is clearly seen that the outer orders as $m \to -n, n$ are in some instances 30 to 40 dB down from the main peaks centred around $m = 0$. Finally, it may be observed that in general lower orders contain more energy than the higher orders. For the KU100 HRTF the focusing is strongest at low orders, relating to the fact that the rigid sphere is a good HRTF approximation at low to mid frequencies. However, with the KU100 HRTF as the order increases the energy is spread considerably more, for example past $n = 7$. This corresponds to approximately $4,250$ Hz as per $N = kr$, further enforcing the accepted validity region of the rigid sphere up to $4,000$ Hz. However, even here there is still a significant reordering of the energy to channels with m close to 0.

6.3 Simulations

To compare the effect of rendering using HOS and thus the $m = 0$ spherical harmonic coefficients only, a number of reproduced HRTFs are shown in Fig. 6.4. Here, the HRTF magnitude for a varying horizontal incident source with zero elevation is shown for the reference KU100 and those synthesised using various order HOS and HOA approaches. Results for a range of frequencies from 100 to 4000 Hz were simulated. The spherical harmonic coefficients used for the re-synthesis are those from the KU100 $N = 40$ estimation after the HOS rotation, and may be considered to not contain any spatial aliasing due to the dense sampling of 2702 points. HOS re-synthesis was performed by using the $m = 0$ coefficients up to the truncation order N resulting in the use of only $(N + 1)$ channels. HOA re-synthesis used all coefficients up to order N corresponding to $(N + 1)^2$ channels. This characterises the main advantage of HOS, that by performing the energy reordering a smaller number of spherical harmonic channels may be used in the rendering. HOS with a truncation order of $N = 1, 4, 7$ was compared to HOA with $N = 1, 4$ to consider matching truncation orders between the two techniques, as well as an order 7 HOS approach that

uses a third of the number of channels required for the order 4 HOA technique but still achieves a higher truncation order.

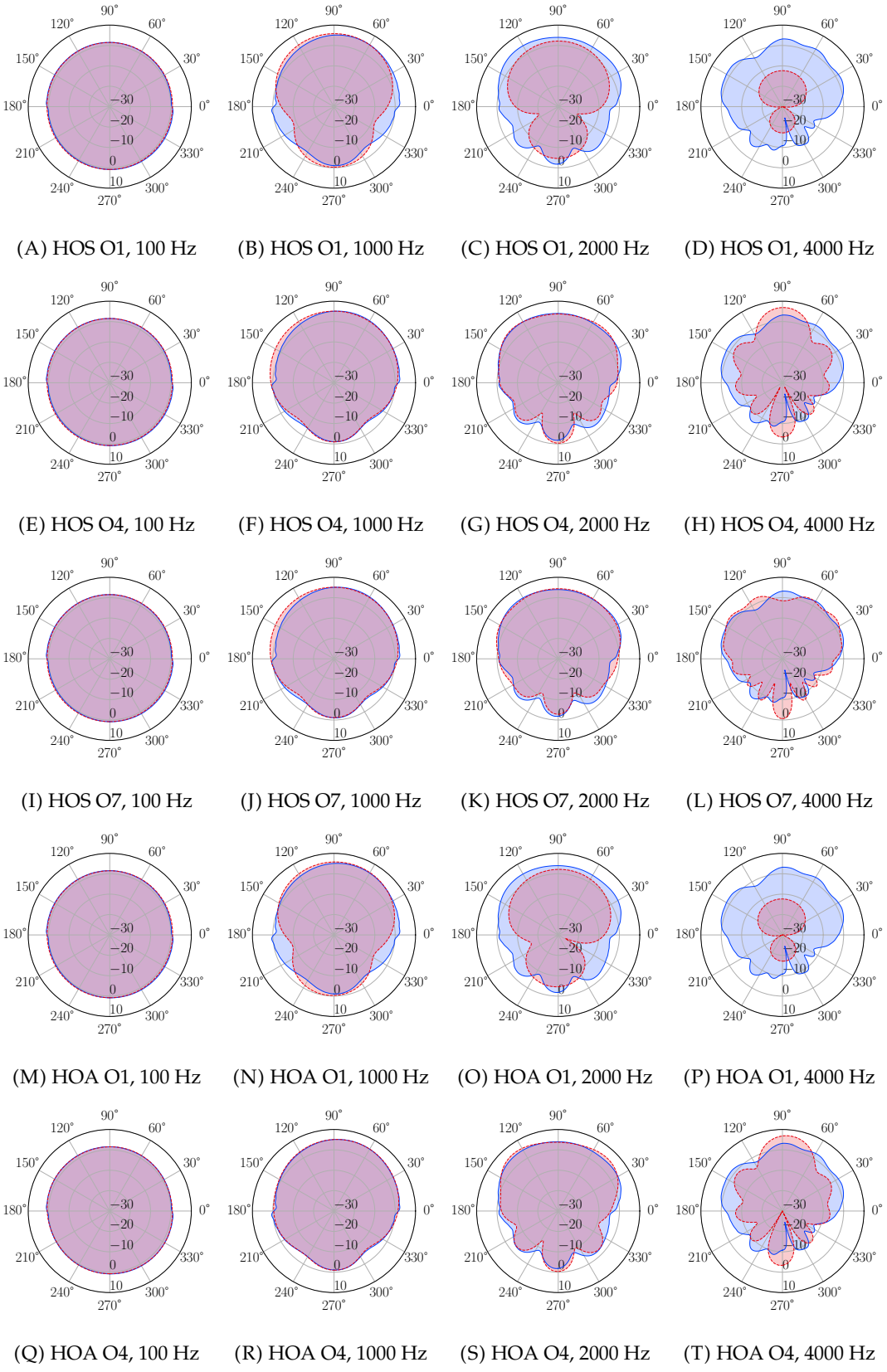
The polar plots demonstrate how by using $m = 0$ channels only, the HOS technique is only capable of reproducing a HRTF that is symmetric about the interaural axis, defined between 90° to 270°. This is expected as the $m = 0$ spherical harmonics fully define functions that are axisymmetric about the z axis, which has here been rotated to coincide with the interaural axis. Thus the HOS approaches most closely match the target reference when the reference itself approximates a symmetric function, which in turn corresponds to lower frequencies. However, even close to 4000 Hz the HOS approach appears to match the reference well most clearly seen through the order 7 technique. This will become more of an issue at high frequencies, for example HOS will be unable to reproduce high frequency spectral cues due to the pinna that aid in resolving front-back confusions.

Increasing the order of the HOS technique increases the frequency limit at which accurate reconstruction occurs, in a similar manner to HOA. The advantage of HOS requiring less channels for higher order reproduction may be seen when comparing order 7 HOS with order 4 HOA at 4000 Hz. At this frequency using HOA with 25 channels appears less optimal than using HOS with just 8 channels but a higher truncation order. This indicates that the energy reordering effect due to the rotation is significant, and results in sufficient information being refocused into the smaller subset of channels.

Fig. 6.5 shows re-synthesised HRTFs again across the horizontal plane except now with 45° elevation. Despite the addition of considering elevated sound sources, similar trends are observed regarding the HOS approach such as axisymmetric responses and increasing accuracy with a higher truncation order.

6.4 Comparative Listening Test

To investigate the viability of the new HOS binaural rendering approach, a listening test was designed and run to compare it to HOA. The MUSHRA approach was utilised, which allows the comparison of multiple options within each page of the test compared to a reference source, as well as a hidden reference and hidden anchor to grade the listener's ability to discern changes in the audio [146]. The aim of the test was to subjectively compare head-tracked binaural HOS to head-tracked binaural HOA using varying orders for both, to see if any difference could be perceived between each approach for both different orders and different audio channel limits.



(A) HOS O1, 100 Hz (B) HOS O1, 1000 Hz (C) HOS O1, 2000 Hz (D) HOS O1, 4000 Hz

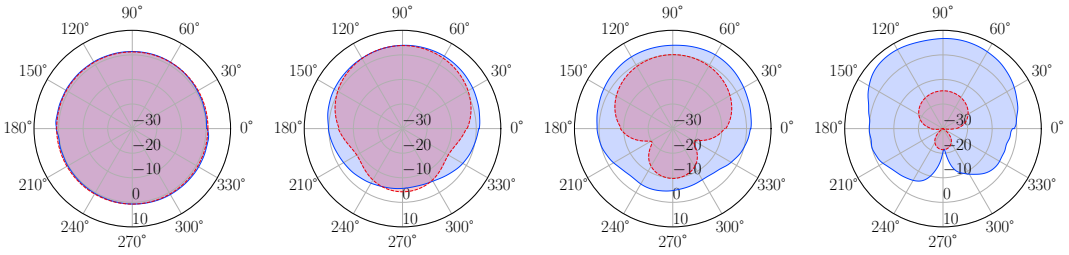
(E) HOS O4, 100 Hz (F) HOS O4, 1000 Hz (G) HOS O4, 2000 Hz (H) HOS O4, 4000 Hz

(I) HOS O7, 100 Hz (J) HOS O7, 1000 Hz (K) HOS O7, 2000 Hz (L) HOS O7, 4000 Hz

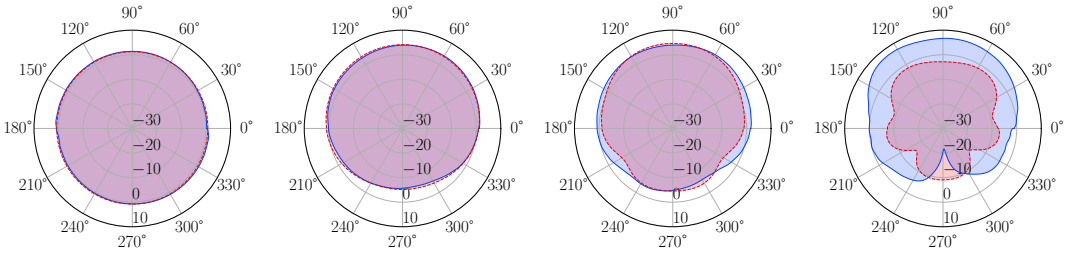
(M) HOA O1, 100 Hz (N) HOA O1, 1000 Hz (O) HOA O1, 2000 Hz (P) HOA O1, 4000 Hz

(Q) HOA O4, 100 Hz (R) HOA O4, 1000 Hz (S) HOA O4, 2000 Hz (T) HOA O4, 4000 Hz

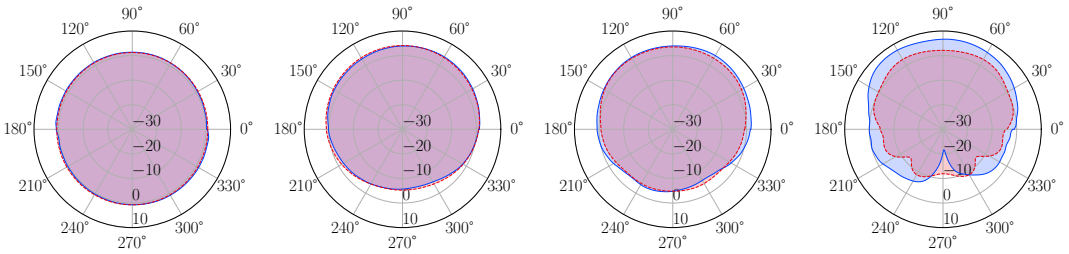
FIGURE 6.4: Reference (blue) and reproduced (red) left ear HRTF magnitudes for the KU100 using various order systems of HOS and HOA for sources around the horizontal plane, with no elevation.



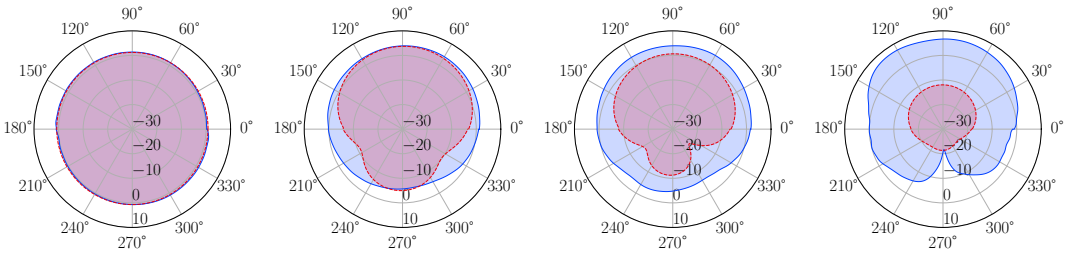
(A) HOS O1, 100 Hz (B) HOS O1, 1000 Hz (C) HOS O1, 2000 Hz (D) HOS O1, 4000 Hz



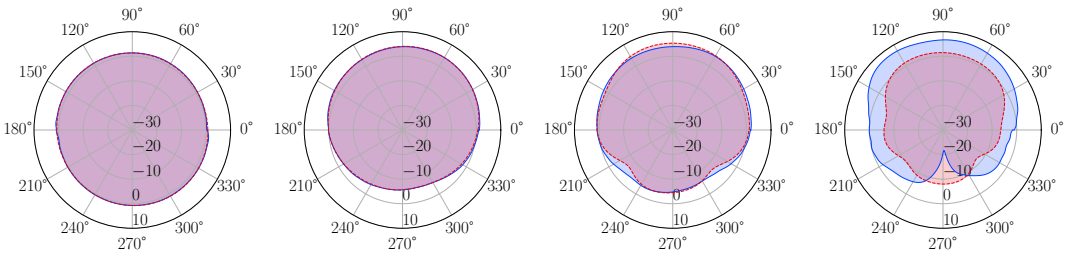
(E) HOS O4, 100 Hz (F) HOS O4, 1000 Hz (G) HOS O4, 2000 Hz (H) HOS O4, 4000 Hz



(I) HOS O7, 100 Hz (J) HOS O7, 1000 Hz (K) HOS O7, 2000 Hz (L) HOS O7, 4000 Hz

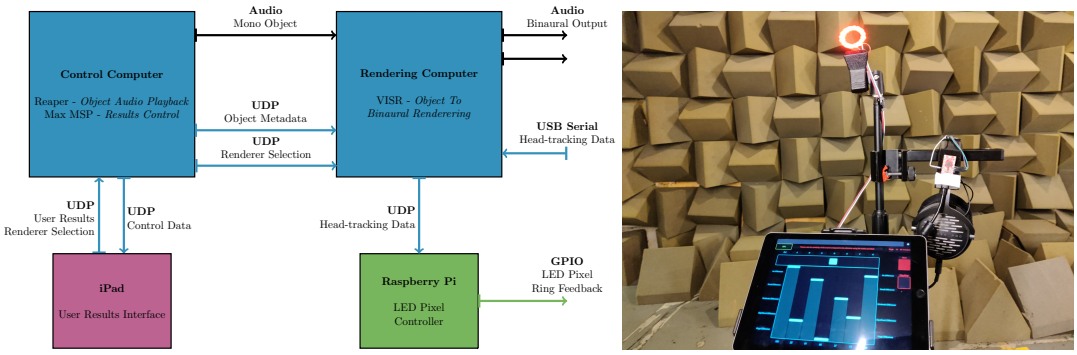


(M) HOA O1, 100 Hz (N) HOA O1, 1000 Hz (O) HOA O1, 2000 Hz (P) HOA O1, 4000 Hz



(Q) HOA O4, 100 Hz (R) HOA O4, 1000 Hz (S) HOA O4, 2000 Hz (T) HOA O4, 4000 Hz

FIGURE 6.5: Reference (blue) and reproduced (red) left ear HRTF magnitudes for the KU100 using various order systems of HOS and HOA for sources around the horizontal plane, with 45 degree elevation.



(A) Diagram of the listening test apparatus and signal flows for audio and metadata.

(B) Picture of the experimental setup showing the iPad user interface, the headphones and head-tracker and the LED ring for listener feedback.

FIGURE 6.6: Experimental setup for the listening test.

6.4.1 Experimental Setup

The listening test was performed in the small anechoic chamber at the ISVR to ensure isolated conditions from external noise. The experimental setup delivered real-time head-tracked binaural renders of a single static source about the listener and is shown in Fig. 6.6. A control computer hosted the mono audio objects in a Reaper DAW session, where the Versatile Interactive Scene Renderer (VISR) object-based audio production suite plugins where used to encode the appropriate object metadata [147]. This object audio stream and associated metadata were sent via MADI and UDP respectively to a separate rendering computer, running a python script also utilising the VISR framework for audio processing. A head-tracker consisting of a micro Arduino and a BNO055 orientation sensor was connected to the rendering computer to provide listener head orientation data. The python script ran a number of spatial audio rendering approaches synchronously, converting the mono audio object input to a binaural output for each approach simultaneously. All rendering approaches were contained within a master switching controller, which allowed the listener to switch between each approach. Triggering the switch resulted in a 5 ms fade out and 5 ms fade in using a cosine ramp of the binaural audio, as required by the MUSHRA standard [146]. Headphone equalisation was applied to the final binaural outputs as specified in [160, 161], based on measurements using the headphones and a Neumann KU100 microphone to ensure consistency with the HRTFs used in the binaural rendering. The binaural output was delivered to the listener by an RME MADIFace pro audio interface and a pair of reference Beyerdynamic DT1990 Pro open back headphones.

The user selected and rated each of the approaches via use of an iPad running TouchOSC, as seen in Fig. 6.6(B), where they could also play/stop the audio as necessary. This communicated using UDP with the control computer, running a Max MSP patch that saved the listener responses. This Max MSP patch also triggered the Reaper session to play and start the audio, as well as the renderer computer to

select the desired audio rendering approach both via UDP. Finally, the rendering computer also sent the head-tracking data via UDP to a Raspberry Pi that controlled a ring of LEDs positioned in front of the listener. As head movements were very important for certain techniques to work, if the user remained stationary for a short period of time the LEDs flashed red to remind the listener to keep rotating their head.

The audio processing at all stages was run at a 48 kHz sampling frequency and with block sizes of 512 samples. The volume was kept to a consistent and normal listening level for all participants.

6.4.2 Experimental Design

The MUSHRA test design was chosen as it allows for the comparison of multiple different stimuli in each page of the listening test. The presence of a hidden reference and an anchor stimuli helps rate the quality of the participants results, including the option to discard unreliable listeners when processing the data. The reference was chosen to be a dynamic HRIR rendering using the densely measured 2702 point grid of anechoic Neumann KU100 HRIR filters. Time domain cross-fading over a single processing block size was employed to ensure a smooth transition between HRIR filters. No reverberation was used in any of the binaural renderers, as a pilot test showed this could mask small differences between the techniques and therefore inhibit the listener's ability to rank the approaches. This came at the cost of using anechoic binaural renderers, which can cause issues when trying to externalise the sources [89]. However, head-tracking has also been demonstrated to aid in externalisation [162]. The low anchor was chosen as a dynamic HRIR renderer, however with the anchor source position fixed to $(\theta, \phi) = (90°, 0°)$. This position was substantially different to all actual source positions used in the test. The anchor was also low-passed with a filter as specified in [146] and with a cut-off frequency of 3.5 kHz. Therefore the anchor was a tonally and spatially impeded binaural renderer.

The spatial audio techniques chosen to compare are detailed in Table 6.1. These techniques were chosen to consider the following questions -

- Is there any perceived difference between HOS and HOA renderers of the same order, but different number of audio rendering channels?
- Given a limit on the number of available audio channels, is there any perceived improvement in using a higher order HOS renderer than HOA?

Thus to answer these questions, HOS and HOA of matching orders ($N = 1, 4$) were used. To consider the question of using a higher order with the same limit on the number of audio channels, HOS $N = 4, L = 5$ and HOS $N = 7, L = 8$ were included to be compared to HOA $N = 4, L = 25$ with L the number of audio channels. The choice of renderers thus matches those used in the numerical simulations.

Renderer Approach	Truncation Order, N	Number Of Audio Channels, L
HOS	1	2
HOS	4	5
HOS	7	8
HOA	1	4
HOA	4	25

TABLE 6.1: Spatial audio techniques chosen for comparison in the listening test.

Both HOA and HOS renderers used a virtual loudspeaker approach with the minimal number of loudspeakers required (value L in Table 6.1). The virtual loudspeaker HRIR filters were chosen from the same Neumann KU100 dataset as used for the reference renderer to avoid bias that could occur if different HRTFs were used across the experiment. The virtual loudspeaker arrays were not adapted with head rotations so they remained in fixed positions relative to the listener regardless of their head orientation. Instead, the virtual source position was adapted to compensate for head rotations. In all renderers a very small amount of Tikhonov regularisation with $\beta = 0.0001$ was used when inverting the plant matrix to define the loudspeaker signals, as per Eqn. 4.16. The HOS technique requires a very simple virtual loudspeaker array, with loudspeakers positioned in just the horizontal plane in front of the listener (all HOS loudspeaker positions $\theta = 90^\circ, \phi \in [-90^\circ, 90^\circ]$) . The HOA approaches used loudspeaker positions over a full sphere as per an equal area sampling regime. A gain normalisation was applied to each renderer to ensure the volume was consistent across all approaches.

Two different source stimuli were tested, in three different positions each. The whole test was also repeated once, which meant there were $2 \times 3 \times 2 = 12$ pages in the whole test. The order of each page and the renderers was randomised for each participant and on average the test took one hour. The stimuli used were looped anechoic recordings of a male speech sample and a short drum beat. Critical signals have been shown to be important in listening tests, to reveal differences between renderers and in particular drums have been shown to be an appropriate critical signal combining broadband sources and sharp transients [149]. The three source positions were

$$(\theta_1,\ \phi_1) = (90^\circ,\ 225^\circ), \qquad (\theta_2,\ \phi_2) = (60^\circ,\ 345^\circ), \qquad (\theta_3,\ \phi_3) = (30^\circ,\ 45^\circ).$$

The participants were asked to rate the 'similarity' of each renderer to the reference. This was scored on a scale from 0-100, with the labels as shown in Fig. 6.7. The participants were encouraged to consider the following properties of the sound, along with any other properties they deemed appropriate - tonal changes, absolute position deviations, changes in apparent source width, stability of position with head

FIGURE 6.7: User interface for controlling and rating the different renderer options.

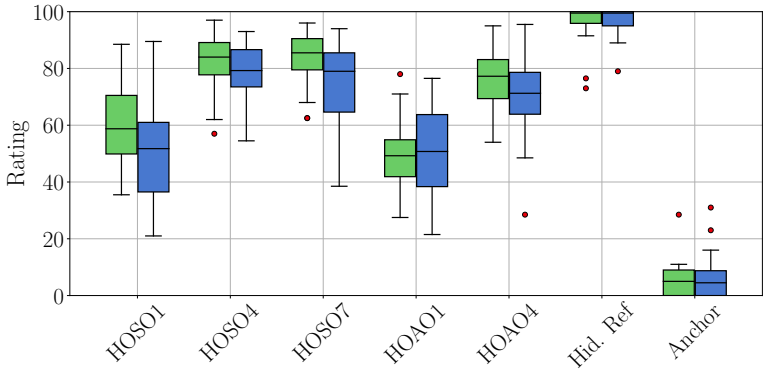
movements. The participants were strongly encouraged to keep rotating their head, with the LED feedback to remind them if they stayed stationary for too long.

Before the test began, a training phase was run for the participants to learn how to use the interface and understand how to translate perceived differences onto the scale provided. The training phase included a representative sample of the breadth of differences in the test, and included the reference, anchor, HOS O7, HOA O1 and HOA O4 renderers for both the drum and speech sample at the $(\theta_1,\ \phi_1) = (90^\circ,\ 225^\circ)$ position only.

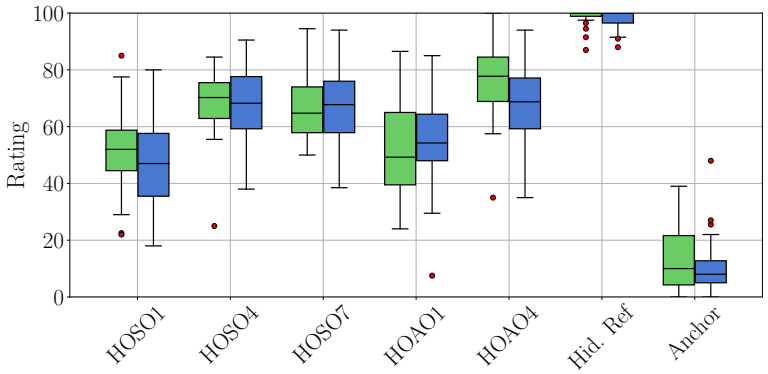
6.4.3 Results

Overall, 29 participants took part in the test, with 28 participants between the ages of 22-50 and one participant over 50. 25 of the subjects were male and 4 female, with all subjects self-reporting having normal hearing. One participant's results were discarded as they did not correctly identify the hidden reference enough times, as the MUSHRA standard dictate. As the experiment contained a full repeat of the test, each participant's final score was averaged over the repeat results. The average deviation between the repeat test scores across all listeners and renderers was calculated as a rating deviation of 12.26 (full score scale from 0-100). This indicates a fair level of consistency in the listener's responses.

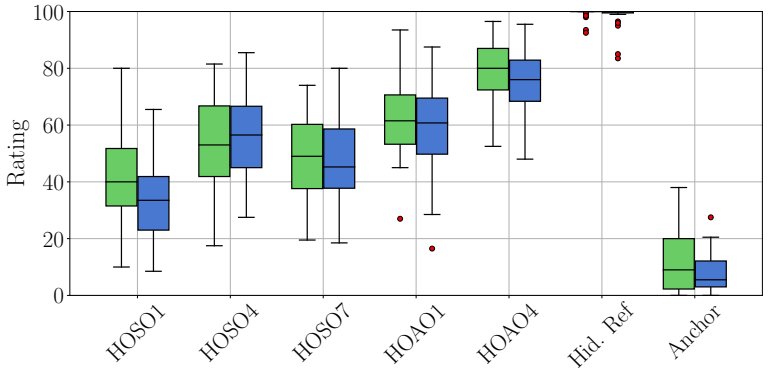
Fig. 6.8 shows the raw data as boxplots indicating the ratings for each renderer type at the three evaluated source positions. It is clear that the hidden reference and anchor were both scored consistently throughout the test. Generally, it might be noted that the HOS O1 technique was scored lower than HOS O4 and HOS O7, which were both roughly equivalent for all source positions. Increasing the HOS order from 4 to 7 did not lead to obvious improvements. Furthermore, HOA O1 was always rated



(A) Results for a source at $(\theta_1, \phi_1) = (90°, 225°)$.



(B) Results for a source at $(\theta_2, \phi_2) = (60°, 345°)$.



(C) Results for a source at $(\theta_3, \phi_3) = (30°, 45°)$.

FIGURE 6.8: Boxplots showing the raw results for each source position, renderer type and signal type. The box indicates the lower and upper quartiles, with the median the black line intersecting each box. The whiskers indicate the minimum/maximum rating for identifying outliers, which are consequently shown as red circles. Each graph shows a single source position. The green boxes illustrate the speech signal results while the blue boxes the drum signal results.

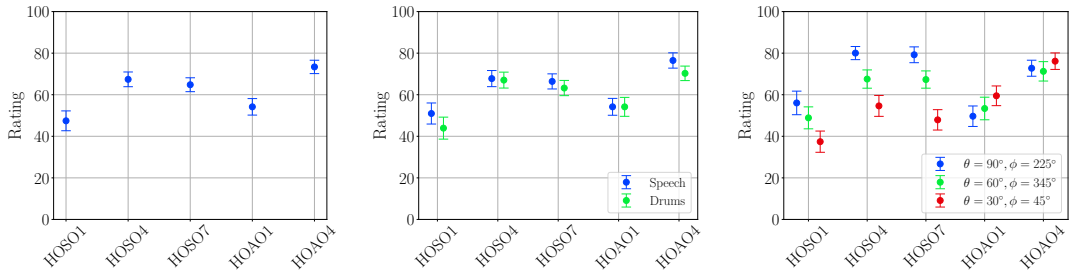
Effect	df	df Error	F	p
Technique *	3.313	89.445	78.884	< 0.001
Source *	1	27	13.38	0.001
Position	2	54	110.257	< 0.001
Technique $\times$ Source	4	108	4.522	0.002
Technique $\times$ Position	8	216	30.679	< 0.001
Source $\times$ Position *	1.732	46.76	1.071	0.343
Technique $\times$ Source $\times$ Position *	8	216	1.333	0.228

TABLE 6.2: Within-subject effects from the three-way repeated measures ANOVA. An asterisk indicates a condition with the Huynh-Feldt correction applied.

significantly lower than HOA O4, as expected. The performance of some HOS techniques appears dependent on the source position. For a source in the horizontal plane the HOS approaches perform strongly, but have lower ratings as elevation is increased. In contrast the HOA approaches perform more consistently with respect to source position. However, the HOA O1 approach surprisingly performed better for the source position with the largest elevation.

Statistical analysis of the results was performed using a repeated measures ANOVA. A Kolmogorov-Smirnov normality test was applied to the residuals of the dataset to test for normality. It rejected the null hypothesis at a significance level of $p = 0.05$ for 28 out of the 30 tests, indicating a high degree of normality. Tests such as ANOVA are rather insensitive to small violations of normality [146]. Within-subjects effects were considered for all combinations of the three independent variables (technique, source type and source position). Three test combinations violated a test for sphericity, therefore for these conditions the Huynh-Feldt corrected values are reported as recommended by the MUSHRA standard [146].

Table 6.2 shows the results of within-subject effects from a three-way repeated measures ANOVA, considering the variation due to the type of renderer (technique), the type of stimuli (source) and the location of the source (position) at a 5% significance level. Only two out of the seven conditions were found to not be significant. These were the source $\times$ position and technique $\times$ source $\times$ position interactions. The main effects due to technique $[F(3.313, 89.445) = 78.884, p < 0.001]$, source $[F(1, 27) = 13.380, p = 0.001]$ and position $[F(2, 54) = 110.257, p < 0.001]$ demonstrate that changing each of these variables has a significant effect on the perceived similarity to the reference. The significant first order interactions due to technique $\times$ source $[F(4, 108) = 4.522, p = 0.002]$ and technique $\times$ position $[F(8, 216) = 30.679, p < 0.001]$ both suggest that not all techniques perform the same depending on the characteristics of the source, that is the type of signal and the position. This is clearly seen as well in Fig. 6.8 where generally the HOS approaches perform worst for an elevated position, whilst the HOA O1 technique performs better as elevation increases yet the HOA O4 renderer appears more consistent.



(A) Overall mean ratings.

(B) Mean ratings for each source stimuli, speech and drums.

(C) Mean ratings for each of the three source positions.

FIGURE 6.9: Mean ratings for each rendering technique. The caps indicate 95% confidence intervals.

Fig. 6.9 shows the calculated mean ratings for each technique. The overall ratings in Fig. 6.9(A) also follow the trends in Fig. 6.8 and the statistical results. In general, HOS O1 performs slightly worse than HOA O1. Increasing the order results in a higher rating, as seen in both HOS and HOA. However, there is little difference between HOS O4 and O7. Whilst the HOA O4 alternative has a slightly higher mean rating there is no statistical difference as both technique's means lie within confidence intervals of each other. The means considering the variation of each technique with the type of source are shown in Fig. 6.9(B). This follows a very similar trend to the overall mean ratings, however in general the drum signal was rated slightly lower for all techniques. This reinforces previous findings that drums are a good critical signal compared to speech for listening experiments [149].

Finally, mean ratings for each technique over the three source positions are shown in Fig. 6.9(C). This result shows the largest variation, and supports that the position of the source has a significant effect on the ratings of each rendering approach. Similar trends with respect to the order of each technique are observed. Interestingly, for a horizontal source position the mean rating for both HOS O4 and O7 exceeds that of HOA O4, as well as when comparing HOS O1 to HOA O1. As the elevation of the source increases the mean scores for all HOS renderers decrease, whilst the HOA O1 technique increases. The HOA O4 mean scores are generally consistent across all three positions.

Post-hoc paired comparisons using a Bonferroni adjustment were also run. Considering varying the technique type these results showed a significant difference for all combinations of techniques except HOS O4 and O7, suggesting there is no difference between the performance of these two approaches, which supports the trends seen in Fig. 6.8 and 6.9. The post-hoc tests for source stimuli and source position all showed significant effects.

6.4.4 Discussion

The aim of this listening test was to answer two questions: does HOS perform as well as HOA at the same order but with a lower number of audio channels, and is there any perceived improvement in using a higher order HOS approach than HOA if the number of audio channels is fixed?

The results of the listening test reveal that HOS exhibits a large dependency on the position of the sound source, performing worse as the elevation is increased. This is likely due to how the HOS approach can only reproduce axisymmetric HRTFs as shown in Section 6.3. This means some elevation specific cues, for example due to the pinna, can not be reproduced. One approach to mitigate this effect, which is suggested for future work, is to perform mixed order HOS where more spherical harmonic coefficients with m close to 0 (for example $m = \pm 1$ coefficients) are also included in the rendering. Doing so would allow non-axisymmetric HRTFs to be reproduced. How many and which extra spherical harmonic channels should be added in this mixed order approach is a key question for the future work to investigate. The ability to perceive some level of elevation could be due to the inclusion of dynamic cues, which HOS can reproduce.

Dependency on source position is not seen with HOA O4. However, HOA O1 is observed to be rated higher as the elevation increases. This could be explained by these elevated source positions being close to a virtual loudspeaker position which when the listener has a pitch and roll rotation of zero is directly above the listener for HOA O1 only. With general head rotations this source would not necessarily be close to this loudspeaker, as the virtual loudspeaker stays fixed relative to the listeners head orientation. However it was observed that most participants focused on yaw head movements in the horizontal plane with just minor pitch and roll rotations. Therefore it is likely most participants would have experienced this source close to the virtual loudspeaker which could explain the apparent increase in score at these positions.

Increasing the order of both types of renderer led to an increase in the ratings. However, a limit was observed with HOS where the O7 renderer did not perform statistically differently to the O4 approach. Interestingly, this suggests a perceptual limit exists beyond which is not advantageous to increase the order of HOS. Furthermore, whilst there was a statistically significant difference between HOS and HOA when considering a fixed order (O1 and O4 for both approaches) the ratings were generally very similar, with HOA performing slightly better although this remains position dependent. This was expected, as explained previously HOS is rendering less information than HOA and from the theory would not be expected to perform better when the order is matched.

Considering the two research questions, the results suggest that HOS performs similarly but slightly worse to HOA of the same order for elevated sources, however

with the saving of a reduced number of audio channels and processing costs. Furthermore, it is advantageous to increase the order of HOS over HOA if there is a limit on the number of audio channels available, but not beyond O4. For example, it would be advantageous to use HOS O4 which requires $(N+1) = 5$ audio channels over HOA O1 with $(N+1)^2 = 4$ channels. However, HOS O7 does not yield an advantage over HOA O4. Furthermore, it is important to also consider that the virtual loudspeaker array for HOS renderers is considerably simpler to implement than for HOA, as just loudspeaker positions in front of the listener in the horizontal plane are required as opposed to full sphere sampling. This is a much simpler arrangement of HRTF positions to consider, particularly if the HRTF is to be measured for an individual person.

Therefore, overall HOS delivers promising results with some limitations particularly regarding the elevation of the virtual source. However, it brings the advantage of more efficient processing and easier to implement virtual loudspeaker arrays compared to HOA.

6.5 Chapter Review

This chapter has extended the HOS technique to include binaural rendering and has investigated the performance of the approach with the inclusion of a HRTF. First, the standard HOA B-format representation of a generalised soundfield through the PWD approach was revised, including the case when general and rigid sphere HRTFs are incorporated. Next, the HOS technique using a specific rotation and just the $m = 0$ spherical harmonic channels was shown to completely reproduce the rigid sphere HRTF. When considering a more complicated and realistic HRTF such as the Neumann KU100, the HOS approach (and in particular the rotation of the interaural axis to the z axis) was shown to reorder a considerable amount of the energy in the spherical harmonic coefficients of the HRTF to channels with m close to 0. Whilst this means channels with $m \neq 0$ are still required for exact rendering of the HRTF, unlike with the rigid sphere, the strength of the energy reordering indicates that up to certain order truncations the HOS approach will perform sufficiently well.

Numerical simulations indicate that rendering using spherical harmonic coefficients with $m = 0$ only results in axisymmetric HRTFs about the z axis, as is the case with the rigid sphere HRTF. Comparing HOA to HOS the key advantage of HOS is that it requires only $(N+1)$ channels for rendering as opposed to $(N+1)^2$. The simulations indicate that in some cases, including when the source is elevated, if the number of audio channels may not exceed a set limit then rendering to a higher truncation order using HOS may be more advantageous than using a lower HOA approach. However, as HOS can only reproduce an axisymmetric HRTF certain localisation cues such as pinna elevation cues are lost. A listening test comparing HOA and HOS techniques was also performed. The results indicate in general matching

the order of HOA and HOS resulted in similar scores by the participants, with HOS having the advantage of requiring less audio channels in the rendering stage. However, for elevated source positions the HOS approach performs worse than HOA. Increasing the order of both techniques led to an increased rating although beyond HOS order 4 no further improvements were noticed, suggesting a perceptual limit to the technique particularly when using virtual loudspeaker rendering.

Finally, a new technique is proposed for future investigation, mixed order HOS. Here it is suggested to add more spherical harmonic coefficients in the rendering. This will improve the accuracy of the HRTF rendering whilst still maintaining a lower channel count than HOA. Due to the energy focusing effect of the HOS rotation it is suggested that coefficients with m close to 0 would be most beneficial. Methods for deciding how many and which coefficients should be included is a key question for the future work. An obvious approach utilised in this work is by assessing the energetic contribution of each coefficient. However, other approaches such as perceptual evaluation could also be used to identify the most important coefficients to include.

Importantly, this initial work has not included the state-of-the-art in pre-processing of HRTFs for optimal performance when using spherical harmonic interpolation. Performing time alignment, such as the Mag-LS technique, at high frequencies above the $N = kr$ limit compacts the HRTF energy into lower orders. This means above the aliasing limit, better HRTF magnitude reproduction can be achieved although with the incorrect phase information, which is assumed to be perceptually unimportant. It is expected that these pre-processing techniques presented originally for HOA would be directly applicable to HOS and are suggested for future work, to improve the overall performance of the binaural HOS approach.

Chapter 7

Conclusions

This thesis has presented a new spatial audio reproduction technique titled Higher Order Stereophony (HOS). This work advances the field primarily through presenting this brand new approach establishing it theoretically, proving it experimentally and testing it subjectively. This new technique is a direct extension of the first ever approach for spatial audio reproduction, Stereophony, many years after its invention which birthed one of the most significant fields in audio research since, spatial audio. HOS is directly compatible with existing spatial audio techniques and formats which makes it applicable in a wide range of situations. The approach may be applied to a number of existing standardised loudspeaker arrangements, therefore may be readily adopted where other techniques were not traditionally applicable. HOS makes extensive use of dynamic listener tracking, which has been rarely applied to loudspeaker based reproduction but is currently the norm for headphone reproduction. HOS has also introduced the concept of soundfield reproduction across a restricted region, a single line. This could have profound effects in other areas of audio research, for example soundfield control and active noise control.

HOS is based on using the Taylor expansion of a plane wave soundfield, which results in terms to the n-th order of the soundfield's derivatives. Applying the expansion across a single axis only results in a representation of the soundfield along a line. Order matching, an approach analogous to mode matching, is used to define a soundfield reproduction technique that ensures accurate reproduction along a single line only. This differs from all other existing soundfield reproduction approaches which are often concerned with reproduction over a region, such as a circle or sphere. The fundamental assumption is that reproducing the soundfield correctly across a line that then coincides with the listener's interaural axis will lead to sufficiently accurate reproduction of the desired binaural signals.

The classic and most common stereo techniques such as the sine law, tangent law and head-tracked sine law were all derived through the framework and shown to be first order HOS systems. HOS was then demonstrated mathematically and through experiments to be the high order extension of these classic stereo techniques, allowing for the use of more loudspeakers and reproduction across a larger distance

along the reproduction line/to higher frequencies. The technique defines simple amplitude panning laws which are therefore easy to implement in a real-time scenario. Furthermore, due to the axial symmetry of the problem, the cone of confusion may be used to an advantage and loudspeakers in front on the listener only are required, resulting in more practical loudspeaker arrangements than other spatial audio techniques such as HOA. One key result of the approach is that for N-th order reproduction, only $(N+1)$ loudspeakers are required, a significant reduction compared to HOA which requires $(2N+1)$ or $(N+1)^2$ for 2D and 3D, respectively. Furthermore, with loudspeakers in front of the listener only it was shown that stable rear virtual sources could be achieved using HOS. HOS exhibits a similar frequency limitation as to HOA, following the $N = kr$ rule setting an upper frequency limit for accurate reproduction. Notably the evaluation of HOS over loudspeakers was limited to 2D loudspeaker arrays and source positions, although the extension to 3D was considered theoretically. An area for future work is therefore evaluating the ability of HOS to render elevated sources when using loudspeakers. A key research question is whether these sources can be perceived as elevated when using horizontal-only loudspeakers, or whether this is a perceptual limitation.

One restriction of the initial technique was the assumption that the listener's head orientation was both known and fixed. Therefore Dynamic HOS was proposed, which extended the technique to allow for any generalised listener head rotation through the use of a head-tracker and dynamic adaptation of the loudspeaker gains. This dynamic panning approach therefore rotates the line of accurate reproduction depending on the listener's look direction to ensure it always coincides with their interaural axis. One limitation of the approach is that head-tracking is generally required for it to work effectively perceptually, however listener head-tracking is now widely utilised in the field of spatial audio reproduction.

HOS was shown to suffer from an issue titled 'the instability condition', which occurs when a loudspeaker pair provide the exact same contribution to the plant matrix due to their angular positions for a specific listener head orientation, which can lead to the problem becoming ill-posed and thus loudspeaker gains which are impractical to implement. To avoid this issue two solutions were proposed: using a larger number of loudspeakers or using Tikhonov regularisation in the inversion of the plant matrix. Careful design of the loudspeaker array and restricting the listener's head rotation to a predetermined range can also be used to avoid the instability issues altogether.

HOS and HOA were demonstrated to share many similarities in their derivations, assumptions about the soundfield and reproduction system and also the behaviour of their solutions. From this fact, decoders to transform signals from the 3D or 2D HOA representation to HOS were derived. The decoders require a rotation of the HOA soundfield to align the listener's interaural axis along the z axis or the x axis (for 3D/2D HOA respectively), followed by reproduction of the $m = 0$ only

spherical harmonics or $\cos(n\theta)$ terms (again for 3D/2D HOA respectively). This demonstrates that all HOA content may be reproduced over a HOS system, where the decoder only maps a subset of the HOA signals to the HOS representation resulting in a more efficient rendering of the content compared. Direct comparison of 2D HOA and HOS through both experiments and subjective listening tests showed HOS could perform similarly compared to HOA when both are truncated to the same order, even when HOS used both less loudspeakers and loudspeakers arranged in front of the listener only. Notably HOS can utilise very simple loudspeaker arrays arranged in a frontal semi-circle, whereas HOA requires fully enclosing circular or spherical loudspeaker array arrangements. Furthermore, when the number of loudspeakers was fixed, HOS could be implemented to a higher truncation order than HOA which was shown to be advantageous.

Binaural rendering using HOS was also considered, and the HOS technique was shown to be able to fully reproduce the binaural signals due to a soundfield incident on a rigid sphere HRTF using only $(N+1)$ spherical harmonic channels. The rotation to align the interaural axis to the z axis was shown to be an energy reordering operation when representing the HRTF using spherical harmonics. This rotation was shown to condense the energy of a generalised HRTF towards spherical harmonic coefficients with m close to 0. Whilst for the rigid sphere only the $m = 0$ channels are required after this rotation (which is the full set of spherical harmonic coefficients that HOS ensures are rendered correctly), for a generalised HRTF other coefficients remain important.

This forms a key suggestion for future work. A mixed order approach that performs the HOS rotation and then keeps channels close to $m = 0$ could provide a good balance between the computational advantage of using HOS and its smaller number of audio channels compared to HOA, whilst also increasing the accuracy of the final rendered HRTF. Results from the subjective comparison suggested HOA outperforms HOS when the source is elevated. Thus including channels with $m \neq 0$ would solve the issue of HOS only being able to reproduce axisymmetric HRTFs, which could lead to HOS better rendering elevated sources. Future work could look at different approaches for deciding which subset of channels should be rendered, for example through energetic or perceptual evaluation.

The concept of mixed order reproduction could also be extended to the soundfield reproduction problem, not just binaural rendering. Mixed order HOA has been lightly investigated in the literature, however the motivation has often been favouring spherical harmonic channels that lead to greater resolution in the horizontal plane. Notably high resolution 2D HOA loudspeaker arrays are considerably more practical (with respect to loudspeaker numbers and positioning) compared to 3D HOA loudspeaker arrays. The understanding of HOS as soundfield reproduction across a single line leads to a large range of new possibilities with mixed order reproduction, where the soundfield can be rendered to different levels of accuracy in

different dimensions. This leads to a completely customisable region in which the soundfield is reproduced accurately. For example, 3D HOA reproduces the soundfield accurately across a sphere, whereas the mixed order HOS approach could lead to reproduction regions of more arbitrary generalised shapes (e.g a bubble elongated along a single axis).

Dynamic listener tracking is a key concept used in the implementation of HOS, with compensation for listener translations and rotations a core part of the technique. Interestingly, whilst listener tracking is currently widely used in binaural reproduction, it is not often employed for loudspeaker-based spatial audio. The exceptions are certain versions of Stereophony, some implementations of Crosstalk Cancellation and (more rarely) HOA where the sweet spot can be moved within a loudspeaker array interior. These approaches often use listener tracking as a compensation technique. However, this work has demonstrated that tracking can be used to render dynamic localisation cues which have proven to be very strong in particularly when solving front-back confusions. This suggests other spatial audio techniques could also benefit from employing listener tracking for similar reasons. Alternatively, listener tracking could be used to further extend HOS to render nearfield sources. This leads to a 6 degrees of freedom renderer adding distance to the virtual source, which is very applicable to recent trends in Virtual and Augmented Reality.

Due to the close similarities HOS exhibits to HOA another suggestion for future work is the adaption of many state-of-the-art techniques developed to improve HOA to HOS. For example, many techniques exist for improving the performance of irregular loudspeaker arrays such as ALLRAD, which could be applied in a dynamic manner to Dynamic HOS to combat the instability condition. Furthermore an extensive number of HRTF pre-processing and optimisation techniques exist, such as Mag-LS which ensures better performance above the $N = kr$ limit. Due to the rederivation of HOS using the spherical harmonics, such techniques should be easily adaptable and could greatly improve the perceptual performance of HOS.

Another area for future work not discussed in this thesis is how the technique might be applied to microphone arrays and soundfield capture. This thesis has focused on deriving the core HOS technique, as well as defining how it may be used for soundfield reproduction. Microphone arrays that capture soundfield derivatives could be used to directly encode a captured soundfield into a format applicable for rendering using HOS. This is one area in the literature where similar work has already been performed, however the resulting microphone signals were generally transformed to a spherical harmonic basis. Instead, the output of these arrays could be linked directly to the basis as defined through HOS. This would then result in native HOS microphone arrays, similar to how the classic mid-side recording technique is used for first order stereo. Here the advantage could be that less microphones are required in just a simple linear arrangement, in comparison to existing approaches for capturing a spatial soundfield often using spherical microphone arrays.


GENELEC
GENELEC

Appendix A

Generalised Low Frequency 3D Audio Reproduction Over Loudspeakers

Audio Engineering Society

Convention Paper

Presented at the 148th Convention
2020 May 25 – 28, Vienna, Austria

This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Generalised Low Frequency 3D Audio Reproduction Over Loudspeakers

Jacob Hollebon[1] and Filippo Maria Fazi[1]

[1] *Institute of Sound and Vibration Research, University of Southampton, SO171BJ, United Kingdom*

Correspondence should be addressed to Jacob Hollebon (J.Hollebon@soton.ac.uk)

ABSTRACT

There exist many different techniques to reproduce 3D audio over loudspeakers, each derived from different models and motivations. However, at low frequencies these varying reproduction methods appear more similar than previously thought. This paper produces an analytical analysis of the stereo sine law, head-tracked stereo sine law, stereo tangent law/vector base amplitude panning, crosstalk cancellation and first order Ambisonics. Many of these techniques are shown mathematically to be equal or subsets of each other, resulting in a more generalised theory for low frequency audio reproduction over loudspeakers. Finally, the performance of each of the reproduction methods is considered under a low frequency analysis framework derived from a soundfield reproduction perspective.

1 Introduction

Reproduction of 3D audio over loudspeakers has remained a key area of research for many years, resulting in the development of a wide range of different techniques for not only reproduction, but also recording, transmission and manipulation of 3D audio. The derivations of each of these techniques are often driven by different motivations or a given reproduction system. However, the overarching goal is always the same: to reproduce a virtual sound source at some given position around a listener that is not limited to the position of the physical reproduction loudspeakers.

Given the wide range of reproduction techniques that have been developed over the years, it is not unreasonable to assume that each technique is vastly different to another in both its performance and methods. However, this is not necessarily the case. In fact, certain reproduction methods may be derived from different approaches which end with the same result, for example the stereo sine law. Hence, looking from only one approach may hide the similarities between any given reproduction method. Often, these similarities exist only in a specific region, for example at low frequencies. In light of this, a more generalised theory for 3D audio reproduction over loudspeakers may be written, encompassing multiple reproduction techniques that were previously thought of as differing.

Comparing the theoretical basis of each of these techniques has been considered before, but often on a case-by-case basis. In [1] the distinction between three different categories of reproduction methods is made;

1. **Soundfield Reconstruction:** Reproducing physical properties of the soundfield over a region of space.

2. **Panning Techniques:** Panpot laws where knowledge of the virtual source position relevant to the reproduction loudspeakers is used to define the loudspeaker gains.

3. **Binaural Techniques:** Reproduction of the pressure at the listener's ears directly through headphones or crosstalk cancellation (CTC) loudspeaker systems.

This paper will show that methods which come under the categories of both 'Panning Techniques' and 'Binaural Techniques' are also methods of 'Soundfield Reconstruction'. There exists much literature considering the subjective performance of various reproduction methods. However, in this work the analysis is of the physical recreated soundfield compared to the target soundfield only.

The aim of this paper is to present a generalised approach and framework for the reproduction of 3D audio over loudspeakers at low frequencies. In doing so, it will be shown that many common reproduction methods are more similar than previously thought in this low frequency regime. First, an analysis framework is presented that utilises a soundfield reconstruction approach, considering reproduction of a plane wave virtual sound source about a point where a listener is situated. A low frequency assumption is made and the analysis is limited to 2D, but might be easily expanded to 3D. Next, six common reproduction techniques are introduced and the similarities between each method shown. The six techniques considered are the stereo sine law, the head-tracked stereo sine law (CAP), the stereo tangent law/2D VBAP, crosstalk cancellation and first order Ambisonics. Where applicable these techniques are proven to be identical, or a subset of another. Finally, a range of metrics are considered that demonstrate each method's ability to reproduce the target soundfield and the performance of each system in an ideal setup are simulated and compared.

2 Analysis Framework

To allow for direct comparisons between any given audio reproduction method, they must each be studied under the same conditions. The following framework considers a low frequency analysis of a soundfield reproduction problem which allows for considerable simplifications to be made. The framework is similar to that presented in [1] and [2]. The analysis is performed in 2D but may easily be expanded to 3D by also considering the elevation of a virtual source or loudspeaker, as opposed to only its azimuth.

Consider a listener positioned such that the head centre and left and right ears are situated at $\mathbf{x}_c, \mathbf{x}_l$ and $\mathbf{x}_r$ respectively, as in Fig. 1a. The centre of the head is fixed at this point in space, however it is allowed to rotate by the quantity θ_{rot}. The Head-Related Transfer Function (HRTF) of the listener is assumed to be that of a rigid sphere for plane wave sources, with a radius of a' [3]. The interaural axis is defined by the unit vector $\hat{\mathbf{n}}$, which points from the listener's head centre to the left ear. As no head translations are considered, $\hat{\mathbf{n}} = [-\sin(\theta_{rot}), \cos(\theta_{rot})]^T$. At low frequencies, when the wavelength is much larger than the radius of the head, the rigid sphere HRTF may be approximated with the shadowless head model with an enlarged head radius of $a = 3a'/2$ [4]. Hence, here a valid HRTF is that of two points in free space separated by $2a$.

At low frequencies, the pressure, P, of a soundfield about a point $\mathbf{x}_0$ may be approximated by a first order Taylor expansion such that [2]

$$P(\mathbf{x}) \approx P(\mathbf{x}_0) + \nabla P(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0). \tag{1}$$

Considering the HRTF of the listener the pressure at the left and right ears is

$$P(\mathbf{x}_{l,r}) \approx P(\mathbf{x}_c) + \nabla P(\mathbf{x}_c) \cdot (\mathbf{x}_{l,r} - \mathbf{x}_c). \tag{2}$$

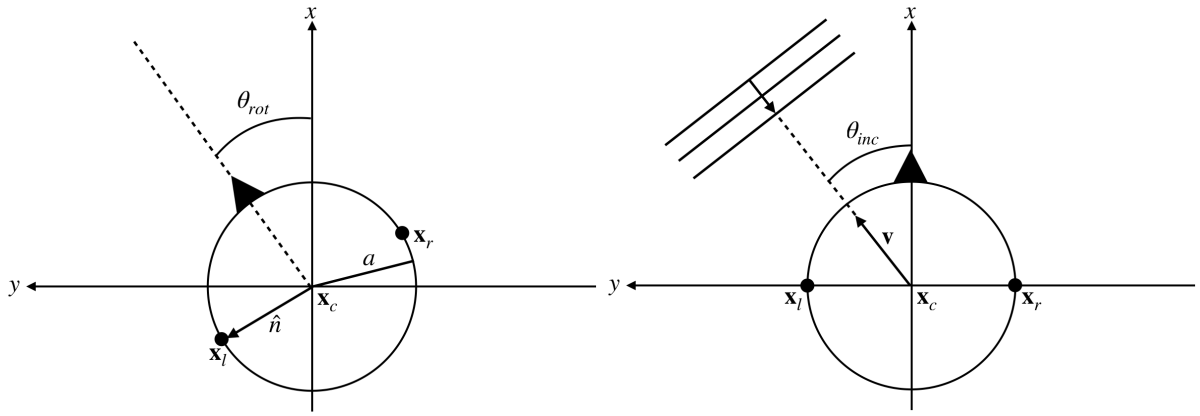
Euler's equation states that [5]

$$\nabla P(\mathbf{x}) = jkZ_0 \mathbf{v}(\mathbf{x}) \tag{3}$$

where k is the wavenumber, Z_0 is the characteristic impedance of the medium and $\mathbf{v}(\mathbf{x})$ is the particle velocity. Combining these definitions the pressure at the listener's ears is simply

$$P(\mathbf{x}_{l,r}) \approx P(\mathbf{x}_c) \pm jkaZ_0 \mathbf{v}(\mathbf{x}_c) \cdot \hat{\mathbf{n}}. \tag{4}$$

Hence, at low frequencies the binaural signals for a listener in a soundfield may be represented simply by the pressure and the particle velocity evaluated at the centre of the listener's head. Noting that $\mathbf{v}(\mathbf{x}_c) = [v_x(\mathbf{x}_c), v_y(\mathbf{x}_c)]^T$, when the listener's ears are aligned along the x or y axis, then $\mathbf{v}(\mathbf{x}_c) \cdot \hat{\mathbf{n}} = v_x(\mathbf{x}_c)$ or $v_y(\mathbf{x}_c)$ respectively. Another view is to consider a modal decomposition of the two ears of the listener as two degrees of freedom. Hence the signal may be split into an in-phase and an out-of-phase component. In this case, the in-phase mode is determined by the pressure at the centre of the head, as $P(\mathbf{x}_l) + P(\mathbf{x}_r) = 2P(\mathbf{x}_c)$. Conversely, the out-of-phase mode is dependent on frequency, the step size of the expansion and the particle velocity; $P(\mathbf{x}_l) - P(\mathbf{x}_r) \propto ka\mathbf{v}(\mathbf{x}_c) \cdot \hat{\mathbf{n}}$.



(a) Listener head orientation. **(b)** Plane wave incident on a listener.

Fig. 1: Geometry for the analysis framework.

2.1 Target Signal

In this work the target signal, or the virtual sound source, to be reproduced is that of a plane wave. Consider a plane wave with unitary amplitude $P_{inc} = 1$ incident with angle θ_{inc} to the centre of the listener's head, as in Fig. 1b. The plane wave produces a pressure and particle velocity

$$P(\mathbf{x}_c) = 1\ ,\ \ \mathbf{v}(\mathbf{x}_c) = \frac{1}{Z_0}\begin{bmatrix}\cos(\theta_{inc})\\ \sin(\theta_{inc})\end{bmatrix} \tag{5}$$

which results in the following binaural signals

$$P(\mathbf{x}_{l,r}) \approx 1 \pm jka\sin(\theta_{inc} - \theta_{rot}). \tag{6}$$

2.2 Reproduction System

To describe the reproduction system in the same framework, consider L loudspeakers positioned equidistantly around $\mathbf{x}_c$ that act as plane wave sources. The ℓth loudspeaker is at an angle γ_ℓ to the centre of the head and the loudspeakers are driven by gains $\mathbf{g} = [g_1, g_2, \dots g_L]^T$. The contributions from each of the loudspeakers at $\mathbf{x}_c$ sum coherently [6] hence

$$\begin{aligned} P(\mathbf{x}_c) &= \sum_{l=1}^{L} g_\ell \\ \mathbf{v}(\mathbf{x}_c) &= \sum_{l=1}^{L} g_\ell \mathbf{v}_\ell(\mathbf{x}_c) = \frac{1}{Z_0}\begin{bmatrix}\sum_{l=1}^{L} g_\ell \cos(\gamma_\ell)\\ \sum_{l=1}^{L} g_\ell \sin(\gamma_\ell)\end{bmatrix}. \end{aligned} \tag{7}$$

Under these assumptions, $P(\mathbf{x}_c)$ is always real-valued. Furthermore, if the gains are real-valued then $\mathbf{v}(\mathbf{x}_c)$ will also be real. The goal of the system is to correctly reproduce $P(\mathbf{x}_c)$ and $\mathbf{v}(\mathbf{x}_c)\cdot\hat{\mathbf{n}}$ such that it matches that of the given virtual sound source. Later, different definitions for the loudspeaker gains will be used to assess each system's ability to reproduce these quantities.

2.3 Performance Metrics

It is clear that two key performance metrics to analyse the reproduction systems are the quantities $P(\mathbf{x}_c)$ and $\mathbf{v}(\mathbf{x}_c)$. However, whilst reproducing all components of $\mathbf{v}(\mathbf{x}_c)$ will reproduce the soundfield in all directions about the centre of the head, it is only the quantity $\mathbf{v}(\mathbf{x}_c)\cdot\hat{\mathbf{n}}$ that matters to reproduce the virtual sound source. This is the projection of the particle velocity across the interaural axis. As will become clear, different reproduction methods take different approaches to reproducing $\mathbf{v}(\mathbf{x}_c)\cdot\hat{\mathbf{n}}$.

A direct map from Eqn. 6 to the reproduced Interaural Time Difference (ITD), a key low-frequency localisation cue, may also be obtained. At low frequencies, the ITD is primarily due to an Interaural Phase Difference (IPD) [7]. Hence

$$\mathrm{ITD} = \frac{\mathrm{IPD}}{\omega} = \frac{\angle P(\mathbf{x}_l) - \angle P(\mathbf{x}_r)}{\omega} \tag{8}$$

where ω is the angular frequency. By making a low frequency assumption that $ka \ll 1$, the ITD may be shown to equal

$$\mathrm{ITD} = \frac{2a}{c}\frac{\mathbf{v}(\mathbf{x}_c)}{P(\mathbf{x}_c)}\cdot\hat{\mathbf{n}} \tag{9}$$

where the derivation of the above is shown in the Appendix. The ITD, being dependent on *differences* between the two ears, is dominated by the out-of-phase mode, that is the factor $\mathbf{v}(\mathbf{x}_c)\cdot\hat{\mathbf{n}}$.

A final useful metric is the Makita localization vector, $\mathbf{r}_v$. The Makita vector was proposed by Gerzon [6] and arises from Makita's sound localization theory [8]. This states that at low frequencies below 700 Hz ,where ITD is the dominant localization cue, the direction a sound is perceived at is the angle perpendicular to the arriving wavefront. The Makita vector is a full description of the reproduced soundfield encompassed in the magnitude, $|\mathbf{r}_v|$, and direction, θ_v, of the vector. If $|\mathbf{r}_v| = 1$ this indicates a fully localized sound source whilst θ_v is related to the direction in which the virtual sound source is perceived. Both metrics are defined by

$$|\mathbf{r}_v| = Z_0\frac{|\mathbf{v}(\mathbf{x}_c)|}{|P(\mathbf{x}_c)|}\quad,\quad \theta_v = \arctan\left(\frac{v_y(\mathbf{x}_c)}{v_x(\mathbf{x}_c)}\right). \tag{10}$$

3 Stereophony

Stereophonic recording and reproduction introduced for the first time the concept of spatial audio. In general, two loudspeakers are positioned equidistantly from a listener, at angles $\gamma_{1,2} = \pm\gamma$ either side of the listener's head. A shadowless head model, plane wave virtual sources and that the loudspeakers radiate as plane waves are also assumed. Stereo is inherently a low frequency technique, valid below 800 Hz [9].

3.1 Sine Law

The stereo sine law considers a fixed listener head position and head rotation with the ears aligned along the y axis. There exists multiple ways to derive the sine law, one approach is to try to reproduce the ITD of the virtual source. The loudspeaker gains are given by [9]

$$\frac{g_1-g_2}{g_1+g_2} = \frac{\sin(\theta_{inc})}{\sin(\gamma)},\ \text{subject to}\ \ g_1+g_2 = 1. \tag{11}$$

Hence the individual loudspeaker gains are

$$\begin{aligned} g_1 &= \frac{1}{2}+\frac{\sin(\theta_{inc})}{2\sin(\gamma)}\\ g_2 &= \frac{1}{2}-\frac{\sin(\theta_{inc})}{2\sin(\gamma)}. \end{aligned} \tag{12}$$

The stereo sine law may also be formulated for any general loudspeaker positioning, as well as any number of loudspeakers. In theory, there is no limit to the position of the virtual source; it may be positioned inside or outside of the span of the loudspeakers.

3.2 Head-tracked Stereo Sine Law (CAP)

Whilst the stereo sine law considers a fixed listener position, the loudspeaker gains may be formulated to adapt for any listener movements or head rotations. This is the motivation for the head-tracked stereo sine law, also presented as Compensated Amplitude Panning (CAP) [10]. Naturally, this reproduction method requires tracking of the listener's head in comparison to the position of the loudspeakers. The loudspeaker gains are adapted depending on the listener's head rotation. To compensate for listener translations, delays may be applied to each of the loudspeakers so they are acoustically equidistant. Consider a standard stereo loudspeaker configuration as before, for a listener positioned equidistantly from both loudspeakers however now compensating for head rotations, then the gains are given by [10]

$$\begin{aligned} g_1 &= \frac{\sin(\gamma+\theta_{rot})+\sin(\theta_{inc}-\theta_{rot})}{2\sin(\gamma)\cos(\theta_{rot})}\\ g_2 &= \frac{\sin(\gamma-\theta_{rot})-\sin(\theta_{inc}-\theta_{rot})}{2\sin(\gamma)\cos(\theta_{rot})} \end{aligned} \tag{13}$$

where it is clear that for $\theta_{rot} = 0^\circ$ (the traditional stereo case) the gains give the stereo sine law.

3.3 Tangent Law and VBAP

An alternative approach for stereo loudspeakers is the stereo tangent law. In 2D, Vector Base Amplitude Panning (VBAP) is equivalent to the tangent law [11]. One starting point for the derivation of the tangent law is assuming the listener is positioned in the sweet spot of a stereo loudspeaker rig, however now the listener's head is assumed to be facing the position of the virtual source. In this case, the loudspeaker gains are [12]

$$\begin{aligned} g_1 &= \frac{1}{2}+\frac{\tan(\theta_{inc})}{2\tan(\gamma)}\\ g_2 &= \frac{1}{2}-\frac{\tan(\theta_{inc})}{2\tan(\gamma)}. \end{aligned} \tag{14}$$

Given that the tangent law assumes the listener is facing the virtual source, the head tracked sine law should be equal to the tangent law when $\theta_{rot} = \theta_{inc}$. Hence, combining this condition with Eqn. 13 and using the identity $\sin(x+y) = \sin(x)\cos(y) + \cos(x)\sin(y)$ then

$$\begin{aligned} g_1 &= \frac{\sin(\gamma)\cos(\theta_{inc}) + \cos(\gamma)\sin(\theta_{inc})}{2\sin(\gamma)\cos(\theta_{inc})} = \frac{1}{2} + \frac{\tan(\theta_{inc})}{2\tan(\gamma)} \\ g_2 &= \frac{\sin(\gamma)\cos(\theta_{inc}) - \cos(\gamma)\sin(\theta_{inc})}{2\sin(\gamma)\cos(\theta_{inc})} = \frac{1}{2} - \frac{\tan(\theta_{inc})}{2\tan(\gamma)} \end{aligned} \quad (15)$$

therefore the tangent law is a special case of the head-tracked sine law where the listener's head rotation follows the virtual source position.

4 Crosstalk Cancellation

Crosstalk cancellation (CTC) aims to correctly reproduce the pressure at the position of the listener's ears only. To do so, soundfield control by means of inverse filtering is employed to ensure sufficient channel separation between each of the listener's ears [13, 14]. Let M be the number of listener ears, or control points, and L be the number of loudspeakers. The reproduced pressure at the control points, $\mathbf{p}$, a column vector of length M, may be written as [15]

$$\mathbf{p} = \mathbf{C}\mathbf{g} = \mathbf{C}\mathbf{H}\mathbf{p}_T \quad (16)$$

where $\mathbf{C}$ is the $M \times L$ plant matrix of acoustic transfer functions that completely describe the reproduction system, the listener's HRTF and the geometry of the problem, $\mathbf{H}$ is the $L \times M$ matrix of CTC filters and $\mathbf{p}_T$ is the length M column vector of target pressures to be reproduced at each control point. The goal in CTC is to design the set of CTC filters, $\mathbf{H}$, such that it is as close to the inverse of $\mathbf{C}$ as possible so that $\mathbf{p}_T = \mathbf{p}$. In reality, this is complicated by issues such as listener and loudspeaker perturbations and ill-conditioning of the plant matrix under inversion [15].

CTC has previously been considered a completely different approach to soundfield reconstruction methods, as suggested in [1]. However, it has recently been shown under the same assumptions as this analysis, for a standard stereo loudspeaker arrangement the CTC loudspeaker gain solutions are equivalent to the head-tracked stereo sine law. The proof of this is shown in [16] and also reproduced in a slightly different format in the Appendix. This means at low frequencies CTC may also be thought of as a local soundfield recreation technique. Hence, the CTC gains at low frequencies under these assumptions equal Eqn. 13.

5 Ambisonics

Ambisonics is a full end-to-end theory covering the recording, transmitting and reproduction of a soundfield [17]. From an analytical view Ambisonics relies on the spherical harmonic expansion of a soundfield about a centre point [18]. As this expansion is an infinite sum of spherical harmonic terms, for a physical system it must be truncated to a finite number of terms, or 'order'. Increasing the order increases the size of the region where the field is correctly reproduced, which is also a function of frequency; hence a first order system is approximately accurate below 700 Hz [19]. Considering a first order system, physical meaning may be attached to each of the spherical harmonics, or 'B-format' terms. Commonly written as W, X, Y and Z these quantities are proportional to $P(\mathbf{x}_c), v_x(\mathbf{x}_c), v_y(\mathbf{x}_c)$ and $v_z(\mathbf{x}_c)$. For 2D reproduction, as will now be considered, the z terms may be ignored.

There exist two main strategies for finding the loudspeaker gains; physical decoding and psychoacoustic decoding [6]. Psychoacoustic decoding assumes the loudspeaker signals sum incoherently at the reproduction point, $\mathbf{x}_c$, and hence considers controlling the energy at this point. This decoding is most effective for high frequencies. For low frequencies, physical decoding is used where the loudspeaker signals are defined to correctly recreate the B-Format signals at $\mathbf{x}_c$.

Assuming the loudspeakers radiate as plane waves, the L loudspeakers contribute to the pressure and particle velocity at the reproduction point as defined in Eqn. 7 - often the normalised particle velocity, $\tilde{\mathbf{v}}$, is used which ignores the factor of $1/Z_0$. In this case the system may then be represented by the following equation

$$\mathbf{B} = \mathbf{\Psi}\mathbf{g}$$

$$\begin{bmatrix} P(\mathbf{x}_c) \\ \tilde{v}_x(\mathbf{x}_c) \\ \tilde{v}_y(\mathbf{x}_c) \end{bmatrix} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \cos(\gamma_1) & \cos(\gamma_2) & \dots & \cos(\gamma_L) \\ \sin(\gamma_1) & \sin(\gamma_2) & \dots & \sin(\gamma_L) \end{bmatrix} \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_L \end{bmatrix} \quad (17)$$

where $\mathbf{B}$ contains the B-Format signals and $\mathbf{\Psi}$ is a plant matrix containing the contribution of each loudspeaker

to the pressure and particle velocity about the reproduction point. Hence the loudspeaker gains are found by inversion of the plant matrix such that $\mathbf{g} = \mathbf{\Psi}^{-1}\mathbf{B}$. If $\mathbf{\Psi}$ is not square, the pseudoinverse is used.

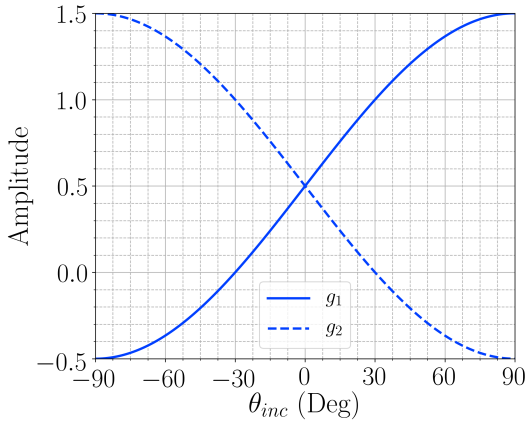
6 Simulations

To compare the performance of each of the reproduction techniques, simulations have been performed for virtual source positions $\theta_{inc} = [-90°, 90°]$ using the analytically formulated gain definitions from the previous sections combined with the low frequency analysis framework presented earlier in Eqn. 7. The techniques analysed are the stereo sine law, the stereo tangent law (inclusive of VBAP), the head-tracked stereo sine law (inclusive of CAP and low frequency CTC as they have been shown to be equal) and first order Ambisonics. For the three two-channel stereo setups, a traditional stereo loudspeaker arrangement of two equidistant loudspeakers positioned at angles $\gamma_{1,2} = \pm 30°$ is assumed. For the Ambisonic system three equidistant loudspeakers are required and assumed to be arranged at angles of $\gamma_{1,2,3} = 0°, 120°$ and $240°$ respectively. First, the listener's ears are assumed to be aligned along the y axis, i.e. $\theta_{rot} = 0$. Next, head rotation such that $\theta_{rot} = [-90°, 90°]$ is also considered. Each system's ability to reproduce the metrics presented in section 2.3 is then compared.

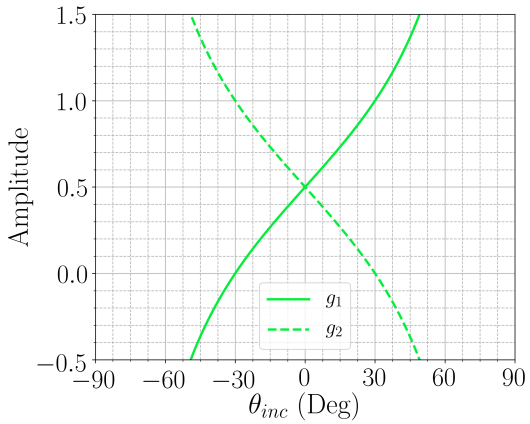
6.1 Results For $\theta_{rot} = 0°$

The loudspeaker gains for the sine law, tangent law and first order Ambisonics are shown in Fig. 2. The head-tracked sine law is not shown as here, where $\theta_{rot} = 0°$, it is identical to the stereo sine law. For all the reproduction methods when the virtual source is positioned at a loudspeaker only that loudspeaker is active, so that here the virtual source becomes a real source.

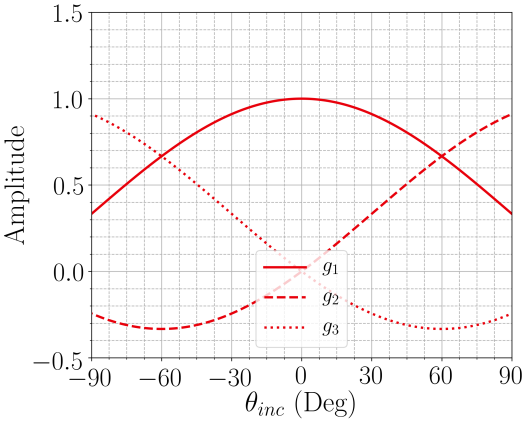
It is interesting to consider the virtual source positions that require negative (or out-of-phase) loudspeaker gains - such regions therefore require cancellation at the reproduction point which often results in a system less robust to perturbations of the loudspeakers or listener position. For both the sine and tangent stereo laws out-of-phase signals are only required when the virtual source is panned outside of the span of the loudspeakers. The sine law gains remain well-behaved for out-of-span panning, however the tangent law gains increase asymptotically indicating it is not a useful reproduction technique for virtual sources outside of the loudspeaker span. Finally, the Ambisonic gains require an out-of-phase loudspeaker at all times, except when the virtual source is at a loudspeaker position.



(a) Stereo sine law.

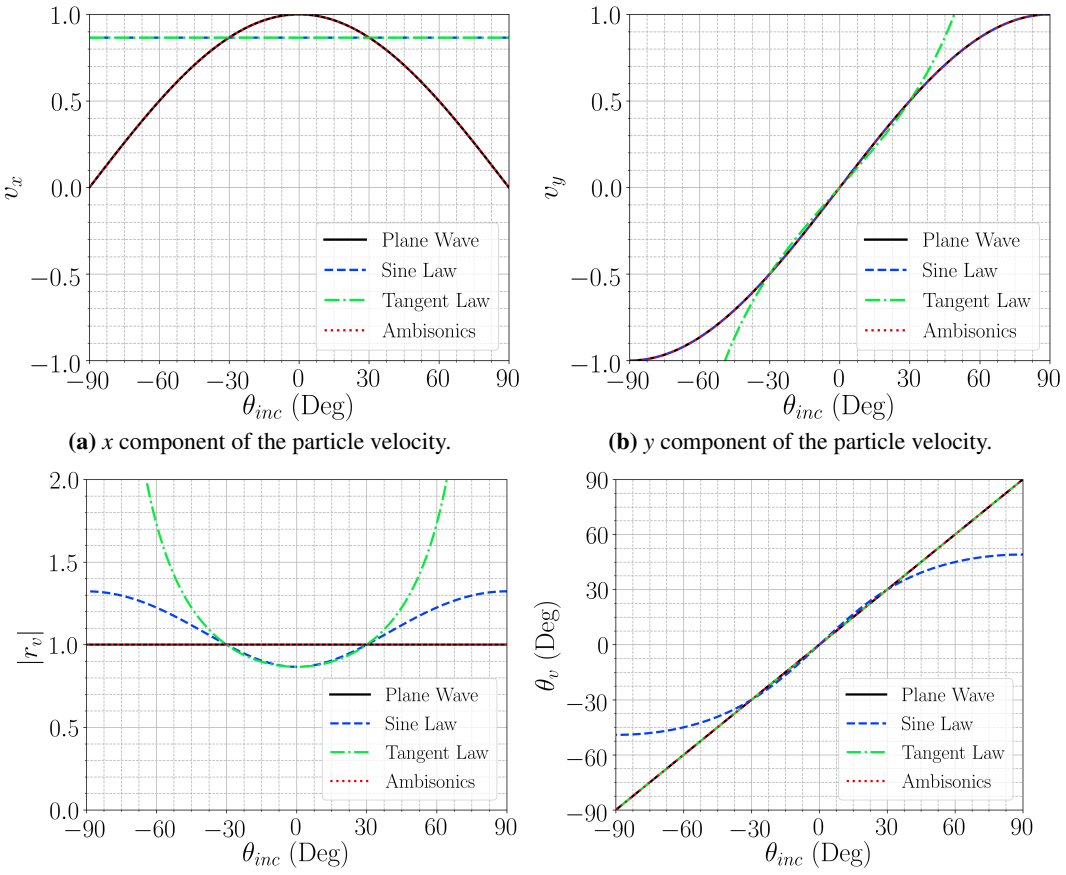


(b) Stereo tangent law.



(c) First order Ambisonics.

Fig. 2: Loudspeaker gains for varying virtual source position for $\theta_{rot} = 0°$.



(a) x component of the particle velocity.

(b) y component of the particle velocity.

(c) Magnitude of the Makita vector.

(d) Angle of the Makita vector.

Fig. 3: Reproduced properties of the soundfield for $\theta_{rot} = 0$.

An overview of the properties of the reproduced soundfield is presented in Table 1, compared to the plane wave target. Firstly, it is clear that the first order Ambisonic solution correctly reproduces all of the quantities. This means the virtual source is reproduced correctly regardless of the listener's head orientation.

All reproduction methods always reproduce $P(\mathbf{x}_c)$ correctly. However, the three stereo based solutions reproduce a constant value for $v_x(\mathbf{x}_c)$ which is governed by the loudspeaker span. This may be visualised by considering the geometric layout of a stereo loudspeaker pair. Both loudspeakers have the same x position, resulting in no degrees of freedom in this axis hence a constant value of $v_x(\mathbf{x}_c)$ regardless of the source position. This is seen visually in Fig. 3a. Considering $v_y(\mathbf{x}_c)$ it is clear that the sine law (and for this head orientation equivalently the head-tracked sine law/CTC) reproduces $v_y(\mathbf{x}_c)$ and therefore the ITD correctly for all virtual source positions. This means the sine law is a soundfield reproduction technique given a forward facing listener for any given virtual source position. However, the tangent law only correctly reproduces $v_y(\mathbf{x}_c)$ when $\theta_{inc} = \theta_{rot}$ which here is at 0°, i.e. when the listener is facing the virtual source. Despite this, considering Fig. 3b when the virtual source is within the loudspeaker span the tangent law approximately reproduces $v_y(\mathbf{x}_c)$ correctly, which means in practice it may still perform well.

Finally, considering the Makita vector for the sine and tangent law it is clear that $|\mathbf{r}_v|$ is only correct when the virtual source is at a loudspeaker. This is due to the incorrect reproduction of all components of the particle velocity vector $\mathbf{v}$. From Fig. 3c for in-span panning both the sine and tangent laws perform similarly, underestimating $|\mathbf{r}_v|$ as here $v_x(\mathbf{x}_c) = \cos(\gamma) < \cos(\theta_{inc})$. For out-of-span panning the sine law is better behaved and remains closer to the target value of 1. From Fig. 3d, θ_v is always perfectly recreated for the stereo tangent law, whilst the stereo sine law recreates θ_v well for in-span panning but poorly for out-of-span positions.

Method	$\mathbf{P}(\mathbf{x}_c)$	$\mathbf{v}_\mathbf{x}(\mathbf{x}_c)$	$\mathbf{v}_\mathbf{y}(\mathbf{x}_c)$	$\boldsymbol{\theta}_v$
Plane Wave Target	1	$\cos(\theta_{inc})$	$\sin(\theta_{inc})$	θ_{inc}
Stereo Sine Law	1	$\cos(\gamma)$	$\sin(\theta_{inc})$	$\arctan\left(\frac{\sin(\theta_{inc})}{\cos(\gamma)}\right)$
Stereo Tangent Law	1	$\cos(\gamma)$	$\cos(\gamma)\tan(\theta_{inc})$	θ_{inc}
Head-tracked Sine Law	1	$\cos(\gamma)$	$\cos(\gamma)\tan(\theta_{rot})+\frac{\sin(\theta_{inc}-\theta_{rot})}{\cos(\theta_{rot})}$	$\arctan\left(\frac{\sin(\theta_{inc}-\theta_{rot})}{\cos(\gamma)\cos(\theta_{rot})}\right)$
1^{st} Order Ambisonics	1	$\cos(\theta_{inc})$	$\sin(\theta_{inc})$	θ_{inc}

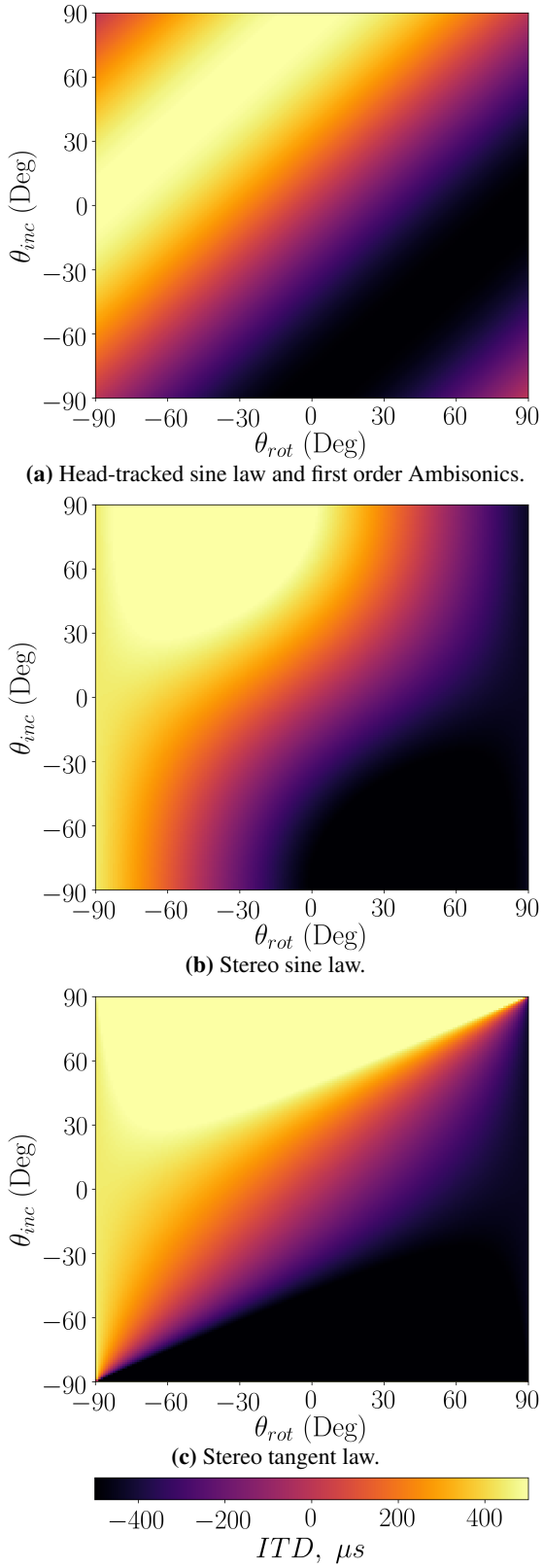
Method	$\lvert\mathbf{r}_\mathbf{v}\rvert$	$\mathbf{v}(\mathbf{x}_c)\cdot\hat{\mathbf{n}}$
Plane Wave Target	1	$\sin(\theta_{inc}-\theta_{rot})$
Stereo Sine Law	$\sqrt{\cos^2(\gamma)+\sin^2(\theta_{inc})}$	$\sin(\theta_{inc})\cos(\theta_{rot})-\cos(\gamma)\sin(\theta_{rot})$
Stereo Tangent Law	$\frac{\cos(\gamma)}{\cos(\theta_{inc})}$	$\tan(\theta_{inc})\cos(\gamma)\cos(\theta_{rot})-\cos(\gamma)\sin(\theta_{rot})$
Head-tracked Sine Law	$\sqrt{\cos^2(\gamma)+\left(\cos(\gamma)\tan(\theta_{inc})+\frac{\sin(\theta_{inc}-\theta_{rot})}{\cos(\theta_{rot})}\right)^2}$	$\sin(\theta_{inc}-\theta_{rot})$
1^{st} Order Ambisonics	1	$\sin(\theta_{inc}-\theta_{rot})$

Table 1: Comparison of the properties of the reproduced soundfield for the different audio reproduction methods. A green cell indicates correct reproduction, matching the plane wave target. A red cell indicates the quantity is only correct when the virtual source is real, located at a loudspeaker position.

6.2 Results With Head Rotation

So far a stationary head position has been considered. Now let the listener rotate their head such that $\theta_{rot} = [-90°, 90°]$. Henceforth the head-tracked sine law is considered separately to the sine law, as they are no longer identical. From Table 1 it is now relevant to consider the quantity $\mathbf{v}(\mathbf{x}_c)\cdot\hat{\mathbf{n}}$, as combining this with the correct $P(\mathbf{x}_c)$ ensures the right signals at the listener's ears and therefore the correct virtual source positioning. The head-tracked sine law/CTC and Ambisonic solutions are the only techniques to reproduce this quantity correctly, however through different means. Whilst the Ambisonic method reproduces the particle velocity correctly in all directions regardless of the listener's head orientation, the head-tracked sine law recognises the listener's head orientation so to only reproduce the particle velocity contributions required for that given head rotation. Due to this, the head-tracked sine law requires an accurate head-tracking implementation, however requires less loudspeakers than Ambisonics.

The reproduced ITD for the reproduction methods as a function of θ_{rot} is shown in Fig. 4. As the head-tracked sine law and Ambisonics solutions equate that of the original target plane wave indicating perfect reproduction this plot is only shown once. The sine law only completely recreates the ITD correctly when $\theta_{rot} = 0°$ however closely matches the correct ITD when either θ_{inc} or θ_{rot} is within the span of the loudspeakers - for any values outside of the loudspeaker span the solution is incorrect. For the tangent law, a straight line where $\theta_{inc} = \theta_{rot}$ is clearly seen indicating correct reproduction only when the listener faces the source, regardless of in-span or out-of-span panning. There again exists a region when θ_{inc} or θ_{rot} is within the loudspeaker span that the ITD is almost correctly recreated.



(a) Head-tracked sine law and first order Ambisonics.

(b) Stereo sine law.

(c) Stereo tangent law.

Fig. 4: Reproduced ITD as a function of head rotation.

7 Conclusions

In conclusion, a low frequency framework motivated by soundfield reproduction has been presented that considers a system's ability to recreate a virtual sound source as a function of the properties of the reproduced soundfield. These properties are simply the pressure and the particle velocity evaluated at the centre of the listener's head. The framework makes key assumptions to simplify the physics of the problem, which include a low frequency assumption, a rigid sphere HRTF for the listener and that the loudspeakers and target virtual source are plane waves. An important result is that to reproduce a virtual sound source it is only the pressure at the head centre and the projection of the particle velocity across the interaural axis that must be recreated - different reproduction methods can take different approaches to achieve this.

A range of different audio reproduction methods have been considered and compared within said framework in 2D. These techniques include the stereo sine law, the head-tracked stereo sine law (also known as CAP), the stereo tangent law (equivalent in 2D to VBAP), crosstalk cancellation (CTC) and first order Ambisonics. First, each of the techniques are introduced mathematically. The head-tracked CTC solution at low frequencies under the same assumptions and a stereo loudspeaker setup is shown to be equivalent to the head-tracked stereo sine law, which is in turn shown to be a subset of first order Ambisonics. The stereo sine law and stereo tangent law are then both shown to be special cases of the head-tracked sine law, under the conditions that the listener's head is rotated to be either facing forwards or facing the virtual sound source respectively. Hence a more generalised low frequency theory for 3D audio reproduction over loudspeakers has been presented, showing that these common 3D audio reproduction methods in this low frequency regime are all subsets of a soundfield reproduction approach.

Next the reproduction methods' performance were considered utilising simulations under a typical two-loudspeaker stereo rig for the sine, head-tracked sine and tangent laws, whilst for first order Ambisonics a three-loudspeaker rig is used as more loudspeakers are required. It is clear that the Ambisonics technique reproduces the soundfield exactly in this low frequency limit, however at the cost of requiring more loudspeakers and energy due to the amount of out-of-phase loudspeaker signals required. As Ambisonics reproduces

all components of the particle velocity, the listener is free to rotate their head and still perceive the correct virtual source position.

The head-tracked sine law acts as a subset of the Ambisonics technique, requiring fewer loudspeakers but with the addition of a head-tracker. This method ensures only the component of the particle velocity across the interaural axis is correctly recreated, which is the only component required to reproduce the virtual sound source. The stereo sine law assumes a stationary listener facing forwards and in doing so is a special case of the head-tracked sine law and therefore Ambisonics. Hence as the listener's head orientation is already known the required component of the particle velocity is also known, here the y component. Therefore, in practice with a well set-up loudspeaker array and the correct listener head orientation the stereo sine law should perform equally as well as first order Ambisonics. The stereo tangent law acts in a similar way, except assumes the listener rotates their head to always face the virtual source position. Hence the required particle velocity component is also known, which is the component perpendicular to the position of the virtual sound source.

For future work, the analysis is to be expanded to higher frequencies where many simplifications made in this paper for the analysis framework are no longer valid. Furthermore, the analysis could be combined with measurements and/or subjective tests to investigate the performance of each reproduction technique in real life.

References

[1] Jot, J.-M., Larcher, V., and Pernaux, J.-M., "A Comparative Study of 3-D Audio Encoding and Rendering Techniques," in *AES Conference: 16th International Conference: Spatial Sound Reproduction*, 1999.

[2] Menzies, D. and Fazi, F. M., "A Theoretical Analysis of Sound Localization, with Application to Amplitude Panning," in *AES Convention 138*, 2015.

[3] Morse, P. and Ingard, K., *Theoretical Acoustics*, International series in pure and applied physics, Princeton University Press, 1986.

[4] Duda, R. O. and Martens, W. L., "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, 104(5), pp. 3048–3058, 1998.

[5] Williams, E. G., *Sound Radiation and Nearfield Acoustical Holography*, Academic Press, London, 1st edition, 1999.

[6] Gerzon, M. A., "General Metatheory of Auditory Localisation," in *AES Convention 92*, 1992.

[7] Blauert, J., *Spatial Hearing*, MIT Press, Cambridge, England, revised edition, 1997.

[8] Makita, Y., "On the Directional Localisation of Sound in the Stereophonic Sound Field," in *E.B.U review*, 73, pp. 102–108, 1962.

[9] Bauer, B. B., "Phasor Analysis of Some Stereophonic Phenomena," *The Journal of the Acoustical Society of America*, 33(11), pp. 1536–1539, 1961.

[10] Menzies, D., Gálvez, M. F. S., and Fazi, F. M., "A Low-Frequency Panning Method With Compensation for Head Rotation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(2), pp. 304–317, 2918.

[11] Pulkki, V., "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc*, 45(6), pp. 456–466, 1997.

[12] Leakey, D. M., "Some Measurements on the Effects of Interchannel Intensity and Time Differences in Two Channel Sound Systems," *The Journal of the Acoustical Society of America*, 31(7), pp. 977–986, 1959.

[13] Damaske, P., "Head-Related Two-Channel Stereophony with Loudspeaker Reproduction," *The Journal of the Acoustical Society of America*, 50(4B), pp. 1109–1115, 1971.

[14] Schroeder, M. R. and Atal, B. S., "Computer simulation of sound transmission in rooms," *Proceedings of the IEEE*, 51(3), pp. 536–37, 1963.

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

[16] Hamdan, E. C. and Maria Fazi, F., "Low Frequency Crosstalk Cancellation and Its Relationship to Amplitude Panning," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 566–570, 2019.

[17] Gerzon, M. A., "Periphony: With-Height Sound Reproduction," *J. Audio Eng. Soc*, 21(1), pp. 2–10, 1973.

[18] Poletti, M. A., "Three-Dimensional Surround Sound Systems Based on Spherical Harmonics," *J. Audio Eng. Soc*, 53(11), pp. 1004–1025, 2005.

[19] Ward, D. B. and Abhayapala, T. D., "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Transactions on Speech and Audio Processing*, 9(6), pp. 697–707, 2001.

Appendix A Proof of Eqn. 9

Starting from the binaural signals defined in Eqn. 4, the phase of the pressure at each ear is

$$\angle P(\mathbf{x}_{l,r}) = \arctan\left(\frac{\pm kaZ_0\mathbf{v}(\mathbf{x}_c)\cdot\hat{\mathbf{n}}}{P(\mathbf{x}_c)}\right) \tag{18}$$

where the fact that both $P(\mathbf{x}_c)$ and $\mathbf{v}(\mathbf{x}_c)$ are real has been assumed. This is true for real-valued loudspeaker gains, as is the case for this work. For complex loudspeaker gains, $P(\mathbf{x}_c)$ may be made real by using $P(\mathbf{x}_c)$ as the definition point for a phase of zero, however $\mathbf{v}(\mathbf{x}_c)$ may be complex. In this case the derivation follows the same steps but results in a more complicated formula for the ITD.

Make a low frequency approximation, that is $ka \ll 1$, then by the small angle approximation $\arctan(x) \approx x$ hence the phase at each ear may be approximated as

$$\angle P(\mathbf{x}_{l,r}) \approx \pm kaZ_0\frac{\mathbf{v}(\mathbf{x}_c)}{P(\mathbf{x}_c)}\cdot\hat{\mathbf{n}} \tag{19}$$

and combining with the definition of the ITD given in Eqn. 8 then the proof is complete.

Appendix B Proof of Low Frequency CTC Gains

To define the target pressures, consider a plane wave incident on a rigid sphere for a forward facing listener. Using a low frequency approximation of the rigid sphere HRTF as before, then $P(\mathbf{x}_c) = e^{-jkr}$, $P(\mathbf{x}_{l,r}) = e^{-jk[r\pm a\sin(\theta_{inc})]}$ hence

$$\mathbf{p}_T = e^{-jkr}\begin{bmatrix} e^{jka\sin(\theta_{inc})} \\ e^{-jka\sin(\theta_{inc})}\end{bmatrix}. \tag{20}$$

The loudspeakers are modelled in the same way, as equidistant sources acting as plane waves, whilst using the same HRTF. Hence the plant matrix is simply

$$\mathbf{C} = e^{-jkr}\begin{bmatrix} e^{jka\sin(\gamma_1)} & e^{jka\sin(\gamma_2)} \\ e^{-jka\sin(\gamma_1)} & e^{-jka\sin(\gamma_2)}\end{bmatrix} \tag{21}$$

where the ℓth loudspeaker is incident with angle γ_ℓ. Consider a stereo arrangement for the loudspeakers, such that $\gamma_1 = \gamma$, $\gamma_2 = -\gamma$ and $M = L = 2$. The loudspeaker gains are given by $\mathbf{g} = \mathbf{H}\mathbf{p}_T = \mathbf{C}^{-1}\mathbf{p}_T$. Therefore using this definition for $\mathbf{C}$ then

$$\mathbf{g} = \frac{-j}{2\sin[2ka\sin(\gamma)]}\begin{bmatrix} 2j\sin\left(ka[\sin(\gamma)+\sin(\theta_{inc})]\right) \\ 2j\sin\left(ka[\sin(\gamma)-\sin(\theta_{inc})]\right)\end{bmatrix} \tag{22}$$

where the identities $\sin(x) = (e^{jx} - e^{-jx})/2j$ and $1/j = -j$ have been used. Finally, making a low frequency approximation such that $\sin(x) \approx x$ the loudspeaker gains are

$$\mathbf{g} = \begin{bmatrix} \frac{1}{2} + \frac{\sin(\theta_{inc})}{2\sin(\gamma)} \\ \frac{1}{2} - \frac{\sin(\theta_{inc})}{2\sin(\gamma)}\end{bmatrix} \tag{23}$$

which is the stereo sine law. The head tracked sine law may be derived in the same way, except for compensating for head rotation in each of the angle definitions, such that any given angle $\theta' = \theta - \theta_{rot}$.

Appendix B

Experimental Study of Various Methods for Low Frequency Spatial Audio Reproduction Over Loudspeakers

Experimental Study of Various Methods for Low Frequency Spatial Audio Reproduction Over Loudspeakers

1st Jacob Hollebon
Institute of Sound and Vibration Research
University of Southampton
Southampton, United Kingdom
J.Hollebon@soton.ac.uk

2nd Filippo Maria Fazi
Institute of Sound and Vibration Research
University of Southampton
Southampton, United Kingdom
Filippo.Fazi@soton.ac.uk

***Abstract*—In previous work it was shown that, at low frequencies, a number of audio reproduction techniques, including the stereo sine law, stereo tangent law, head-tracked sine law (or Compensated Amplitude Panning) and crosstalk cancellation all form a subset of first order Ambisonics and can therefore be analysed under one generalised framework. This paper expands and validates these results through objective experiments in an anechoic environment. For each of the spatial audio techniques the spherical harmonic components of the reproduced sound field of a single virtual source, as well as the reproduced binaural cues, are analysed and compared to those of an equivalent real source. The results of these measurements are in good agreement with the theoretical advances previously developed further validating, at low frequencies, the link between the otherwise disparate panning laws, binaural techniques and soundfield reconstruction approaches to spatial audio.**

***Index Terms*—Spatial Audio, Soundfield Reconstruction, Binaural, Ambisonics, Panning**

I. Introduction

Since the development of stereophony, spatial audio reproduction using loudspeaker arrays has been a key yet complex problem in the field that has resulted in the development of a broad range of reproduction techniques. Each of these approaches are derived using different assumptions about the soundfield and/or loudspeaker array, and generally each have their own advantages and disadvantages. As proposed in [1], these approaches may be classed into three broad categories:

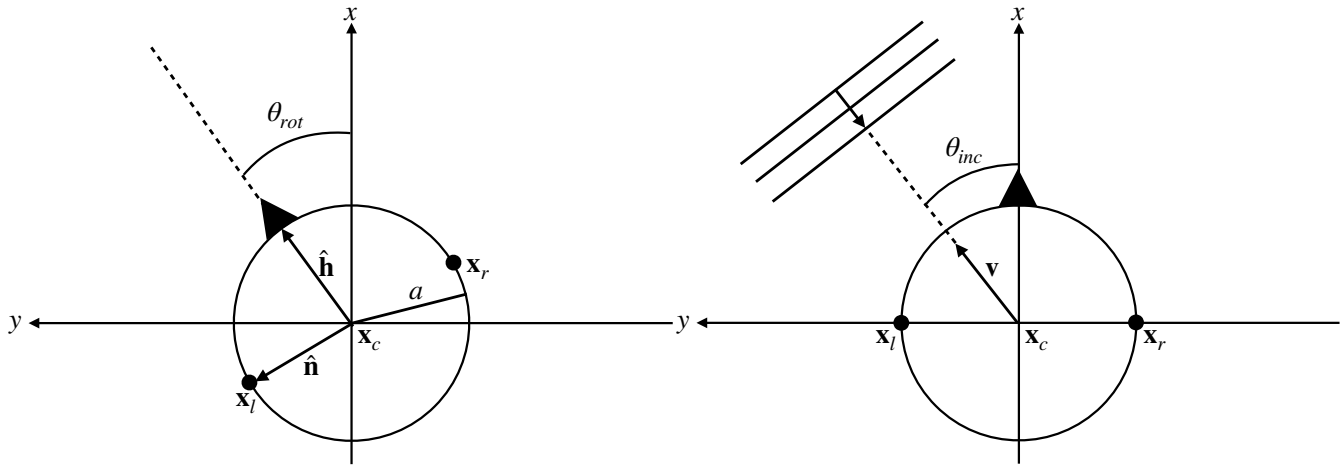
1) **Soundfield Reconstruction:** Reproducing physical properties of the soundfield over a region of space.
2) **Panning Techniques:** Panpot laws where knowledge of the virtual source position relevant to the reproduction loudspeakers is used to define the loudspeaker gains.
3) **Binaural Techniques:** Reproduction of the pressure at the listener's ears directly through headphones or crosstalk cancellation (CTC) loudspeaker systems.

Despite this, all techniques have the same goal - to create the illusion of one or more virtual sound sources at positions around the listener not limited to the physical positions of the reproduction loudspeakers. Furthermore, as these theories have been further developed, natural links between otherwise disparate reproduction techniques have arisen. For example, at low frequencies and assuming the virtual source and loudspeakers are plane waves, CTC has been shown to equal the stereo sine law solution [2]. Thus in certain frequency regions, a binaural approach has been shown to equal that of a panning technique. This is advantageous to understand, because at low frequencies CTC might thus be implemented in a simpler manner using panning as opposed to employing inverse filtering. Further links between other approaches have also been identified across the literature, but never unified under one framework to compare multiple techniques at the same time.

This paper is an extension of previous theoretical and simulation work proposing a generalised theory of low frequency spatial audio reproduction over loudspeakers [3]. The five key techniques covered previously and in this paper include the stereo sine law, the head-tracked stereo sine law (Compensated Amplitude Panning or CAP), the stereo tangent law, crosstalk cancellation, and first order Ambisonics. The aim of this work is to validate the previous theoretical contribution through experimental procedures. The paper is structured as follows. First, the results of the generalised low frequency spatial audio framework and the reproduction techniques it encompasses are reviewed. Next, the experimental set up, procedure and data processing techniques are presented. Following this, the results of the measurements are presented and analysed and finally conclusions and suggestions for future work are made.

II. Generalised Low Frequency Spatial Audio

This section is a summary of the previous theoretical work, and for full details of the derivations the reader is referred to the previous paper in [3]. The aim is to analyse a number of spatial audio techniques under the same analysis framework, to consider how and where they are similar or differ. The framework describes the contribution of each reproduction loudspeaker to the reproduced binaural signals at a listener's ears, with the aim of recreating a set of target binaural signals to create the illusion of a virtual source. The analysis is



(a) Listener head orientation.

(b) Plane wave incident on a listener.

Fig. 1. Geometry for the analysis framework.

performed in 2D, but might be easily expanded to 3D. The framework is similar to that presented in [1] and [4].

A low frequency assumption is made, such that the wavelength is much larger than the radius of the listener's head. Under these conditions, a good approximation of the Head-Related Transfer Function (HRTF) is of two points in free space with a head radius enlarged by a factor of $3/2$, which is in turn the low frequency approximation of a rigid sphere HRTF [5], [6]. The listener head orientation is shown in Fig. 1a. The left and right ears and head centre are defined by the positions $\mathbf{x}_l, \mathbf{x}_r$ and $\mathbf{x}_c$, respectively, whilst the unit vector $\hat{\mathbf{n}}$ points from $\mathbf{x}_c$ to $\mathbf{x}_l$, defining the interaural axis. To consider head rotations it is also convenient to define the unit vector $\hat{\mathbf{h}}$, which is orthogonal to $\hat{\mathbf{n}}$ and points from $\mathbf{x}_c$ in the direction the listener is facing. Thus the listener may rotate the head by an angle θ_{rot} which in turn defines both $\hat{\mathbf{n}}$ and $\hat{\mathbf{h}}$. These two vectors define a coordinate system that rotates with the listener's head rotation, the *listener frame of reference*.

Under these assumptions, the resulting binaural signals at the listener's ears are derived through combining a first order Taylor expansion about the centre of the listener's head and the Euler equation. Thus it may be shown that the binaural signals, $P(\mathbf{x}_{l,r})$ are approximately [3]

$$P(\mathbf{x}_{l,r}) \approx P(\mathbf{x}_c) \pm jkaZ_0\mathbf{v}(\mathbf{x}_c) \cdot \hat{\mathbf{n}} \tag{1}$$

where j is the imaginary unit, k is the wavenumber, a is the enlarged radius of the listener's head (enlarged by a factor of $3/2$), Z_0 is the characteristic impedance of the medium and $\mathbf{v}(\mathbf{x}_c) = [v_x(\mathbf{x}_c), v_y(\mathbf{x}_c)]^T$ is the particle velocity sampled at the centre of the listener's head. The term $\mathbf{v}(\mathbf{x}_c)\cdot\hat{\mathbf{n}}$ corresponds to the component of the particle velocity in the direction of the interaural axis, as defined by the orientation of the head due to θ_{rot}. In this sense, the only particle velocity component that contributes to the final binaural signals is that across the interaural axis. Therefore, to a first order approximation, the binaural signals are defined by this particle velocity component and the pressure at the centre of the head only.

A. Target Signals

The target virtual source is assumed to act like a plane wave with unitary amplitude and arrives from an incident angle θ_{inc}, as shown in Fig. 1b. In this case the field properties due to the plane wave are

$$P_T(\mathbf{x}_c) = 1, \quad \mathbf{v}_T(\mathbf{x}_c) = \frac{1}{Z_0}\begin{bmatrix}\cos(\theta_{inc})\\ \sin(\theta_{inc})\end{bmatrix} \tag{2}$$

where the subscript T indicates these are the target properties to be reproduced to create the virtual sound source. This leads to the following target binaural signals

$$P_T(\mathbf{x}_{l,r}) \approx 1 \pm jka\sin(\theta_{inc} - \theta_{rot}). \tag{3}$$

The goal of the spatial audio reproduction system is to recreate these signals by correctly reproducing the pressure at the listener's head centre and the particle velocity component across the interaural axis that matches the desired virtual source position.

B. Reproduction System

With the desired target signals defined, the reproduction system is now considered. The listener is assumed to be placed in the middle of a loudspeaker array, with L equidistant loudspeakers positioned at angles $\gamma_1, \gamma_2, \ldots \gamma_L$ respectively. The loudspeakers are also assumed to act as plane waves and are driven with gains $\mathbf{g} = [g_1, g_2, \ldots g_L]^T$. Assuming that the loudspeaker contributions sum coherently, a valid assumption due to the low frequency regime [7], the reproduced (subscript R) components are

$$P_R(\mathbf{x}_c) = \sum_{\ell=1}^{L} g_\ell$$
$$\mathbf{v}_R(\mathbf{x}_c) = \sum_{\ell=1}^{L} g_\ell \mathbf{v}_\ell(\mathbf{x}_c) = \frac{1}{Z_0}\begin{bmatrix}\sum_{\ell=1}^{L} g_\ell \cos(\gamma_\ell) \\ \sum_{\ell=1}^{L} g_\ell \sin(\gamma_\ell)\end{bmatrix}. \quad (4)$$

In this case, the reproduced binaural signals are given by

$$P_R(\mathbf{x}_{l,r}) \approx \sum_{\ell=1}^{L} g_\ell \pm jka \sum_{\ell=1}^{L} g_\ell \sin(\gamma_\ell - \theta_{rot}). \quad (5)$$

The goal of the reproduction system is to find a set of loudspeaker gains that correctly match the reproduced field to the target field, so that the variables $P(\mathbf{x}_c)$ and $\mathbf{v}(\mathbf{x}_c) \cdot \hat{\mathbf{n}}$ equal that of the desired virtual source. From hereon in, the pressure and particle velocity will only be considered at the centre of the head, hence the dependency on position of these quantities will be dropped for brevity, that is $P = P(\mathbf{x}_c)$.

C. Previous Results

The results of the previous theoretical work will now be presented. The following systems were compared in the framework above using the stated loudspeaker arrays:

1) Stereo Sine Law: **Panning Technique**. Standard stereo panning using a stereo pair of loudspeakers positioned at $\gamma_{1,2} = \pm\gamma$ [8], [9]. Assumes the listener faces forward at all times.

2) Stereo Tangent Law: **Panning Technique**. 2D formulation of Vector Base Amplitude Panning (VBAP) [10] using a stereo pair of loudspeakers positioned at $\gamma_{1,2} = \pm\gamma$ [11]. Assumes the listener faces the incident angle of the virtual source at all times.

3) Head-Tracked Stereo Sine Law: **Panning Technique**. Adaptive stereo panning expanding the sine law approach by utilising head-tracking to compensate for listener head rotations, using a stereo pair of loudspeakers positioned at $\gamma_{1,2} = \pm\gamma$. Also known as Compensated Amplitude Panning (CAP) [12].

4) Crosstalk Cancellation: **Binaural Technique**. Reproduction of the exact binaural signals at the listeners ears through the use of inverse filtering [13]–[15]. Assuming the loudspeakers and virtual source act as plane waves and head-tracking is used, at low frequencies it is identical to the head-tracked sine law [2]. Uses a stereo pair of loudspeakers positioned at $\gamma_{1,2} = \pm\gamma$.

5) First Order Ambisonics: **Soundfield Reconstruction/Panning Technique**. Mode matching approach through reproduction of the spherical harmonic components (the B-format signals) of a specified target soundfield [16], [17]. To first order, B-format signals are specified as W, X, Y, Z channels corresponding to P, v_x, v_y, v_z, respectively. In 2D the Z component is omitted and a three loudspeaker array may be used that uniformly samples a circle.

For this work, CAP and CTC are utilised in setups where they have been shown to be identical, and thus the common name, the 'head-tracked sine law', is used to refer to both techniques. Furthermore, the B-format channels W, X, Y, Z might be used interchangeably with the soundfield properties P, v_x, v_y, v_z. Strictly speaking, however, the velocity components and the B-format channels are only proportional and not directly equal.

Method	**P**	v_x	v_y	$\mathbf{v} \cdot \hat{\mathbf{n}}$
Stereo Sine Law	✓	×	✓	×
Stereo Tangent Law	✓	×	×	✓ only if $\theta_{rot} = \theta_{inc}$
Head-tracked Sine Law	✓	×	×	✓
First Order Ambisonics	✓	✓	✓	✓

TABLE I
RESULTS FROM THE PREVIOUS THEORETICAL STUDY. GREEN TICKS INDICATE CORRECT REPRODUCTION OF THE VALUE TO MATCH A PLANE WAVE TARGET. RED CROSSES INDICATE INCORRECT REPRODUCTION.

A summary of the results is shown in Table I, considering how each system reproduces the pressure at the head centre, the x and the y particle velocity component and also the projection of the particle velocity across the interaural axis. Two exceptions exist for these results. Firstly, with these loudspeaker array definitions all techniques activate a single loudspeaker when the virtual source is positioned at that loudspeaker position. When this occurs all soundfield properties are correctly reproduced as the virtual source becomes a real source. Secondly, the term $\mathbf{v} \cdot \hat{\mathbf{n}}$ is only signified as correct in the table if it works for any generalised head position. For example, if a technique reproduces v_y correctly then the quantity $\mathbf{v} \cdot \hat{\mathbf{n}}$ is only correct for one given head orientation $\hat{\mathbf{n}} = \hat{\mathbf{y}}$. Thus in the table this would result in a red cross as it does not hold for all head orientations, just when the interaural axis aligns along the y axis.

All approaches reproduce the pressure at the center of the head correctly. This may be viewed in their gain definitions as all use a gain normalisation such that $\sum_{\ell=1}^{L} g_\ell = 1$. The Ambisonics approach reproduces all components, as expected, as it is a soundfield reconstruction technique aiming to reproduce the exact soundfield properties under discussion. This allows the listener to rotate their head freely and always receive the correct particle velocity contribution across the interaural axis. However, in doing so Ambisonics requires more loudspeakers than the stereo solutions.

All stereo techniques take different approaches to reproducing the particle velocity components. However, they all only control one component and therefore may use one less loudspeaker than first order Ambisonics. The sine law reproduces just the y component correctly, as the gain derivation assumes the listener is always forward facing, with their ears aligned along the y axis. The tangent law dynamically changes across which direction the particle velocity is correct, by assuming the

listener is always facing the direction of the incident virtual source such that $\theta_{rot} = \theta_{inc}$. This becomes an issue when multiple sources from different directions are rendered, as the head orientation dictates a single source position for which the correct particle velocity projection across the interaural axis can be reproduced. Finally, the head-tracked sine law utilises head tracking and adapts the loudspeaker gains according to θ_{rot}, such that regardless of the direction the listener is facing the interaural particle velocity contribution is always correct thus obtaining the correct binaural signals.

Therefore, whilst some links between individual techniques have previously been formulated, these results link all the approaches in a generalised framework. Furthermore, whilst the derivations of each technique classifies them as either a panning, binaural or soundfield reconstruction approach, this demonstrates that at low frequencies all the methods might be considered as soundfield reproduction and also implemented using simple panning. Thus, the first order Ambisonic approach reproduces all components and provides the full soundfield reproduction. The head-tracked sine law is a subset of this idea, only reproducing the particle velocity component across the interaural axis through the use of head-tracking. Finally, the sine and tangent laws are special cases of the head-tracked approach where the head rotation is assumed to be either forward facing, or facing the position of the virtual source.

III. Experimental Validation

A. Measurements and Post-Processing

Measurements were performed to further validate these theoretical results. To ensure freefield conditions, the measurements took place in the anechoic chamber at the Institute of Sound and Vibration Research (ISVR), University of Southampton. A database of impulse responses (IRs) for a sound source positioned horizontally around a spherical microphone array was measured, such that any given 2D horizontal loudspeaker array might then be simulated. This database is made freely available for further research, and contains IRs up to Ambisonic order 4. A Genelec 8020C loudspeaker was used as the sound source whilst an Eigenmike EM32 spherical microphone array was used as the capture device, placed on an automated turntable, as shown in Fig. 2. The loudspeaker was placed 3 metres away from the microphone at the same height, so that it might approximate a plane wave source in the horizontal plane. IRs were measured using the sine sweep approach [18] for a 1 degree resolution in source position in the horizontal plane by rotating the microphone between each measurement, resulting in 360 measured source positions.

Following this, frequency-dependent windowing (FDW) [19], [20] was employed to remove small reflections due to the floor of the anechoic chamber. This was used as opposed to standard windowing to ensure a short window might be applied at high frequencies to remove the prominent reflections due to the chamber floor, whilst at low frequencies a longer window might be used to ensure the loudspeaker response would not be cut off. The FDW was applied using a Tukey

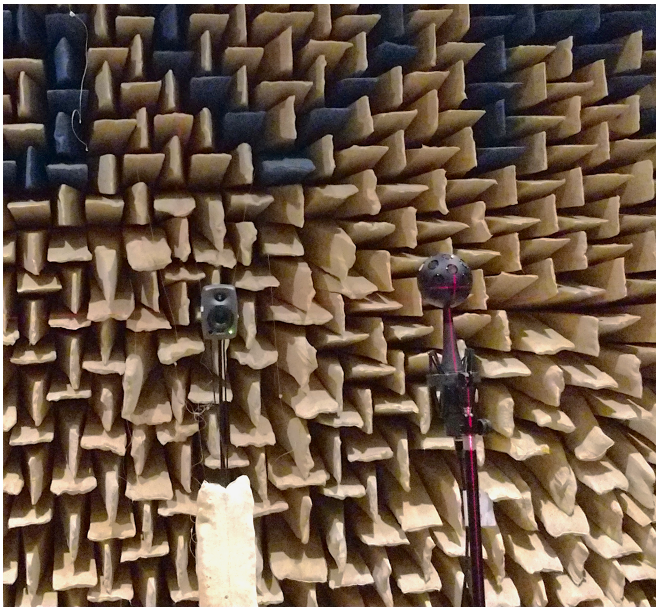
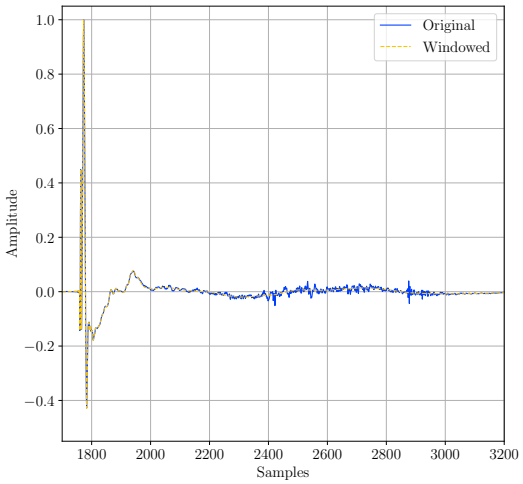


Fig. 2. Picture of the experimental apparatus.

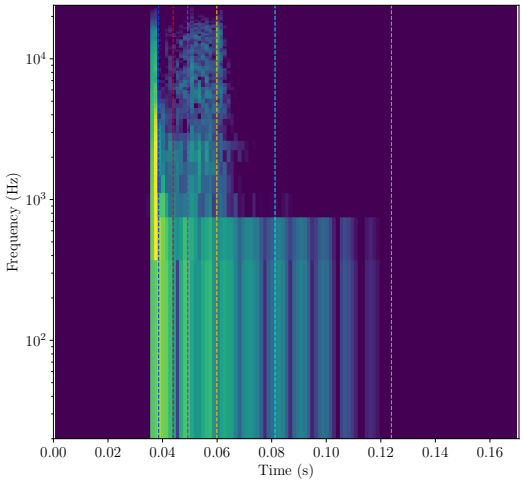
window, and the window length for five set cutoff frequencies was manually tuned and is shown in Table II. Unlike previous implementations of FDW, for all frequencies between these values the window length was calculated through linear interpolation so to ensure a smoothly decreasing window length as the frequency increased. This was done so to minimise artefacts at the otherwise abrupt change in window length. Much care was taken in the window design to ensure only reflections were removed. Fig. 3a demonstrates an example measured IR before and after FDW. The reflections observed later in the IR tail are removed and the envelope of the IR is retained, effectively smoothing this section of the IR. Fig. 3b and Fig. 3c show spectrograms of the two IRs respectively. It is clear that FDW removes the reflections which are prominently an issue above 700 Hz in the spectrogram. At low frequencies, however, the structure of the IR is maintained, unlike if traditional windowing had been used to remove the high frequency reflections. The trade-off to this technique is that some level of pre-ringing is introduced to the IR.

Following the windowing, the loudspeaker-to-microphone IRs were processed using the Eigenmike VST plugin to convert them to loudspeaker-to-B-format IRs. Only the first three B-format channels were retained - the W, X, Y channels corresponding to P, v_x, v_y. However, the full set of IRs up to order 4 are available in the online database. Thus from this measured IR database the contributions from any given horizontal loudspeaker array to the reproduced B-format components can be simulated.

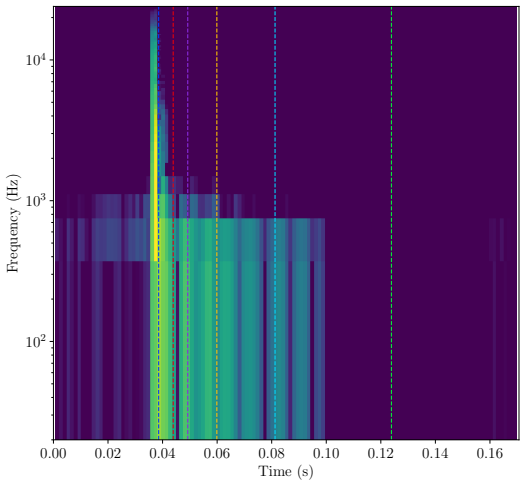
Finally, time alignment was applied across the measurement dataset to remove any sources of error due to positional offsets, for example in case the Eigenmike was not exactly centered on the turntable's axis of rotation. A variation across the dataset of approximately 4 samples was observed due



(a) Example IR before and after FDW.



(b) Spectrogram of the IR before windowing.



(c) Spectrogram of the IR after FDW.

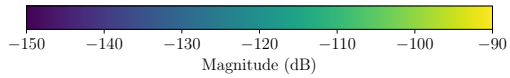


Fig. 3. Demonstration of the FDW technique. The coloured dashed lines indicate the windows as specified in Table II.

Cutoff Frequency (Hz)	Window Length (ms)	Window Length (samples)
0	85.3	4096
150	62.5	3000
250	25.0	1200
500	7.3	350
700	5.3	256

TABLE II
SET FREQUENCIES AT WHICH THE WINDOW LENGTHS WERE SPECIFIED. IN-BETWEEN THESE FREQUENCIES THE WINDOW LENGTHS WERE CALCULATED BY INTERPOLATING BETWEEN THE TWO ADJACENT CUTOFF FREQUENCIES.

to this issue. The procedure utilised was onset-based time alignment in a similar manner to [21], and is commonly used to remove the interaural time difference from head-related impulse response datasets. The time of arrival was calculated for each loudspeaker position using the W channel only - any delay on the W channel was assumed to be consistent across the X and Y channels. The time of arrival (TOA) was defined as when the IR passed a threshold of -20 dB relative to the main peak of that given IR, on the 10 times upsampled and lowpass filtered (8^{th} order butterworth, cutoff frequency at 3 kHz) IR of the W channel . Having identified the TOA's, these were converted into delays relative to the first measurement position with the loudspeaker at 0 degrees. Thus the absolute TOA was not removed from the IR and only any variation in TOA across the dataset was equalised. These delays were then removed using the fractional delay implementation in the SUpDEq toolbox [21].

B. Simulation Setup

Following the measurement of the IR dataset, simulations were performed utilising the measured data to reproduce the setup of different loudspeaker array geometries. A given loudspeaker reproduction system might be represented by a matrix equation, computed for each frequency and often referred to as the *forward problem*. This relates the vector of length L, whose components are the gains driving each loudspeaker $\mathbf{g} = [g_1, g_2, \dots g_L]^T$, to the reproduced properties of the soundfield, $\mathbf{p}_R$. Here $\mathbf{p}_R$ is a vector of three 3 elements corresponding to the reproduced W, X, Y components due to the loudspeaker array, at a given frequency. The forward problem is thus

$$\mathbf{p}_R = \mathbf{\Psi}\mathbf{g} \tag{6}$$

where $\mathbf{\Psi}$ is a $3 \times L$ plant matrix of transfer functions from each reproduction loudspeaker to the soundfield components.

Thus, different spatial audio systems might be simulated by using the corresponding measured transfer functions to populate the plant matrix for that given loudspeaker array. As all loudspeakers are always assumed to be equidistant only the parameter γ_ℓ defines each loudspeaker position.

The gain definitions are calculated through the corresponding reproduction approach under investigation.

The effects of listener head rotation might also be investigated by redefining the loudspeaker's angular positions (and loudspeaker gains if the technique utilises head-tracking). For a given head rotation θ_{rot}, a loudspeaker position is compensated such that $\gamma'_\ell = \gamma_\ell - \theta_{rot}$ and the plant matrix is adjusted correspondingly. In this case, the reproduced velocity components are $\mathbf{v} \cdot \hat{\mathbf{h}}$ and $\mathbf{v} \cdot \hat{\mathbf{n}}$ ensuring the particle velocity across the interaural axis is always considered. For a forward facing listener such that $\theta_{rot} = 0^\circ$ then $\mathbf{v} \cdot \hat{\mathbf{h}} = v_x, \mathbf{v} \cdot \hat{\mathbf{n}} = v_y$.

Following the simulation of the reproduced soundfield properties the complex reproduction error, ϵ, is used as a performance metric. The error is calculated for each reproduced element of $\mathbf{p}_R$, that is each individual soundfield property denoted by p_R. This is compared to $\mathbf{p}_T$, the target signals that are defined by the measurement of a loudspeaker arranged at the desired virtual source position. Thus

$$\epsilon = \frac{|p_R - p_T|^2}{|W_T|^2} \tag{7}$$

where p_T is the target signal for that given quantity (e.g. P, $\mathbf{v} \cdot \hat{\mathbf{h}}$, etc.) and the normalisation term is always the target W channel for that given desired source position. The error encompasses the reproduction of both the magnitude and phase of the soundfield components, and is often presented using a decibel scale.

A similar approach might also be used except with binaural transfer functions populating a plant matrix $\mathbf{\Psi_B}$. In this case $\mathbf{\Psi_B}$ is a $2 \times L$ matrix of binaural transfer functions from each loudspeaker to the two ears of a binaural microphone, whilst $\mathbf{p}_R$ is a 2-element vector of reproduced binaural signals. From these signals, localisation cues such as the interaural time difference (ITD) [22] might be calculated. For this work, far-field anechoic measurements of the KU100 binaural microphone were utilised [23] to create the matrix $\mathbf{\Psi_B}$. From these reproduced binaural signals the ITD calculated by interaural cross correlation was evaluated [24]. However, the binaural signals were first lowpassed at 600 Hz to ensure the ITD at low frequencies only was considered. A higher cutoff frequency than 600 Hz results in incorrect ITD recreation for all approaches, as the soundfield model used relies on a first order approximation, which is valid to approximately this cutoff frequency.

The techniques and loudspeaker arrays considered are as previously discussed and are summarised in Table III. A loudspeaker span greater than the standard ± 30 degree was utilised for the stereo approaches. This was to achieve a closer match to the Ambisonics array, for better comparison.

Method	**Loudspeaker Positions (degrees)**
Stereo Sine Law	± 60
Stereo Tangent Law	± 60
Head-tracked Sine Law	± 60
First Order Ambisonics	$\pm 60, 180$

TABLE III
SPATIAL AUDIO REPRODUCTION APPROACHES AND CORRESPONDING LOUDSPEAKER ARRAYS INVESTIGATED.

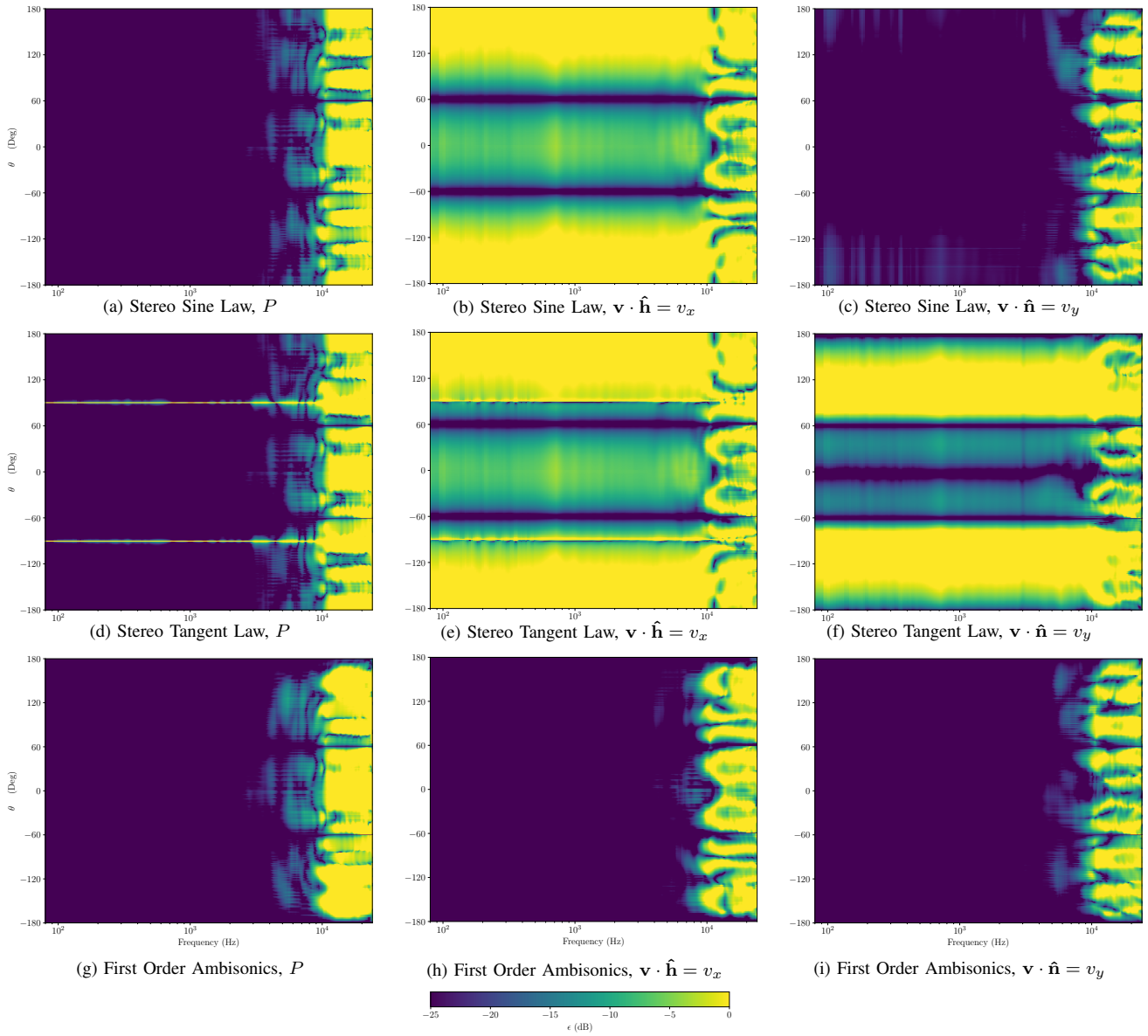
IV. RESULTS

A. Zero Head Rotation

The first scenario considered was that of a forward facing listener such that $\theta_{rot} = 0^\circ$. In this case, the head-tracked sine law and stereo sine law are exactly equal so the head-tracked sine law results are omitted. The results of the reproduced soundfield components due to the stereo sine law, stereo tangent law and first order Ambisonics attempting to recreate a virtual source positioned in the range $\theta_{inc} = [-180^\circ, 180^\circ]$ are shown in Fig. 4. The reproduction error ϵ (in decibels as per the colour bar) is presented as a function of the frequency (x-axis) and of the target virtual source position (y-axis). A fundamental limit of the Eigenmike might be observed above 6000 Hz, where spatial aliasing occurs, resulting in large errors for all techniques in this frequency band. Furthermore, there is zero error at all frequencies for a select few source positions. This is because with these loudspeaker arrays all techniques activate a single loudspeaker when the virtual source is at a loudspeaker position, thus here the virtual source becomes real and there is zero error in the reproduced soundfield. If more loudspeakers were used with these approaches, the problem would become underdetermined and this would no longer be the case as the minimum energy solution would favour activating more loudspeakers all the time [25].

As expected, all techniques correctly reproduce P but only up to approximately 4000 Hz. Above this limit, some levels of error are introduced by all the reproduction methods which might be due to incoherent summation of the loudspeaker contributions due to small misalignments from the measurement setup. Furthermore, the tangent law loudspeaker gains are undefined at $\theta_{inc} = 90^\circ$ thus the error is significant at this virtual source position.

For the particle velocity components, as anticipated the Ambisonics solution reproduces all elements correctly. Neither the stereo sine nor tangent law reproduce the v_x component correctly, except for at the loudspeaker positions. Notably, when the virtual source is within the loudspeaker span the error is small, suggesting there is some robustness to small head rotations in this region. The stereo sine law exhibits very little error in reproducing v_y, as this is the property the technique aims to portray correctly. This is due to the assumption that the listener is facing forwards, thus the interaural axis aligns with the y axis. Finally, the stereo tangent law reproduces the correct v_y component when $\theta_{inc} = 0^\circ$. This is because the tangent law assumes $\theta_{inc} = \theta_{rot}$ and thus reproduces the desired interaural particle velocity component for this condition only. As per the cone of confusion this v_y contribution



(a) Stereo Sine Law, P (b) Stereo Sine Law, $\mathbf{v} \cdot \hat{\mathbf{h}} = v_x$ (c) Stereo Sine Law, $\mathbf{v} \cdot \hat{\mathbf{n}} = v_y$

(d) Stereo Tangent Law, P (e) Stereo Tangent Law, $\mathbf{v} \cdot \hat{\mathbf{h}} = v_x$ (f) Stereo Tangent Law, $\mathbf{v} \cdot \hat{\mathbf{n}} = v_y$

(g) First Order Ambisonics, P (h) First Order Ambisonics, $\mathbf{v} \cdot \hat{\mathbf{h}} = v_x$ (i) First Order Ambisonics, $\mathbf{v} \cdot \hat{\mathbf{n}} = v_y$

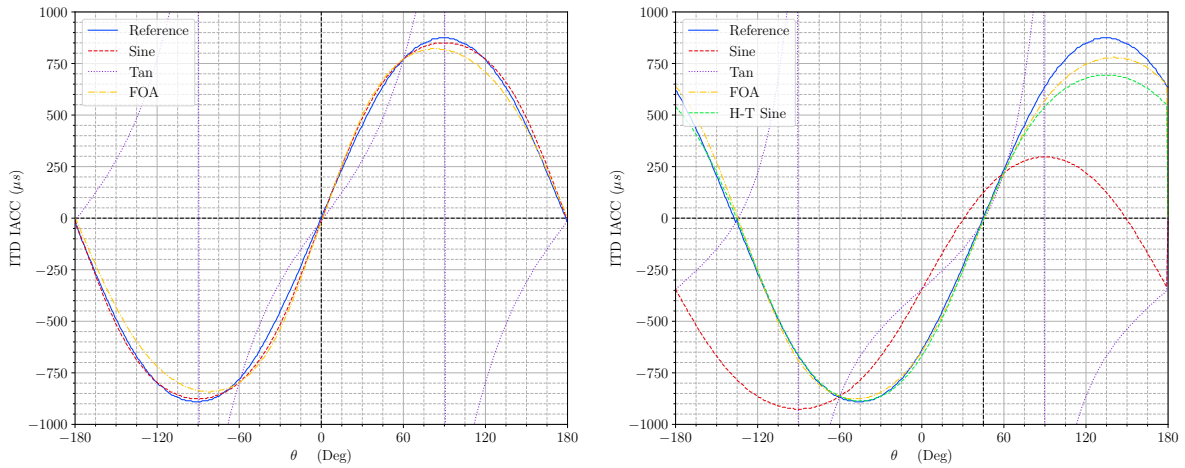
Fig. 4. Reproduction error for different soundfield components for $\theta_{rot} = 0°$.

is equivalent to a source from $\theta_{inc} = 180°$ also, so v_y is reproduced correctly for this incident angle as well.

Similar trends might be viewed in the reproduced ITD demonstrated in Fig. 5a. As it is only the v_y particle velocity component that contributes to the binaural signals for this head orientation, the reproduced ITD by both the sine and Ambisonic solutions is approximately correct if compared to the reference. Notably, the sine law requires one less loudspeaker than first order Ambisonics and both perform very similarly. However, interestingly the Ambisonics approach slightly underestimates the ITD when the source is positioned directly between some of the loudspeakers, around $\theta_{inc} = \pm 120°$. Furthermore, the tangent law only correctly reproduces the ITD at $\theta_{inc} = 0°$ when the listener faces the virtual source.

B. Including Head Rotation

Next, a head rotation was applied such that $\theta_{rot} = 45°$. To replicate this condition, all loudspeaker positions and their corresponding plant matrix entries were modified by the head rotation angle. For example, the stereo loudspeaker array became two loudspeakers positioned at $\gamma_{1,2} = 15°, 105°$. Furthermore, the loudspeaker gain definitions for the head-tracked sine law now differ to the stereo sine law, thus it is now included as a separate item in the analysis. The head-tracked sine law is the only approach considered in this work that utilises head-tracking and thus adapts the loudspeaker gains accordingly. As the listener is no longer facing forward, the projection of the particle velocity across the interaural axis, $\mathbf{v} \cdot \hat{\mathbf{n}}$, is no longer equal to v_y as per the reference geometry



(a) ITD for $\theta_{rot} = 0°$

(b) ITD for $\theta_{rot} = 45°$

Fig. 5. Interaural time differences calculated by interaural cross correlation for a range of incident source positions. The black dashed lines indicate the source position corresponding to an ITD of zero for that given head orientation.

in Fig. 1. Similarly $\mathbf{v} \cdot \hat{\mathbf{h}}$ is no longer equal to v_x.

The resulting reproduced soundfield components for the four spatial audio techniques are shown in Fig. 6. Once again, every approach produces the correct pressure P at the head centre up to 4 kHz. Moreover, as predicted the first order Ambisonic approach reproduces both particle velocity components correctly.

With the inclusion of head rotation the stereo sine and tangent law both perform poorly. The sine law does not reproduce any particle velocity component correctly, except at the loudspeaker positions. This is a result of assuming (now incorrectly) that $\mathbf{v} \cdot \hat{\mathbf{n}} = v_y$. The tangent law broadly reproduces the incorrect particle velocity components except for a few unique conditions. Following the assumption that $\theta_{rot} = \theta_{inc}$ then the proper interaural particle velocity projection is achieved when the source is positioned at $\theta_{inc} = 45°$. This is also seen with the corresponding source position on the cone of confusion where the $\mathbf{v} \cdot \hat{\mathbf{n}}$ contribution is identical, at $\theta_{inc} = 135°$ (a rotation of 180°). Interestingly, the orthogonal particle velocity contribution is also reproduced as desired for a few set source positions including when the source is positioned at a loudspeaker.

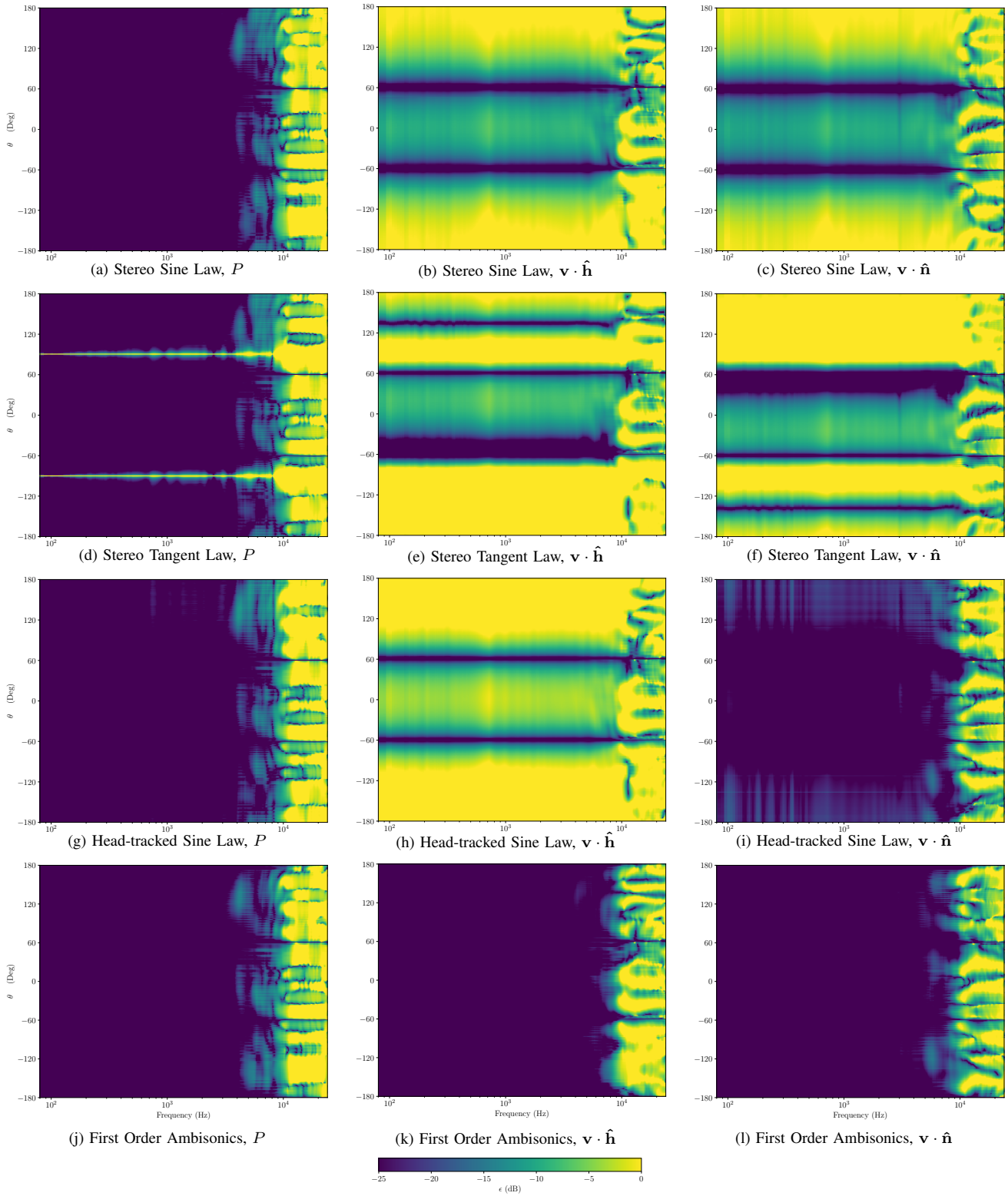
However, the head-tracked sine law always recreates the correct $\mathbf{v} \cdot \hat{\mathbf{n}}$ component. This is because the listener head position is tracked, so the loudspeaker gains are adapted to ensure that for any given head orientation the correct interaural particle velocity component is produced. As to a first order approximation this is the only relevant particle velocity component that contributes to the final binaural signals, the head-tracked sine law thus performs as well as first order Ambisonics, but with one less loudspeaker and with the inclusion of a listener head-tracker.

This is clearly seen in the corresponding ITD simulations in Fig. 5b. Here, the black dashed lines indicate at what virtual source position a zero ITD is expected, which for $\theta_{rot} = 45°$ is at $\theta_{inc} = 45°$. As before the stereo tangent law is correct at this zero ITD position when the listener faces the source. However, now the stereo sine law cannot recreate the desired ITD due to the listener head rotation. The head-tracked sine law and Ambisonic approach both match the reference ITD well, except around $\theta_{inc} = 135°$. This is when the source is positioned at a maximum lateral position to the listener for this head orientation.

V. Conclusions

This work has validated through anechoic measurements a theoretical generalised framework covering a number of spatial audio techniques at low frequencies. The techniques considered are the stereo sine law, stereo tangent law, head-tracked sine law (encompassing Compensated Amplitude Panning and crosstalk cancellation) and first order Ambisonics. All techniques are shown to be a form of soundfield reconstruction at low frequencies reproducing, to a first order approximation, the pressure and varying components of the particle velocity at the center of a loudspeaker array/the listener's head. The comparison is performed in 2D, but might readily be extended to 3D if required.

The theory was tested using measurements sampling the spherical harmonic components of a soundfield due to a loudspeaker in anechoic conditions. The measurements form a database of loudspeaker positions in the horizontal plane to a 1 degree resolution. The measurement data is processed using time alignment to remove any relative delays in the dataset due to misalignment, as well as frequency-dependent windowing to remove high frequency reflections whilst retaining the low frequency behaviour of the data. The database is freely available for future work, and includes data to Ambisonics order 4, however a truncation order of 1 was used in this work.



(a) Stereo Sine Law, P (b) Stereo Sine Law, $\mathbf{v} \cdot \hat{\mathbf{h}}$ (c) Stereo Sine Law, $\mathbf{v} \cdot \hat{\mathbf{n}}$

(d) Stereo Tangent Law, P (e) Stereo Tangent Law, $\mathbf{v} \cdot \hat{\mathbf{h}}$ (f) Stereo Tangent Law, $\mathbf{v} \cdot \hat{\mathbf{n}}$

(g) Head-tracked Sine Law, P (h) Head-tracked Sine Law, $\mathbf{v} \cdot \hat{\mathbf{h}}$ (i) Head-tracked Sine Law, $\mathbf{v} \cdot \hat{\mathbf{n}}$

(j) First Order Ambisonics, P (k) First Order Ambisonics, $\mathbf{v} \cdot \hat{\mathbf{h}}$ (l) First Order Ambisonics, $\mathbf{v} \cdot \hat{\mathbf{n}}$

Fig. 6. Reproduction error for different soundfield components for $\theta_{rot} = 45°$.

Using these measurements, the reproduced pressure and particle velocity due to each of the corresponding spatial audio techniques was analysed. The first order Ambisonics approach reproduces all elements of the pressure and particle velocity correctly, resulting in the listener being able to rotate their head freely and hear the correct binaural signals to reproduce a target virtual source. However, to achieve this Ambisonics requires more loudspeakers than the other approaches. The head-tracked sine law recognises that, in the low-frequency range considered in this work, only the projection of the particle velocity across the interaural axis is required to reproduce the binaural signals thus controls just this particle velocity component. This allows the listener to rotate their head freely as with Ambisonics but using less loudspeakers, however requiring a listener head-tracker implementation. The stereo sine and tangent law are shown to be special cases of the head-tracked sine law, under the conditions that the listener is facing forwards or facing the virtual sound source, respectively. Thus these techniques reproduce the correct particle velocity across the interaural axis only when these conditions are met.

A similar procedure was then performed using measurements of a KU100 binaural microphone to analyse the interaural time difference, a key low frequency binaural cue. These results also agree with the trends explained above.

Thus it has been shown that at low frequencies, the key spatial audio techniques widely used in the field are all a subset of a soundfield reproduction approach. Furthermore, the loudspeaker gain solutions in this low-frequency region are all simple panning laws for all techniques. Moreover, it is demonstrated how the use of head-tracking, an increasingly popular feature in spatial audio, may be utilised in an advantageous manner not just for headphone reproduction, but also loudspeaker reproduction of spatial audio. For future work, this analysis might be expanded to 3D considering that a number of the approaches have a natural 3D extension.

VI. License and Access

The measured impulse response dataset supporting this study is openly available from the University of Southampton repository at https://doi.org/10.5258/SOTON/D1857 under a Creative Commons CC BY license. The repository contains data beyond that used in the paper, including impulse responses up to Ambisonics order 4 and for varying loudspeaker radii of $1, 2$ and 3 metres.

References

[1] J.-M. Jot, V. Larcher, and J.-M. Pernaux, "A comparative study of 3-D audio encoding and rendering techniques," in *AES Conference: 16th International Conference: Spatial Sound Reproduction*, 03 1999.

[2] E. C. Hamdan and F. Maria Fazi, "Low frequency crosstalk cancellation and its relationship to amplitude panning," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 566–570.

[3] J. Hollebon and F. M. Fazi, "Generalised low frequency 3D audio reproduction over loudspeakers," in *AES 148th Convention*, 2020.

[4] D. Menzies and F. M. Fazi, "A theoretical analysis of sound localization, with application to amplitude panning," in *AES 138th Convention*, 2015.

[5] P. Morse and K. Ingard, *Theoretical Acoustics*. Princeton University Press, 1986.

[6] R. O. Duda and W. L. Martens, "Range dependence of the response of a spherical head model," *The Journal of the Acoustical Society of America*, vol. 104, no. 5, pp. 3048–3058, 1998.

[7] M. A. Gerzon, "General metatheory of auditory localisation," in *AES Convention 92*, Mar 1992.

[8] B. B. Bauer, "Phasor analysis of some stereophonic phenomena," *The Journal of the Acoustical Society of America*, vol. 33, no. 11, pp. 1536–1539, 1961.

[9] A. D. Blumlein, "British patent specification 394,325 (improvements in and relating to sound-transmission, sound-recording and sound-reproducing systems)," *J. Audio Eng. Soc*, vol. 6, no. 2, pp. 91–98, 1958.

[10] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc*, vol. 45, no. 6, pp. 456–466, 1997.

[11] D. M. Leakey, "Some measurements on the effects of interchannel intensity and time differences in two channel sound systems," *The Journal of the Acoustical Society of America*, vol. 31, no. 7, pp. 977–986, 1959.

[12] D. Menzies, M. F. S. Gálvez, and F. M. Fazi, "A low-frequency panning method with compensation for head rotation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 304–317, 2918.

[13] P. Damaske, "Head-related two-channel stereophony with loudspeaker reproduction," *The Journal of the Acoustical Society of America*, vol. 50, no. 4B, pp. 1109–1115, 1971.

[14] M. R. Schroeder and B. S. Atal, "Computer simulation of sound transmission in rooms," *Proceedings of the IEEE*, vol. 51, no. 3, pp. 536–37, 03 1963.

[15] O. Kirkeby, P. A. Nelson, H. Hamada, and F. Orduna-Bustamante, "Fast deconvolution of multichannel systems using regularization," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 189–194, 03 1998.

[16] M. A. Gerzon, "Periphony: With-height sound reproduction," *J. Audio Eng. Soc*, vol. 21, no. 1, pp. 2–10, 1973.

[17] F. Zotter and M. Frank, *A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer International Publishing, 2019, vol. 19.

[18] A. Farina, "Advancements in impulse response measurements by sine sweeps," in *Audio Engineering Society Convention 122*, May 2007.

[19] M. Karjalainen and T. Paatero, "Frequency-dependent signal windowing," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, 2001, pp. 35–38.

[20] F. Denk, B. Kollmeier, and S. D. Ewert, "Removing reflections in semianechoic impulse responses by frequency-dependent truncation," *J. Audio Eng. Soc*, vol. 66, no. 3, pp. 146–153, March 2018.

[21] J. M. Arend, F. Brinkmann, and C. Pörschmann, "Assessing spherical harmonics interpolation of time-aligned head-related transfer functions," *J. Audio Eng. Soc*, vol. 69, no. 1/2, pp. 104–117, Jan 2021.

[22] J. Blauert, *Spatial Hearing*, revised ed. MIT Press, Cambridge, England, 1997.

[23] B. Bernschütz, "A spherical far field hrir/hrtf compilation of the neumann ku 100," in *AIA-DAGA Conference on Acoustics*, 2013.

[24] B. F. G. Katz and M. Noisternig, "A comparative study of interaural time delay estimation methods," *The Journal of the Acoustical Society of America*, vol. 135, no. 6, pp. 3530–3540, 2014.

[25] E. C. Hamdan and F. Maria Fazi, "Three-channel crosstalk cancellation mode efficiency for sources in the far-field," in *AES Conference On Immersive and Interactive Audio*, 2019.

Appendix C

Properties Of The Spherical Bessel Functions

This derivation follows the results in [116]. The following properties of the Legendre polynomials are utilised. The Legendre polynomials form an orthogonal basis over the region $[-1,1]$ [115]

$$\int_{-1}^{1} P_n(x)P_{n'}(x)dx = \frac{2}{(2n+1)}\delta_{nn'} \tag{C.1}$$

whilst the parity condition leads to the following useful relation

$$P_n(-x) = (-1)^n P_n(x) \implies P_n(-1) = (-1)^n. \tag{C.2}$$

Begin with the plane wave expansion [118]

$$e^{-jkr\cos(\Theta)} = \sum_{n=0}^{\infty}(-j)^n(2n+1)j_n(kr)P_n(\cos\Theta). \tag{C.3}$$

and set the variables so that $kr = x$, $y = \cos Theta$

$$e^{-jxy} = \sum_{n=0}^{\infty}(-j)^n(2n+1)j_n(x)P_n(y). \tag{C.4}$$

Multiply by a dummy variable $P_{n'}(y)$ and integrate over the region $[-1,1]$ to utilise the orthogonality condition

$$\begin{aligned}\int_{-1}^{1} e^{-jxy}P_{n'}(y)dy &= \int_{-1}^{1}\sum_{n=0}^{\infty}(-j)^n(2n+1)j_n(x)P_n(y)P_{n'}(y)dy \\ &= (-j)^n(2n+1)j_n(x)\frac{2}{(2n+1)}\delta_{nn'}.\end{aligned} \tag{C.5}$$

Rearranging gives the integral representation of the spherical Bessel function (the Legendre polynomial to spherical Bessel transform)

$$j_n(x) = \frac{j^n}{2}\int_{-1}^{1} P_n(y)e^{-jxy}dy. \tag{C.6}$$

Next, consider the product of two spherical Bessel functions using the integral form above

$$j_n(x)j_{n'}(x) = \frac{j^{n+n'}}{4}\int_{-1}^{1}\int_{-1}^{1} P_n(y)P_{n'}(y')e^{-jxy}e^{-jxy'}dy'dy. \tag{C.7}$$

Integrating both sides with respect to x over $[-\infty,\infty]$ and using the relation $\int_{-\infty}^{\infty} e^{-j\alpha\beta}d\alpha = 2\pi\delta(\beta)$ (see [115])

$$\begin{aligned}\int_{-\infty}^{\infty} j_n(x)j_{n'}(x)dx &= \frac{j^{n+n'}}{4}\int_{-1}^{1}\int_{-1}^{1} P_n(y)P_{n'}(y')\int_{-\infty}^{\infty} e^{-jx(y+y')}dxdy'dy \\ &= \frac{j^{n+n'}}{4}\int_{-1}^{1}\int_{-1}^{1} P_n(y)P_{n'}(y')2\pi\delta(y+y')dy'dy \\ &= \pi\frac{j^{n+n'}}{2}\int_{-1}^{1}\int_{-1}^{1} P_n(y)P_{n'}(y')\delta(y+y')dy'dy.\end{aligned} \tag{C.8}$$

Due to the delta function the inner integral is non-zero when $y' = -y$ only. Using this fact, as well as the parity and the orthogonality of the Legendre polynomials then

$$\begin{aligned}\int_{-\infty}^{\infty} j_n(x)j_{n'}(x)dx &= \pi\frac{j^{n+n'}}{2}\int_{-1}^{1} P_n(y)P_{n'}(-y)dy \\ &= \pi(-1)^{n'}\frac{j^{n+n'}}{2}\int_{-1}^{1} P_n(y)P_{n'}(y)dy \\ &= \frac{\pi}{(2n+1)}\delta_{nn'}\end{aligned} \tag{C.9}$$

which defines the orthogonality relation for the spherical Bessel functions. To derive the integral representation of the Legendre polynomials (the spherical Bessel to Legendre polynomial transform) begin with the plane wave expansion, multiply by a dummy variable $j_{n'}(x)$, integrate with respect to x over $[-\infty,\infty]$ and use the orthogonality relation for the spherical Bessel functions above

$$\begin{aligned}\int_{-\infty}^{\infty} e^{-jxy} j_{n'}(x)dx &= \int_{-\infty}^{\infty} \sum_{n=0}^{\infty} (-j)^n (2n+1) j_n(x) j_{n'}(x) P_n(y) dx \\ &= (-j)^n \pi P_n(y)\end{aligned} \tag{C.10}$$

and finally rearranging gives

$$P_n(y) = \frac{j^n}{\pi} \int_{-\infty}^{\infty} j_n(x) e^{-jxy} dx. \tag{C.11}$$

The following relationship holds for negative arguments [117]

$$j_n(-x) = (-1)^n j_n(x). \tag{C.12}$$

Appendix D

Proof Of Regularised Pseudoinverse

The following proof demonstrates that when using Tikohonov regularisation, the left and right pseudoinverse definitions are identical.

D.1 Left Pseudoinverse

Let $\Psi_{M\times L}$ be an M by L matrix. The Singular Value Decomposition (SVD) is given by

$$\Psi_{M\times L} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H. \tag{D.1}$$

$\mathbf{\Sigma}$ is a diagonal matrix containing the non-negative and real singular values, σ, in order of decreasing magnitude. $\mathbf{U}$ and $\mathbf{V}$ are size $M \times M$, $L \times L$ respectively and are the sets of left and right singular vectors.

The left psuedoinverse with Tikhonov regularisation is

$$\Psi^{\dagger,left}_{L\times M} = \Psi^H_{L\times M}[\Psi_{M\times L}\Psi^H_{L\times M} + \beta\mathbf{I}_{M\times M}]^{-1} \tag{D.2}$$

with β a real, non-negative scalar. Employing the SVD

$$\begin{aligned}
\Psi^{\dagger,left} &= \mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^H[\mathbf{U}\mathbf{\Sigma}\mathbf{V}^H\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^H + \beta\mathbf{I}]^{-1} \\
&= \mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^H[\mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^T\mathbf{U}^H + \beta\mathbf{U}\mathbf{I}\mathbf{U}^H]^{-1} \\
&= \mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^H\mathbf{U}[\mathbf{\Sigma}\mathbf{\Sigma}^T + \beta\mathbf{I}]^{-1}\mathbf{U}^H \\
&= \mathbf{V}\mathbf{\Sigma}^T[\mathbf{\Sigma}\mathbf{\Sigma}^T + \beta\mathbf{I}]^{-1}\mathbf{U}^H.
\end{aligned} \tag{D.3}$$

Consider $\mathbf{\Sigma}^T[\mathbf{\Sigma}\mathbf{\Sigma}^T + \beta\mathbf{I}]^{-1}$. Note that

$$\begin{aligned}
\mathbf{\Sigma}_{M\times L} &= \begin{bmatrix} \sigma_1 & & \\ & \ddots & \mathbf{0} \\ & & \sigma_M \end{bmatrix} \\
\mathbf{\Sigma}^T_{L\times M} &= \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_M \\ & \mathbf{0} & \end{bmatrix} \\
\implies \mathbf{\Sigma}^T[\mathbf{\Sigma}\mathbf{\Sigma}^T + \beta\mathbf{I}]^{-1} &= \begin{bmatrix} \frac{\sigma_1}{\sigma_1^2+\beta} & & \\ & \ddots & \\ & & \frac{\sigma_M}{\sigma_M^2+\beta} \\ & \mathbf{0} & \end{bmatrix} \text{ an } L \times M \text{ matrix}
\end{aligned} \tag{D.4}$$

D.2 Right Pseudoinverse

Now consider the same applied to the right psuedoinverse with Tikhonov regularisation

$$\mathbf{\Psi}^{\dagger,right}_{L\times M} = [\mathbf{\Psi}^H_{L\times M}\mathbf{\Psi}_{M\times L} + \beta\mathbf{I}_{L\times L}]^{-1}\mathbf{\Psi}^H_{L\times M}. \tag{D.5}$$

Hence

$$\begin{aligned}
\mathbf{\Psi}^{\dagger,right} &= [\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^H\mathbf{U}\mathbf{\Sigma}\mathbf{V}^H + \beta\mathbf{I}]^{-1}\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^H \\
&= [\mathbf{V}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{V}^H + \beta\mathbf{V}\mathbf{I}\mathbf{V}^H]^{-1}\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^H \\
&= \mathbf{V}[\mathbf{\Sigma}^T\mathbf{\Sigma} + \beta\mathbf{I}]^{-1}\mathbf{V}^H\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^H \\
&= \mathbf{V}[\mathbf{\Sigma}^T\mathbf{\Sigma} + \beta\mathbf{I}]^{-1}\mathbf{\Sigma}^T\mathbf{U}^H.
\end{aligned} \tag{D.6}$$

Consider $[\mathbf{\Sigma}^T\mathbf{\Sigma} + \beta\mathbf{I}]^{-1}\mathbf{\Sigma}^T$

$$[\mathbf{\Sigma}^T\mathbf{\Sigma} + \beta\mathbf{I}]^{-1}\mathbf{\Sigma}^T = \begin{bmatrix} \frac{\sigma_1}{\sigma_1^2+\beta} & & \\ & \ddots & \\ & & \frac{\sigma_M}{\sigma_M^2+\beta} \\ & \mathbf{0} & \end{bmatrix} \text{ an } L \times M \text{ matrix} \tag{D.7}$$

Therefore

$$\mathbf{\Sigma}^T[\mathbf{\Sigma}\mathbf{\Sigma}^T + \beta\mathbf{I}]^{-1} = [\mathbf{\Sigma}^T\mathbf{\Sigma} + \beta\mathbf{I}]^{-1}\mathbf{\Sigma}^T \tag{D.8}$$

which means the singular values of the left and right psuedoinverse are equal. It follows that

$$\begin{aligned} \Psi^{\dagger,left} &= \Psi^{\dagger,right} \\ \text{as} \quad \mathbf{V}\mathbf{\Sigma}^T[\mathbf{\Sigma}\mathbf{\Sigma}^T + \beta\mathbf{I}]^{-1}\mathbf{U}^H &= \mathbf{V}[\mathbf{\Sigma}^T\mathbf{\Sigma} + \beta\mathbf{I}]^{-1}\mathbf{\Sigma}^T\mathbf{U}^H \end{aligned} \tag{D.9}$$

and thus when using Tikhonov regularisation, the left and right pseudoinverse are equivalent.

Appendix E

Derivation Of The Rigid Sphere HRTF

The rigid sphere is a useful approximation of a HRTF up to a certain frequency limit. This HRTF models the human head as a symmetric rigid sphere, with the ears situated at two diametrically opposed positions on either side of the head. The rigid sphere HRTF therefore includes estimations of range of localisation cues, importantly including estimates for the ITD and ILD [133, 154]. No model of the pinna is taken into account, therefore the model is useful up to approximately 4000 Hz where the pinna effects become prominent. The advantage of the rigid sphere HRTF is that it is a good approximation of an actual HRTF in a wide frequency range, whilst remaining an analytical model which may be exploited in the mathematics. In this thesis, the scattering due to a rigid sphere for plane wave sources is used, however point sources may also be considered. The analytical expression for this HRTF will be now derived.

For a rigid sphere of radius a in the presence of a plane wave source, the total pressure field , $p_{tot}(\mathbf{r}, \omega)$ evaluated at $r \geq a$ is a linear summation of the incident field, $p_i(\mathbf{r}, \omega)$ and the scattered field due to the rigid sphere, $p_s(\mathbf{r}, \omega)$ [115]

$$p_{tot}(\mathbf{r}, \omega) = p_i(\mathbf{r}, \omega) + p_s(\mathbf{r}, \omega). \tag{E.1}$$

This may be expressed as a summation of incoming and outgoing waves [115]

$$p_{tot}(\mathbf{r}, \omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \left(A_n^m(\omega) j_n(kr) + D_n^m(\omega) h_n^{(2)}(kr) \right) Y_n^m(\hat{\mathbf{r}}). \tag{E.2}$$

To find the coefficients A_n^m and D_n^m consider the boundary conditions due to the rigid sphere. At the rigid sphere boundary the particle velocity, $\mathbf{u}$, must be 0 so that $\mathbf{u}|_{r=a} = 0$. From the Euler equation the gradient of the pressure is related to its particle velocity for a plane wave by [115]

$$\nabla p = -j\omega\rho_0 \mathbf{u} \implies \nabla\, p_{tot}|_{r=a} = 0. \tag{E.3}$$

Applying this condition to Eqn. E.2 gives

$$\begin{aligned}\left.\frac{\partial p_{tot}}{\partial r}\right|_{r=a} = 0 &= \left[\sum_{n=0}^{\infty}\sum_{m=-n}^{n} k\left(A_n^m(\omega) j_n'(kr) + D_n^m(\omega) h_n^{(2)\prime}(kr)\right) Y_n^m(\hat{\mathbf{r}})\right]\Bigg|_{r=a} \\ &= \sum_{n=0}^{\infty}\sum_{m=-n}^{n} k\left(A_n^m(\omega) j_n'(ka) + D_n^m(\omega) h_n^{(2)\prime}(ka)\right) Y_n^m(\hat{\mathbf{r}}).\end{aligned} \tag{E.4}$$

Next, multiply by a dummy variable $Y_{n'}^{m'}(\hat{\mathbf{r}})^*$ and integrate over the unit sphere

$$\int_{\Omega}\sum_{n=0}^{\infty}\sum_{m=-n}^{n} k\left(A_n^m(\omega) j_n'(ka) + D_n^m(\omega) h_n^{(2)\prime}(ka)\right) Y_n^m(\hat{\mathbf{r}}) Y_{n'}^{m'}(\hat{\mathbf{r}})^* d\Omega = 0 \tag{E.5}$$

and by exploiting the orthogonality of the spherical harmonics

$$\begin{aligned}A_n^m(\omega) j_n'(ka) + D_n^m(\omega) h_n^{(2)\prime}(ka) &= 0 \\ \implies D_n^m(\omega) &= -A_n^m(\omega)\frac{j_n'(ka)}{h_n^{(2)\prime}(ka)}.\end{aligned} \tag{E.6}$$

Having now found a relation for the coefficients, the total pressure field may be expressed as

$$\begin{aligned}p_{tot}(\mathbf{r},\omega) &= \sum_{n=0}^{\infty}\sum_{m=-n}^{n} A_n^m(\omega)\left[j_n(kr) - \frac{j_n'(ka)}{h_n^{(2)\prime}(ka)} h_n^{(2)}(kr)\right] Y_n^m(\hat{\mathbf{r}}) \\ &= \sum_{n=0}^{\infty}\sum_{m=-n}^{n} A_n^m(\omega) R_n(kr) Y_n^m(\hat{\mathbf{r}}) \\ \text{where } R_n(kr) &= j_n(kr) - \frac{j_n'(ka)}{h_n^{(2)\prime}(ka)} h_n^{(2)}(kr).\end{aligned} \tag{E.7}$$

$R_n(kr)$ is defined as the n-th order radial filter for the rigid sphere. Finally, to get the pressure at the ear positions to define the HRTF, evaluate at $r = a$

$$p_{l,r}(\mathbf{r}_{l,r},\omega) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} A_n^m(\omega) R_n(ka) Y_n^m(\theta_{l,r},\phi_{l,r}). \tag{E.8}$$

For this special case, when $r = a$ the radial filter's definition may be simplified by using the Wronskian Relation [115]

$$j_n(x)h_n^{(2)\prime}(x) - j_n'(x)h_n^{(2)}(x) = -\frac{j}{x^2} \implies R_n(ka) = -\frac{j}{(ka)^2 h_n^{(2)\prime}(ka)}. \tag{E.9}$$

The final expression for the radial functions is thus

$$R_n(kr) = \begin{cases} j_n(kr) - \frac{j_n'(ka)}{h_n^{(2)\prime}(kr)} h_n^{(2)}(kr) & \text{if } r > a \\ -\frac{j}{(ka)^2 h_n^{(2)\prime}(ka)} & \text{if } r = a \\ \text{Undefined} & \text{if } r < a \end{cases}. \tag{E.10}$$

Bibliography

[1] F. Zotter and M. Frank. *A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Vol. 19. Springer International Publishing, 2019.

[2] Jean-Marc Jot, Veronique Larcher, and Jean-Marie Pernaux. "A Comparative Study of 3-D Audio Encoding and Rendering Techniques". In: *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*. Mar. 1999.

[3] H. Fletcher. "Symposium on wire transmission of symphonic music and its reproduction in auditory perspective: Basic requirements". In: *The Bell System Technical Journal* 13.2 (Apr. 1934), pp. 239–244.

[4] J. C. Steinberg and W. B. Snow. "Physical factors". In: *The Bell System Technical Journal* 13.2 (Apr. 1934), pp. 245–258.

[5] Alan D. Blumlein. "British Patent Specification 394,325 (Improvements in and relating to Sound-transmission, Sound-recording and Sound-reproducing Systems)". In: *J. Audio Eng. Soc* 6.2 (1958), pp. 91–98.

[6] Manfred R. Schroeder. "An Artificial Stereophonic Effect Obtained from a Single Audio Signal". In: *J. Audio Eng. Soc* 6.2 (1958), pp. 74–79.

[7] Benjamin B. Bauer. "Phasor Analysis of Some Stereophonic Phenomena". In: *The Journal of the Acoustical Society of America* 33.11 (1961), pp. 1536–1539.

[8] H. A. M. Clark, G. F. Dutton, and P. B. Vanderlyn. "The 'stereosonic' recording and reproducing system. A two-channel system for domestic tape records". In: *Proceedings of the IEE - Part B: Radio and Electronic Engineering* 104.17 (Sept. 1957), pp. 417–432.

[9] Mikko-Ville Laitinen et al. "Gain Normalization in Amplitude Panning as a Function of Frequency and Room Reverberance". In: *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*. Aug. 2014.

[10] Benjamin Bernfeld. "Simple Equations for Multichannel Stereophonic Sound Localization". In: *J. Audio Eng. Soc* 23.7 (1975), pp. 553–557.

[11] B. Xie. "Signal Mixing for a 5.1-Channel Surround Sound System - Analysis and Experiment". In: *J. Audio Eng. Soc* 49.4 (2001), pp. 263–274.

[12] Dylan Menzies, Marcos F. Simón Gálvez, and Filippo Maria Fazi. "A Low-Frequency Panning Method With Compensation for Head Rotation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.2 (2018), pp. 304–317.

[13] Dylan Menzies and Filippo Maria Fazi. "A Complex Panning Method for Near-Field Imaging". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.9 (2018), pp. 1539–1548.

[14] Dylan Menzies and Filippo Maria Fazi. "Surround Sound Without Rear Loudspeakers: Multichannel Compensated Amplitude Panning And Ambisonics". In: *Proceedings of the 21st International Conference on Digital Audio Effects (DAFx-18), Portugal*. 2018.

[15] Dylan Menzies and Filippo Maria Fazi. "A Theoretical Analysis of Sound Localization, with Application to Amplitude Panning". In: *Audio Engineering Society Convention 138*. May 2015.

[16] Dylan Menzies and Filippo Maria Fazi. "Implementation Of Dynamic Panning Reproduction With Adaption For Head Rotation". In: *Proceedings of the Institute of Acoustics: Reproduced Sound*. 2015.

[17] Dylan Menzies and Filippo Maria Fazi. "Spatial Reproduction Of Near Sources At Low Frequency Using Adaptive Panning". In: *Tecni Acustica - 46th National Congress on Acoustics*. 2015.

[18] Dylan Menzies and Filippo Maria Fazi. "Ambisonic Decoding for Compensated Amplitude Panning". In: *IEEE Signal Processing Letters* 26.3 (Mar. 2019), pp. 470–474.

[19] Dylan Menzies and Filippo Maria Fazi. "3D Ambisonic Decoding for Stereo Loudspeakers with Headtracking". In: *Audio Engineering Society Convention 146*. Mar. 2019.

[20] Dylan Menzies and Filippo Mari Fazi. "Small Array Reproduction Method For Ambisonic Encodings using Headtracking". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019.

[21] D. M. Leakey. "Some Measurements on the Effects of Interchannel Intensity and Time Differences in Two Channel Sound Systems". In: *The Journal of the Acoustical Society of America* 31.7 (1959), pp. 977–986.

[22] Ville Pulkki and Matti Karjalainen. "Localization of Amplitude-Panned Virtual Sources I: Stereophonic Panning". In: *J. Audio Eng. Soc* 49.9 (2001), pp. 739–752.

[23] Benjamin B. Bauer. "Broadening the Area of Stereophonic Perception". In: *J. Audio Eng. Soc* 8.2 (1960), pp. 91–94.

[24] Ville Pulkki. "Compensating Displacement of Amplitude-Panned Virtual Sources". In: *AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*.

[25] Gaál Dezsó. "Calculation of the Stereophonical Localization Area". In: *Audio Engineering Society Convention 53*. 1976.

[26] Thomas Lund. "Enhanced Localization in 5.1 Production". In: *Audio Engineering Society Convention 109*. 2000.

[27] Abhiram Lg, Aman Tayyab, and Ankith V. "Determination and Enlarging of the Acoustic SweetSpot in an Auditorium". In: *International Journal of Innovative Science and Research Technology* 2.7 (2017).

[28] Ronald M. Aarts. "Enlarging the Sweet Spot for Stereophony by Time/Intensity Trading". In: *Audio Engineering Society Convention 94*. 1993.

[29] Hyunkook Lee and Francis Rumsey. "Level and Time Panning of Phantom Images for Musical Sources". In: *J. Audio Eng. Soc* 61.12 (2013).

[30] H. Mertens. "Directional Hearing In Stereophony". In: *E.B.U. Review - Part A - Technical* 31.92 (1965), pp. 146–158.

[31] Hiroyuki Miyata and Shigeaki Aoki. "Stereo Reproduction with Good Localization in a Wide Listening Area". In: *Audio Engineering Society Convention 85*. 1988.

[32] Mark F. Davis. "Loudspeaker Systems with Optimized Wide-Listening-Area Imaging". In: *J. Audio Eng. Soc* 35.11 (1987), pp. 888–896.

[33] Sebastian Merchel and Stephan Groth. "Adaptively Adjusting the Stereophonic Sweet Spot to the Listener's Position". In: *J. Audio Eng. Soc* (2010).

[34] Sebastian Merchel and Stephan Groth. "Adaptive Adjustment of the "Sweet Spot" to the Listener's Position in a Stereophonic Play Back System - Part 1". In: *Int. Conf. on Acoustics (NAG/DAGA)*. 2009.

[35] Sebastian Merchel and Stephan Groth. "Analysis and Implementation of a Stereophonic Play Back System for Adjusting the "Sweet Spot" to the Listener's Position". In: *AES 126th Convention, Munich*. 2009.

[36] Sebastian Merchel and Stephan Groth. "Evaluation of a New Stereophonic Reproduction Method with Moving "Swet Spot" using a Binaural Localization Model". In: *Proceedings of the ISAAR, Copenhagen*. 2009.

[37] Sebastian Merchel and Stephan Groth. "Adaptive Adjustment of the "Sweet Spot" for Head Rotation". In: *Proceedings of the 22nd International Congress on Acoustics, ICA*. 2010.

[38] Marcos Felipe Simón Gálvez et al. "Object-Based Audio Reproduction Using A Listener-Position Adaptive Stereo System". In: *J. Audio Eng. Soc* (2016).

[39] Marcos Felipe Simón Gálvez, Dylan Menzies, and Filippo Maria Fazi. "A Listener Position Adaptive Stereo System For Object Based Reproduction". In: *AES 138th Convention*. 2015.

[40] Michael A. Gerzon. "Practical Periphony: The Reproduction of Full-Sphere Sound". In: *Audio Engineering Society Convention 65*. 1980.

[41] Michael A. Gerzon. "Periphony: With-Height Sound Reproduction". In: *J. Audio Eng. Soc* 21.1 (1973), pp. 2–10.

[42] Michael A. Gerzon. "General Metatheory of Auditory Localisation". In: *Audio Engineering Society Convention 92*. 1992.

[43] Mark A. Poletti. "Three-Dimensional Surround Sound Systems Based on Spherical Harmonics". In: *J. Audio Eng. Soc* 53.11 (2005), pp. 1004–1025.

[44] Jerome Daniel and Sebastien Moreau. "Further Study of Sound Field Coding with Higher Order Ambisonics". In: *Audio Engineering Society Convention 116*. 2004.

[45] J. Bamford. "An Analysis Of Ambisonic Sound Systems Of First And Second Order". PhD thesis. University Of Waterloo, 1995.

[46] J. Bamford and J. Vanderkooy. "Ambisonic Sound For Us". In: *Audio Engineering Society 99th Convention, New York*. 1995.

[47] J. Daniel. "Représentation de champs acoustiques, application á la transmission et la reproduction de scénes sonores complexes dans un contexte multimédia." PhD thesis. Université Paris, 2000.

[48] D. B. Ward and T. D. Abhayapala. "Reproduction of a plane-wave sound field using an array of loudspeakers". In: *IEEE Transactions on Speech and Audio Processing* 9.6 (2001), pp. 697–707.

[49] Mark Poletti. "Robust Two-Dimensional Surround Sound Reproduction for Nonuniform Loudspeaker Layouts". In: *J. Audio Eng. Soc* 55.7/8 (2007), pp. 598–610.

[50] J. Ahrens and S. Spors. "An analytical approach to sound field reproduction with a movable sweet spot using circular distributions of loudspeakers". In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2009, pp. 273–276.

[51] Filippo Maria Fazi and Philip A. Nelson. "A Theoretical Study Of Sound Field Reconstruction Techniques". In: *19th International Congress On Acoustics, Madrid*. 2007.

[52] Sascha Spors and Jens Ahrens. "A Comparison of Wave Field Synthesis and Higher-Order Ambisonics with Respect to Physical Properties and Spatial Sampling". In: *Audio Engineering Society Convention 125*. 2008.

[53] Jerome Daniel. "Spatial Sound Encoding Including Near Field Effect: Introducing Distance Coding Filters and a Viable, New Ambisonic Format". In: *Audio Engineering Society 23rd International Conference, Copenhagen, Denmark*. 2003.

[54] O. Kirkeby et al. "Fast deconvolution of multichannel systems using regularization". In: *IEEE Transactions on Speech and Audio Processing* 6.2 (Mar. 1998), pp. 189–194.

[55] Franz Zotter and Matthias Frank. "All-Round Ambisonic Panning and Decoding". In: *J. Audio Eng. Soc* 60.10 (2012), pp. 807–820.

[56] Audun Solvang. "Spectral Impairment of Two-Dimensional Higher Order Ambisonics". In: *J. Audio Eng. Soc* 56.4 (2008), pp. 267–279.

[57] Filippo Maria Fazi and Jacob Hollebon. "The Ring of Silence in Ambisonics and Binaural Audio Reproduction". In: *International Conference on Audio for Virtual and Augmented Reality*. 2022.

[58] B. Bernschütz et al. "Binaural Reproduction of Plane Waves With Reduced Modal Order". In: *Acta Acustica united with Acustica* 100.5 (2014), pp. 972–983.

[59] Johann-Markus Batke and Florian Keiler. "Investigation of Robust Panning Functions for 3-D Loudspeaker Setups". In: *Audio Engineering Society Convention 128*. 2010.

[60] Johann-Markus Batke and Florian Keiler. "Using VBAP-Derived Panning Functions For 3D Ambisonics Decoding". In: *Proc. of the 2nd Int. Conference on Ambisonics and Spherical Acoustics*. 2010.

[61] Fei Chen and Qinghua Huang. "Sparsity-based higher order ambisonics reproduction via LASSO". In: *IEEE China Summit and International Conference on Signal and Information Processing*. 2013, pp. 151–154.

[62] Gyanajyoti Routray and Rajesh M Hedge. "Sparsity Based Framework for Spatial Sound Reproduction in Spherical Harmonic Domain". In: *26th European Signal Processing Conference (EUSIPCO)*. 2018.

[63] Markus Noisternig et al. "A 3D Ambisonic Based Binaural Sound Reproduction System". In: *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. 2003.

[64] Boaz Rafaely and Amir Avni. "Interaural cross correlation in a sound field represented by spherical harmonics". In: *The Journal of the Acoustical Society of America* 127.2 (2010), pp. 823–828.

[65] Ville Pulkki. "Virtual Sound Source Positioning Using Vector Base Amplitude Panning". In: *J. Audio Eng. Soc* 45.6 (1997), pp. 456–466.

[66] Ville Pulkki. "Spatial sound generation and perception by amplitude panning techniques". PhD thesis. Helsinki University of Technology, 2001.

[67] V. Pulkki. "Uniform spreading of amplitude panned virtual sources". In: *Proceedings of the 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. WASPAA'99 (Cat. No.99TH8452)*. 1999, pp. 187–190.

[68] Jean-Marie Pernaux, Patrick Boussard, and Jean-Marc Jot. "Virtual Sound Source Positioning and Mixing in 5.1 Implementation on the Real-Time System Genesis". In: *In Proc. Conf. Digital Audio Effects (DAFx-98)*. 1998, pp. 76–80.

[69] Trond Lossius, Pascal Baltazar, and Théo de la Hogue. "DBAP - Distance-Based Amplitude Panning". In: *International Computer Music Conference (ICM)*. 2009.

[70] Dimitar Kostadinov, Joshua D. Reiss, and Vlaeri Mladenov. "Evaluation Of Distance Based Amplitde Panning For Spatial Audio". In: *ICASSP 2010 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2010.

[71] Mark R. P. Thomas and Charles Q. Robinson. "Amplitude Panning and the Interior Pan". In: *Audio Engineering Society Convention 143*. 2017.

[72] Mikko-Ville Laitinen et al. "Auditory Distance Rendering Using a Standard 5.1 Loudspeaker Layout". In: *Audio Engineering Society Convention 139*. 2015.

[73] P. Damaske. "Head-Related Two-Channel Stereophony with Loudspeaker Reproduction". In: *The Journal of the Acoustical Society of America* 50.4B (1971), pp. 1109–1115.

[74] M. R. Schroeder and B. S. Atal. "Computer simulation of sound transmission in rooms". In: *Proceedings of the IEEE* 51.3 (Mar. 1963), pp. 536–37.

[75] Marcos F. Simón Gálvez and Filippo Maria Fazi. "Loudspeaker Arrays For Transaural Reproduction". In: *The 22nd International Congress of Sound and Vibration, Florence*. 2015.

[76] Marcos F. Simón Gálvez and Filippo Maria Fazi. "Room Compensation for Binaural Reproduction with Loudspeaker Arrays". In: *Euro Regio 2016*. 2016.

[77] Jacob Hollebon, Filippo Maria Fazi, and Marcos F. Simón Gálvez. "A Multiple Listener Crosstalk Cancellation System Using Loudspeaker Dependent Regularization". In: *J. Audio Eng. Soc* 69.3 (2021), pp. 191–203.

[78] Jacob Hollebon, Marcos F. Simón Gálvez, and Filippo Maria Fazi. "Multiple Listener Crosstalk Cancellation Using Linear Loudspeaker Arrays For Binaural Cinematic Audio". In: *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. 2019.

[79] Marcos F. Simón Gálvez, Takashi Takeuchi, and Filippo Maria Fazi. "Low-Complexity, Listener's Position-Adaptive Binaural Reproduction Over a Loudspeaker Array". In: *Acta Acustica united with Acustica* 103.5 (2017).

[80] Marcos F. Simón Gálvez, Dylan Menzies, and Filippo Maria Fazi. "Dynamic Audio Reproduction with Linear Loudspeaker Arrays". In: *J. Audio Eng. Soc* 67.4 (2019), pp. 190–200.

[81] E. C. Hamdan and F. Maria Fazi. "Low Frequency Crosstalk Cancellation and Its Relationship to Amplitude Panning". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 566–570.

[82] Jacob Hollebon and Filippo Maria Fazi. "Generalised Low Frequency 3D Audio Reproduction Over Loudspeakers". In: *AES 148th Convention*. 2020.

[83] Henrik Møller. "Fundamentals of binaural technology". In: *Applied Acoustics* 36.3 (1992), pp. 171–218.

[84] B. Xie and J. Blauert. *Head-Related Transfer Function and Virtual Auditory Display: Second Edition*. A Title in J. Ross Publishing's Acoustics: Information and Communication Series. J. Ross Publishing, 2013.

[85] Henrik Møller et al. "Binaural Technique: Do We Need Individual Recordings?" In: *J. Audio Eng. Soc* 44.6 (1996), pp. 451–469.

[86] Elizabeth M. Wenzel et al. "Localization using nonindividualized head-related transfer functions". In: *The Journal of the Acoustical Society of America* 94.1 (1993), pp. 111–123.

[87] Ravish Mehra et al. "Comparison of localization performance with individualized and non-individualized head-related transfer functions for dynamic

listeners". In: *The Journal of the Acoustical Society of America* 140.4 (2016), pp. 2956–2957.

[88] Michele Geronazzo et al. "Improving Elevation Perception With A Tool For Image-Guided Head-Related Transfer Function Selection". In: *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17)*. 2017.

[89] Durand R. Begault et al. "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source". In: *Audio Engineering Society Convention 108*. 2000.

[90] Akio Honda et al. "Transfer effects on sound localization performances from playing a virtual three-dimensional auditory game". In: *Applied Acoustics* 68.8 (2007), pp. 885–896.

[91] David Poirier-Quinot and Brian F.G. Katz. "Assessing the Impact of Head-Related Transfer Function Individualization on Task Performance: Case of a Virtual Reality Shooter Game". In: *J. Audio Eng. Soc* 68.4 (2020), pp. 248–260.

[92] David Poirier-Quinot and Brian F. G. Katz. "Impact of HRTF Individualization on Player Performance in a VR Shooter Game I". In: *Audio Engineering Society Conference: 2018 AES International Conference on Spatial Reproduction - Aesthetics and Science*. 2018.

[93] David Poirier-Quinot and Brian F.G. Katz. "Impact of HRTF Individualization on Player Performance in a VR Shooter Game II". In: *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. 2018.

[94] Michele Geronazzo et al. "The Impact of an Accurate Vertical Localization with HRTFs on Short Explorations of Immersive Virtual Reality Scenarios". In: *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 2018, pp. 90–97.

[95] Lord Rayleigh. "XII. On our perception of sound direction". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 13.74 (1907), pp. 214–232.

[96] Jens Blauert. *Spatial Hearing*. Revised Edition. MIT Press, Cambridge, England, 1997.

[97] Frederic L. Wightman and Doris J. Kistler. "The dominant role of low-frequency interaural time differences in sound localization". In: *The Journal of the Acoustical Society of America* 91.3 (1992), pp. 1648–1661.

[98] G. Bruce Henning. "Detectability of interaural delay in high-frequency complex waveforms". In: *The Journal of the Acoustical Society of America* 55.1 (1974), pp. 84–90.

[99] Hans Wallach. "The Role of Head Movements and Vestibular and Visual Cues in Sound Localization." In: *Journal of Experimental Psychology* 27.4 (1940), p. 339.

[100] Frederic L. Wightman and Doris J. Kistler. "Resolution of Front-Back Ambiguity in Spatial Hearing by Listener and Source Movement". In: *The Journal of the Acoustical Society of America* 105.5 (1999), pp. 2841–2853.

[101] Douglas S. Brungart and William M. Rabinowitz. "Auditory localization of nearby sources. Head-related transfer functions". In: *The Journal of the Acoustical Society of America* 106.3 (1999), pp. 1465–1479.

[102] Enrique A. Lopez-Poveda and Ray Meddis. "A physical model of sound diffraction and reflections in the human concha". In: *The Journal of the Acoustical Society of America* 100.5 (1996), pp. 3248–3259.

[103] John Middlebrooks, James Makous, and David Green. "Directional sensitivity of sound-pressure levels in the human ear canal". In: *The Journal of the Acoustical Society of America* 86 (Aug. 1989), pp. 89–108.

[104] Jack Hebrank and D. Wright. "Spectral cues used in the localization of sound sources on the median plane". In: *The Journal of the Acoustical Society of America* 56.6 (1974), pp. 1829–1834.

[105] V. Ralph Algazi, Carlos Avendano, and Richard O. Duda. "Elevation localization and head-related transfer function analysis at low frequencies". In: *The Journal of the Acoustical Society of America* 109.3 (2001), pp. 1110–1122.

[106] Michael J. Evans, James A. S. Angus, and Anthony I. Tew. "Analyzing head-related transfer function measurements using surface spherical harmonics". In: *The Journal of the Acoustical Society of America* 104.4 (1998), pp. 2400–2411.

[107] Fabian Brinkmann and Stefan Weinzierl. "Comparison Of Head-Related Transfer Function Pre-Processing Techniques For Spherical Harmonics Decomposition". In: *AES International Confernce On Audio For Virtual And Augemented Reality*. 2018.

[108] Alonso Martinez J Engel, D Goodman, and L Picinali. "Assessing HRTF preprocessing methods for Ambisonics rendering through perceptual models". In: *Acta Acustica -Peking-* 6 (2022). DOI: aacus/2021055.

[109] Tim Lübeck et al. "Perceptual Evaluation of Mitigation Approaches of Impairments due to Spatial Undersampling in Binaural Rendering of Spherical Microphone Array Data". In: *J. Audio Eng. Soc* 68.6 (2020), pp. 428–440.

[110] Christian Schörkhuber, Markus Zaunschirm, and Robert Holdrich. "Binaural Rendering of Ambisonic Signals via Magnitude Least Squares". In: *44th DAGA*. 2018.

[111] Thomas Deppisch, Hannes Helmholz, and Jens Ahrens. "End to end Magnitude Least Squares Binaural Rendering Of Spherical Microphone Array Signals". In: *Immersive and 3D Audio: from Architecture to Automotive (I3DA)*. 2021.

[112] P.M.C. Morse and K.U. Ingard. *Theoretical Acoustics*. International series in pure and applied physics. Princeton University Press, 1986.

[113] F. Zotter. "Analysis and Synthesis of Sound-Radiation with Spherical Arrays". PhD thesis. Institute of Electronic Music, Acoustics University of Music, and Performing Arts, Austria, 2009.

[114] G. Aubert. "An alternative to Wigner d-matrices for rotating real spherical harmonics". In: *AIP Advances* 3.6 (2013).

[115] E. G. Williams. *Sound Radiation and Nearfield Acoustical Holography*. 1st. London: Academic Press, 1999.

[116] R. Mehrem. "The plane wave expansion, infinite integrals and identities involving spherical Bessel functions". In: *Applied Mathematics and Computation* 217.12 (2011), pp. 5360 –5365.

[117] Jian-Ming Jin. *Theory and Computation of Electromagnetic Fields*. John Wiley and Sons, 2010.

[118] D. Colton and R. Kress. *Inverse Acoustic and Electromagnetic Scattering Theory*. Springer, 2013.

[119] Adi Ben-Israel and Thomas Greville. *Generalized Inverses Theory and Applications*. Springer, 2003.

[120] Franz Zotter, Hannes Pomberger, and Matthias Frank. "An Alternative Ambisonics Formulation: Modal Source Strength Matching and the Effect of Spatial Aliasing". In: *AES 126th Convention*. 2009.

[121] George B. Arfken and Hans J. Weber. *Mathetical Methods For Physicists*. Sixth Edition. Boston: Academic Press, 2005.

[122] Glenn Dickins and Rodney Kennedy. "Towards Optimal Soundfield Representation". In: *Audio Engineering Society Convention 106*. 1999.

[123] Glenn Dickins. "Soundfield Representation, Reconstruction and Perception". PhD thesis. Research School of Information Sciences and Engineering, The Australian National University, 2003.

[124] Stefan Bilbao, Jens Ahrens, and Brian Hamilton. "Incorporating source directivity in wave-based virtual acoustics: Time-domain models and fitting to measured data". In: *The Journal of the Acoustical Society of America* 146.4 (2019), pp. 2692–2703.

[125] S. Bilbao, A. Politis, and B. Hamilton. "Local Time-Domain Spherical Harmonic Spatial Encoding for Wave-Based Acoustic Simulation". In: *IEEE Signal Processing Letters* 26.4 (2019), pp. 617–621.

[126] Nail A. Gumerov and Ramani Duraiswami. *Fast Multipole Methods For The Helmholtz Equation In Three Dimensions*. Elsevier, 2005.

[127] Phillip Cotterell. "On the Theory of Second-Order Soundfield Microphone". PhD thesis. University of Reading, 2002.

[128] Mihailo Kolundzija, Martin Vetterli, and Christof Faller. "Spatio-Temporal Gradient Analysis of Differential Microphone Arrays". In: *Audio Engineering Society Convention 126*. 2009.

[129] Mihailo Kolundzija, Christof Faller, and Martin Vetterli. "Sound Field Recording by Measuring Gradients". In: *Audio Engineering Society Convention 128*. 2010.

[130] G. W. Elko. *Audio Signal Processing for Next-Generation Multimedia Communication Systems*. Ed. by Y. Huang and J. Benesty. 1st ed. Springer, 2004.

[131] G. W. Elko. *Acoustic Signal Processing for Telecommunication*. Ed. by S. L. Gay and J. Benesty. 1st ed. Springer, 2000.

[132] Benjamin A. Cray, Victor M. Evora, and Albert H. Nuttall. "Highly directional acoustic receivers". In: *The Journal of the Acoustical Society of America* 113.3 (2003), pp. 1526–1532.

[133] Richard O. Duda and William L. Martens. "Range dependence of the response of a spherical head model". In: *The Journal of the Acoustical Society of America* 104.5 (1998), pp. 3048–3058.

[134] Filippo Maria Fazi, M. Noisternig, and O. Warusfel. "Representation of Sound Fields for Audio Recording and Reproduction". In: *Acoustics 2012*.

[135] Jung-Woo Choi and Kim Yang-Hann. *Sound Visualization and Manipulation*. John Wiley and Sons, 2013.

[136] ITU-R BS775-1. *Multichannel Sterephonic Sound System with and without Accompanying Picture*. Tech. rep. International Telecommunications Union, 1994.

[137] JBL. *Cinema Sound System Manual*. Tech. rep. JBL Professional Audio, 2003.

[138] Jacob Hollebon and Filippo Maria Fazi. "Experimental Study of Various Methods for Low Frequency Spatial Audio Reproduction Over Loudspeakers". In: *I3DA: International Conference on Immersive and 3D Audio*. 2021.

[139] Angelo Farina. "Advancements in Impulse Response Measurements by Sine Sweeps". In: *Audio Engineering Society Convention 122*. 2007.

[140] Jens Meyer and Gary Elko. "Spherical Microphone Array For Spatial Sound Recording". In: *Audio Engineering Society Convention 115*. 2003.

[141] Jens Meyer and Gary Elko. "A Highly Scalable Spherical Microphone Array Based On An Orthonormal Decomposition Of The Soundfield". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 2. 2002.

[142] M. Karjalainen and T. Paatero. "Frequency-dependent signal windowing". In: *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*. 2001, pp. 35–38.

[143] Florian Denk, Birger Kollmeier, and Stephan D. Ewert. "Removing reflections in semianechoic impulse responses by frequency-dependent truncation". In: *J. Audio Eng. Soc* 66.3 (2018), pp. 146–153.

[144] Johannes M. Arend, Fabian Brinkmann, and Christoph Pörschmann. "Assessing Spherical Harmonics Interpolation of Time-Aligned Head-Related Transfer Functions". In: *J. Audio Eng. Soc* 69.1/2 (2021), pp. 104–117.

[145] Benjamin Bernfeld. "Attempts for Better Understanding of the Directional Stereophonic Listening Mechanism". In: *Audio Engineering Society Convention 44*. Mar. 1973.

[146] ITU-R BS.1534-2. *Method For The Subjective Assessment Of Intermediate Quality Level Of Audio Systems*. Standard. ITU-R, 2014.

[147] Andreas Franck et al. "An Open Realtime Binaural Synthesis Toolkit for Audio Research". In: *Audio Engineering Society Convention 144*. 2018.

[148] Michio Woirgardt et al. *Cologne University of Applied Sciences*. Tech. rep. Anechoic Recordings, 2012.

[149] Jens Ahrens and Carl Andersson. "Perceptual Evaluation Of Headphone Auralization Of Rooms Captured With Spherical Microphone Arrays With Respect To Spaciousness And Timbre". In: *The Journal of the Acoustical Society of America* 145.4 (2019), pp. 2783–2794.

[150] B. Rafaely. *Fundamentals of Spherical Array Processing*. 1st. Vol. 8. Springer-Verlag Berlin Heidelberg, 2015.

[151] Filippo Maria Fazi. "Sound Field Reproduction". PhD thesis. University Of Southampton, 2010.

[152] Matthias Kronlachner and Franz Zotter. "Spatial Transformations For The Enhancement Of Ambisonic Recordings". In: *2nd International Conference On Spatial Audio (ICSA)*. 2014.

[153] Hannes Pomberger and Franz Zotter. "Warping Of 3D Ambisonic Recordings". In: *Ambisonics Symposium*. 2011.

[154] George F. Kuhn. "Model for the interaural time differences in the azimuthal plane". In: *The Journal of the Acoustical Society of America* 62.1 (1977), pp. 157–167.

[155] Richard O. Duda. "Modeling Head Related Transfer Functions". In: *The 27th Asilomar Conference On Signals, Systems and Computers*. 1993.

[156] Michele Geronazzo, Simone Spagnol, and Federico Avanzini. "Mixed structural modeling of head-related transfer functions for customized binaural audio delivery". In: *2013 18th International Conference on Digital Signal Processing (DSP)*. 2013, pp. 1–8.

[157] Jacob Hollebon, Eric Hamdan, and Filippo Maria Fazi. "A Comparison Of The Performance Of HRTF Models In Inverse Filter Design For Crosstalk Cancellation". In: *Proceedings of the Institute of Acoustics: Reproduced Sound*. Vol. 41. 3. 2019.

[158] Johannes Zaar. "Phase Unwrapping For Spherical Interpolation Of Head-Related Transfer Functions". PhD thesis. Institute of Electronic Music, Acoustics University of Music, and Performing Arts, Graz, 2011.

[159] Benjamin Bernschütz. "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100". In: *AIA-DAGA Conference on Acoustics*. 2013.

[160] Bruno Masiero and Janina Fels. "Perceptually Robust Headphone Equalization for Binaural Reproduction". In: *Audio Engineering Society Convention 130*. 2011.

[161] Bruno Masiero, Janina Fels, and Michael Vorlander. "Equalization For Binaural Synthesis With Headphones". In: *DAGA*. 2011.

[162] Etienne Hendrickx et al. "Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis". In: *The Journal of the Acoustical Society of America* 141.3 (2017), pp. 2011–2023.