ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/278404811

Binaural Reproduction of Higher Order Ambisonics A Real-Time Implementation and Perceptual Improvements

Thesis · July 2014

DOI: 10.13140/RG.2.1.4624.4007

CITATION
1

READS
5,691

1 author:

Jakob Vennerød
SINTEF

4 PUBLICATIONS **9** CITATIONS

SEE PROFILE

All content following this page was uploaded by Jakob Vennerød on 16 June 2015.

The user has requested enhancement of the downloaded file.

NTNU – Trondheim
Norwegian University of
Science and Technology

Binaural Reproduction of Higher Order Ambisonics

A Real-Time Implementation and Perceptual Improvements

Jakob Vennerød

Electronics System Design and Innovation
Submission date: June 2014
Supervisor: Peter Svensson, IET

Norwegian University of Science and Technology
Department of Electronics and Telecommunications

Problem description

A spherical microphone array is a scalable array with a certain number of microphones, where the number of microphones determines how directive the array can be implemented. The microphone signals can be processed by means of spherical harmonic (SH) modal functions, e.g. in the Higher Order Ambisonics format. This format is easily scalable between different array sizes, and sound fields can easy be rotated in this format. One can also convert between Higher Order Ambisonics and the binaural format for headphone reproduction.

In this project, the student should study how signals from a spherical microphone array, in the SH format, can be used to create rotatable sound fields that can be reproduced through headphones, with a head-tracking device. A real-time system shall, if possible, be implemented.

Preface

This thesis marks the end of six years of studying in Trondheim, and my completion of the MSc. degree in Electronics at the Norwegian University of Science and Technology (NTNU). The thesis work was done from January to June, 2014.

Initially, the topic of this thesis developed from a wish to study and work with microphone arrays, which is a rather popular, but also broad, field of acoustics today. I soon discovered that spatial audio and binaural sound would be the main ingredients of the thesis, which was not my primary field of interest prior to this work. However, this gave me the opportunity to learn a lot and it actually brought me back to the reason for why I started studying acoustics – music. Spatial audio opens up new dimensions (actually, two!) in music, compared to the stereo sound that we are so used to.

I am ever so grateful for having the opportunity to study and live here, at a great university and with so many wonderful people. I have learned that studying is not just to study. Studying is also to make new friends and colleagues, develop as a person and prepare for the world that comes after. I feel well prepared. Thus, there are many people I would like to thank.

First of all, thanks to my supervisor, Prof. Peter Svensson, for guiding me, inspiring me, generating new ideas and thoughts all the time, and counselling me on to my further career. Also, thanks to Audun Solvang for valuable discussions throughout the thesis work, and for the ideas that founded a lot of the work.

Thanks to all my friends that have supported me through these years, especially those at the Student Society in Trondheim and my classmates at NTNU. And finally, thanks to my family who always supports me.

Jakob Vennerød
Trondheim, June 2014

Abstract

During the last decade, Higher Order Ambisonics has become a popular way of capturing and reproducing sound fields. It can be combined with the theory of spherical microphone arrays to record sound fields, and this three-dimensional audio format can be reproduced with loudspeakers or headphones and even rotated around the listener. A drawback is that near perfect reproduction is only possible inside a sphere of radius r given by $kr < N$, where N is the Ambisonics order and k is the wavenumber.

In this thesis, the theory of spherical harmonics and Higher Order Ambisonics has been reviewed and expanded, which serves as a foundation for a real-time system that was implemented. This system can record signals from a commercial spherical microphone array, convert them to the Higher Order Ambisonics format, and reproduce the sound field through headphones. To compensate for head motion, a head-tracking device is used. The real-time system operates with a latency of around 95 milliseconds between head motion and consequent sound field rotation.

Further, two new methods for improving the headphone reproduction were assessed. These methods do not need to be applied in real-time, so no further system resources are used. Simulations of headphone reproduction with Higher Order Ambisonics show that both methods yield quantitative improvements in binaural cues such as the Interaural Level Difference, spectral cues and spectral coloration of the sound field. Median error values are reduced as much as 50 % between 4 and 7 kHz.

The findings indicate that Higher Order Ambisonics reproduction over headphones can be improved at frequencies above limit frequency given by $kr < N$, but these findings need to be confirmed by subjective assessments, such as listening tests. The work conducted in this thesis has also resulted in a comprehensive basis for further development of a real-time three-dimensional audio reproduction system.

Sammendrag

I løpet av det siste tiåret har Higher Order Ambisonics blitt en populær metode for å gjøre opptak av lydfelt og gjenskape det for lytting. Metoden kan kombineres med teori om kulemikrofoner for å gjøre opptak, og det tre-dimensjonale lydformatet gjenskapes ved hjelp av høyttalere eller hodetelefoner. Det kan også roteres rundt lytteren. En ulempe er at tilnærmet perfekt reproduksjon kan kun gjøres innenfor en sfære med radius r gitt av $kr < N$, hvor N er Ambisonics-ordenen og k er bølgetallet.

I denne masteroppgaven har teorien bak sfæriske harmoniske funksjoner og Higher Order Ambisonics blitt gjennomgått og utvidet, og dette legger et grunnlag for et sanntidssystem som ble implementert. Dette systemet kan ta opp lydsignaler fra en kommersielt tilgjengelig kulemikrofon, konvertere de til Higher Order Ambisonics-formatet, og spille av lydfeltet via hodetelefoner. En head-tracker ble brukt for å kompensere for hodebevegelser. Sanntidssystemet fungerer med en forsinkelse på rundt 95 millisekunder mellom hodebevegelser og påfølgende rotasjon av lydfeltet.

Videre har to nye metoder for å forbedre gjengivelsen med hodetelefoner blitt undersøkt. Disse metodene trenger ikke å kjøres i sanntid, så det er ikke behov for mer systemressurser. Simuleringer av Higher Order Ambisonics-reproduksjon med hodetelefoner viser at begge metodene gir kvantitative forbedringer i binaurale egenskaper slik som nivåforskjeller mellom ørene, spektrale mønstre og spektral farging av lydfeltet. Medianverdien av feilen ble redusert med opptil 50 % mellom 4 og 7 kHz.

Funnene indikerer at Higher Order Ambisonics-reproduksjon med hodetelefoner kan forbedres for frekvenser høyere enn grensefrekvensen gitt av $kr < N$, men disse funnene må bekreftes av subjektive eksperimenter, for eksempel lytteforsøk. Arbeidet som har blitt gjort i denne masteroppgaven har også resultert i et solid grunnlag for videreutvikling av et sanntids tredimensjonalt lydgjengivelsessystem.

Contents

List of Figures

List of Tables

List of Acronyms

API	Application Programming Interface
CDF	Cumulative Distribution Function
CPU	Central Processing Unit
DSP	Digital Signal Processor
EQ	Equalizer
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GCD	Great Circle Distance
HID	Human Interface Device
HOA	Higher Order Ambisonics
HRIR	Head-Related Impulse Response
HRTF	Head-Related Transfer Function
IFFT	Inverse Fast Fourier Transform
ILD	Interaural Level Difference
ITD	Interaural Time Difference
MPA	Median Plane Angle
NFC	Near Field Compensation
ROV	Region Of Validity
WFS	Wave Field Synthesis

CHAPTER 1 Introduction

3D sound is a continuously evolving research field in acoustics. During the past hundred years or so, scientists have studied how humans perceive sound sources in space with various experiments. Since the birth of modern stereophonic sound in the 1930s and its introduction on LP records in the late 1950s, 3D audio reproduction has evolved to comprise tenths of loudspeakers in today's movie theatres. On the contrary, personal sound systems have not had the same evolution, and still most sound reproduction is done with stereo sound or simple surround systems.

Why do we want 3D sound? Many people cannot really relate to this term, because terms like *stereo* and *surround* is used more frequently to describe spatial audio. So, 3D sound is merely a collection of reproduction techniques, which describes the ability for a sound reproduction system to render audio sources with spatial content. Countless experiments have shown that our sensory experiences are well affected by the spatial content of the sound. It catches our attention and helps us to distinguish between sources. Thus, when pursuing a realistic and sensuous experience, spatial audio is an important ingredient. And, recent development in 3D Virtual Reality (VR) visualisation addresses the need for improved 3D audio.

In addition to VR, other applications of spatial audio include movie theatres, spatial music listening, teleconferencing, and even communication in noisy, hazardous environments. The latter two relies on the fact that our ability to understand speech is improved by spatial separation of the speech and other noise sources.

When it comes to full-3D audio reproduction, three primary methods have been established in the last decades, namely Wave Field Synthesis (WFS) [1], Vector Based Amplitude Panning (VBAP) [2] and Higher Order Ambisonics (HOA) [3]. WFS relies on Huygens' principle that a wave front can be approximated with a distribution of secondary sources. VBAP is a generalisation of stereo panning methods, with a two- or three-dimensional loudspeaker setup. Ambisonics, introduced in its simplest form by Michael Gerzon in the 1970s, decomposes the sound field into a modal structure and seeks to reproduce these modes at the listener's position, called the ”sweet spot”. Ambisonics has further been developed to include higher order modes, which is HOA, based on a spherical harmonics decomposition of the sound field. Ideally, an infinite order is needed to perfectly reconstruct a sound field, but in practice, as in many cases, a finite order must be used. This is due to a restricted number of loudspeakers.

Both WFS and HOA, along with alternative methods, suffer from the fact that a large number of loudspeakers are required to provide high spatial resolution. This drawback can be overcome by simulating the loudspeaker signals with headphones, but that introduces further complications as will be discussed in this thesis.

1.1 Motivation

With HOA, one can obtain near perfect sound field reconstruction inside a sphere limited by the radius $r = N/k$, where N is the Ambisonics order and k is the wavenumber. At least $(N+1)^2$ loudspeakers are required for this. The quadratic relation between the radius and loudspeaker count is costly, especially if more than one listener is present. As an example, for decent reproduction below 5 kHz, at least 100 loudspeakers are required for one listener. In addition, this only applies for the ”sweet spot” in the centre of the array.

A possible solution suggested by SINTEF ICT[1] is to convert the HOA signals to the binaural format, so the sound reproduction can be done through headphones or earplugs. Head-Related Transfer Functions (HRTFs) can be constructed or measured to relate a sound source with the sound pressure at the ears, thus facilitating binaural synthesis, or *auralization*[2], of spatial sound with headphones.

This combination of HOA and auralization would provide a very convenient 3D audio format. The strength of HOA is that it is independent of the loudspeaker geometry and source positions, so it is very flexible. In addition, it is scalable in terms of transmission or storage. Auralization is cheap in terms of hardware and good quality can be obtained if the HRTFs are a good match to the listener.

So far, relatively few studies have been performed on how HOA can be combined with auralization. Though the main concepts have been previously described, few have studied how the objective and perceived quality is, and very few suggestions for improving such a system has been found in the literature.

In addition, the ongoing research on HOA in multiple scientific communities indicates that there is still need for improvement and evaluation of the method.

1.2 Previous work

While original 4-channel Ambisonics dates back to the 1970s [5, 6], HOA was further developed by Jérôme Daniel et al. [7] in the late 1990s and early 2000s, based on spherical harmonics. He and his colleagues proposed alternative decoding methods [3], expanded the theory with spherical microphone array processing [8–11] and introduced near-field coding [12, 13]. Their work constitutes a significant part of the knowledge of HOA today.

Simultaneously, loudspeaker reproduction of sound fields was studied by Ward and Abhayapala [14], by using an array of loudspeakers to construct a plane wave source. 3D loudspeaker reproduction techniques based on HOA (and WFS) has further been

[1] www.sintef.no

[2] Auralization was introduced by Kleiner et al. [4], and aims to *”recreate the aural impression of the acoustic characteristics of a space”*.

studied in detail by Ahrens and Spors [15–17], often with a more theoretical approach.

Spherical microphone array processing became popular following Meyer and Elko's paper [18], which resulted in the commercial Eigenmike® microphone. Rafaely et al. have studied such arrays in depth [19–25], in particular with respect to beamforming applications and error analysis. Advanced beamforming methods were further developed by Sun [26], describing the ability to construct more advanced beamforming patterns. Duraiswami et al. has also studied such arrays with attention to microphone positioning and array robustness [27–30], but also different shapes such as rigid hemispherical microphone arrays [31].

For readers with little experience on the subject, a good review on 3D sound field recording and reproduction was written by Poletti [32]. Modal array processing is thoroughly covered in the book by Teutsch [33].

Previous work on binaural reproduction with HOA is highly relevant for this study. Landone and Sandler [34] first introduced binaural processing of Ambisonic sound fields, for the purpose of evaluating multi-channel systems. However, the decoding of Ambisonic signals to virtual loudspeakers convolved with HRTFs was first introduced by Noistering et al. [35], further expanded with a spherical microphone array by Duraiswami et al. [36, 37], claiming that the results sounded convincing. Menzies and Ai-Akaidi [38] showed that binaural rendering of near field sources might suffer from errors due to scattering from objects (i.e. the listener's body) outside the region of validity in HOA. Further developments included conversion of HRTFs to the spherical harmonics domain (Duraiswami et al., [39], Pollow et al. [40]), which simplifies binaural HOA rendering and facilitates calculation of HRTFs at arbitrary points in space.

Evaluation of HOA reproduction has mainly been focused on loudspeaker systems. Several approaches have been taken, for example localisation accuracy [41, 42] or spectral impairment [43]. Recently, binaural HOA has also been evaluated, mainly in terms of localisation [44,45]. Shabtai and Rafaely [46] also studied speech intelligibility with a binaural beamforming method based on a spherical microphone array.

Few studies have been conducted seeking to evaluate binaural cues such as the Interaural Time/Level Difference (ITD/ILD). Epain et al. [47] evaluated a 32-loudspeaker array with an acoustic manikin, focusing on ILD and ITD, also with out-of-centre head locations. They found that the broadband ITD was well preserved below 2 kHz, while the ILD and spectral cues had significant errors leading to worse localisation. Very recently, Clapp et al. [48] studied binaural cues, to evaluate a spherical microphone array perceptually. Bertet et al. [42] also touched this topic, comparing various HOA systems with a combined ILD and ITD localisation model.

Most recently, binaural HOA reproduction has been evaluated in terms of more subjective measures with listening tests. Sheaffer, Rafaely and Villeval [49] studied *externalisation, localisation blur* and *timbre*, as well as suggesting a timbre correction filter to compensate for the high frequency loss resulting from a finite order N. Preliminary results show that this compensation yields a better subjective experience.

1.3 Problem and purpose

To evaluate a binaural HOA reproduction system, it is essential to incorporate the listener's head movements into the auralization process. Thus, a real-time system with a head-tracking device is needed. The problem formulation is then divided in two parts:

- A real-time system for binaural reproduction of HOA is needed to do subjective tests. Thus, such a system shall be implemented. A spherical microphone array will be used to obtain the HOA signals in a real sound field. The system must be able to compensate for head movements with a head-tracker.

- The quality of reproduction will be evaluated with focus on the binaural cues that constitute spatial hearing. This is done by evaluating objective measures such as interaural differences and the spectral behaviour of the sound field. In addition, possible improvements of the reproduction technique are investigated.

A purpose that arose throughout the thesis work was to provide a consistent theoretical framework for the HOA method and the addition of auralization. Many previous studies are rather limited on this end, partly because different authors use different conventions, and often only specific parts of the theory are addressed. The purpose is not to do a complete theoretical review of HOA, but to provide the necessary theory and tools to implement and analyse the system in question. Further work with the binaural HOA project will also benefit from this documentation.

1.4 Contribution

Though the main objectives of this project are to implement a binaural HOA rendering system and evaluate its performance, the work contains some new theory, insight and results.

Firstly, the study connects HOA sound field capture to binaural reproduction both in terms of implementation, but also in terms of evaluating the performance. To the author's knowledge, the full chain from a spherical microphone array to a binaural reproduction system has not previously been thoroughly covered. Thus, limitations on both the recording and reproduction side can be discussed and compared.

In addition, further insight has been gained on the effects of truncation error. A new formula for the normalised truncation error on a rigid sphere has been derived. The truncation error affects binaural cues, which has received little attention in the literature. New insights on how these errors affect localisation and audio quality has been obtained with objective, numerical measures.

Finally, two new methods for improving the binaural reproduction have been developed. By manipulating the phase response of the HRTF database, high frequency reproduction is improved only at the cost of phase response. However, the current established psychoacoustic models allow a degraded phase response at high frequencies since the localisation is mainly influenced by level differences in this frequency range.

1.5 Outline

Chapter 2 covers the theoretical framework behind HOA, starting with spherical acoustics. Then, the HOA encoding process is covered, including spherical microphone arrays and virtual sources. Sound field rotation is briefly covered. Further, methods for decoding the HOA signals to loudspeakers and headphones are presented, and the main error sources are addressed. Finally, the basics of binaural hearing are revisited, and new correction methods for improving the binaural cues are proposed.

Chapter 3 goes through the real-time implementation, including hardware and software choices, and a detailed description of the system. Processing algorithms for the encoding, motion handling and binaural synthesis are discussed. A brief overview of system requirements and performance is presented, along with a few key numbers on resource use and latency.

Chapter 4 presents the numerical results from various simulations of the HOA system. This includes spherical microphone analysis in terms of aliasing and noise, truncation error and binaural representation error. The ILD, ITD and spectral cues are investigated, and the proposed correction methods are evaluated.

Chapter 5 discusses the implementation, quantitative results and implications. Some remarks are made with respect to further development and evaluation of the system.

Chapter 6 sums up the main findings and implications.

Appendix A contains a derivation of the normalised truncation error on a rigid sphere.

Appendix B contains an overview of the MATLAB code, and the most important scripts and functions that were developed.

Attachments (`zip`-file) contain all the MATLAB scripts and necessary data files for the real-time system and analysis. Psychophysics Toolbox has to be downloaded separately. Sample audio files were not included due to file size limitations.

CHAPTER 2 Theoretical framework

This chapter seeks to give the reader a thorough introduction to Higher Order Ambisonics (HOA). The basics of spherical acoustics are revisited, and it is shown how this is used to represent 3D sound with HOA. Encoding is covered, both from virtual sources and spherical microphone arrays. It is shown how the HOA signals (or *channels*) can be decoded with a loudspeaker array or binaurally with headphones. In addition, some basic theory of binaural hearing is presented. Finally, a novel method for improving the binaural HOA reproduction at high frequencies is introduced, taking into account psychoacoustic features of binaural listening.

There are two main goals with this chapter. The first one is to provide a framework for the reader that has general knowledge of acoustics, but limited knowledge of HOA. The second is to provide a complete theoretical description of a HOA system that can be used in future research and development.

2.1 Spherical acoustics

Spherical acoustics describes the treatment of acoustics in spherical geometries. The wave equation can be solved in the Cartesian, cylindrical or spherical coordinate system. With spherical geometries, the latter is an obvious choice, and a sound field can be described in a very elegant way with *Spherical Harmonics*. In the following, decomposition of sound fields in the spherical coordinate system is considered.

For a more thorough description of spherical acoustics, the reader is referred to Williams [50].

2.1.1 Spherical coordinates

A point (x, y, z) in spherical coordinates can be described as a vector with length r, elevation θ and azimuth ϕ. Figure 2.1 shows the classical definition[1].

The relation between the Cartesian coordinate system and the classical spherical

[1] Other definitions include swapping θ and ϕ, or defining θ as the angle from the xy-plane.

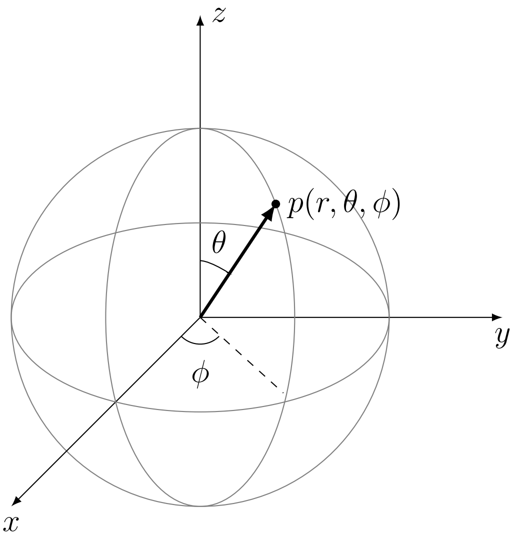


Figure 2.1: Definition of the spherical coordinate system used in this study.

coordinate system is

$$\begin{aligned} x &= r\cos\phi\sin\theta \\ y &= r\sin\phi\sin\theta \\ z &= r\cos\theta \end{aligned} \tag{2.1}$$

so a coordinate transformation can easily be done.

2.1.2 Spherical harmonics

The sound field will be decomposed into frequency, radial and angular functions. Spherical harmonics constitute the angular functions. Any arbitrary, square integrable function on a sphere can be described as

$$f(\theta,\phi) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} a_n^m Y_n^m(\theta,\phi) \tag{2.2}$$

where a_n^m are complex coefficients and Y_n^m are the *complex* spherical harmonics defined as:

$$Y_n^m(\theta,\phi) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}}P_n^m(\cos\theta)e^{\mathrm{i}m\phi} \tag{2.3}$$

Here, $P_n^m(x)$ are the associated Legendre functions [50, pp. 187]. If the function $f(\theta,\phi)$ is known, the complex coefficients can be found by

$$a_n^m = \int_0^{2\pi}\int_0^{\pi} Y_n^m(\theta,\phi)^* f(\theta,\phi)\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\phi \tag{2.4}$$

Different definitions of the spherical harmonics exist, and the real-valued spherical harmonics [8] are of particular interest:

$$\Upsilon_n^m(\theta,\phi) = \sqrt{(2n+1)(2-\delta_{0m})\frac{(n-m)!}{(n+m)!}}P_n^m(\cos\theta) \times \begin{cases} \cos(m\phi) & m>0 \\ 1 & m=0 \\ \sin(m\phi) & m<0 \end{cases} \tag{2.5}$$

δ_{nm} is the Kronecker delta function. The real-valued definition is practical when dealing with real-valued audio signals. For reference, the relation between the complex definition in Equation (2.3) and Daniel's real definition in Equation (2.5) is

$$\Upsilon_n^m = \begin{cases} \sqrt{2\pi}(Y_n^m + Y_n^{m*}) & m > 0 \\ \sqrt{4\pi} Y_n^m & m = 0 \\ -\mathrm{i}\sqrt{2\pi}(Y_n^{|m|} - Y_n^{|m|*}) & m < 0 \end{cases} \tag{2.6}$$

Two important properties of the spherical harmonics are the orthonormality property

$$\int_0^{2\pi}\int_0^{\pi} Y_n^m(\theta,\phi) Y_{n'}^{m'}(\theta,\phi)\sin\theta \,\mathrm{d}\theta\,\mathrm{d}\phi = \delta_{nn'}\delta_{mm'} \tag{2.7}$$

which means that they form a complete set, and the addition theorem [14], which states that:

$$\sum_{m=-n}^{n} Y_n^m(\theta_1,\phi_1) Y_n^m(\theta_2,\phi_2)^* = \frac{2n+1}{4\pi} P_n^0(\cos\Omega) \tag{2.8}$$

where Ω is the central angle between (θ_1, ϕ_1) and (θ_2, ϕ_2). This is particularly interesting in the axisymmetric case (an acoustic wave arriving along the z-axis), where Ω reduces to θ.

2.1.3 Directivity patterns of the spherical harmonics

To understand how the spherical harmonics work and how the sound field can be represented in the spherical harmonics domain, one can look at the directivity plots of the first few spherical harmonics. Figure 2.2 show the magnitude of the real spherical harmonics up to order 4. The directivity plots can be interpreted as a monopole (omnidirectional) element (Y_0^0), dipole elements (Y_1^m), quadrupole elements (Y_2^m) and so on. Thus, the spherical harmonics components of a wave field around the origin can be interpreted as signals corresponding to an infinite number of microphones with different directivities.

2.1.4 Solution of the wave equation in spherical coordinates

The linear, homogeneous wave equation in spherical coordinates is given by

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial p}{\partial r}\right) + \frac{1}{r^2\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial p}{\partial\theta}\right) + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 p}{\partial\phi^2} - \frac{1}{c^2}\frac{\partial^2 p}{\partial t^2} = 0 \tag{2.9}$$

where c is the speed of sound. A solution can be found by separating the variables such that:

$$p(r,\theta,\phi,t) = R(r)\Theta(\theta)\Phi(\phi)T(t) \tag{2.10}$$

A complete derivation of how to solve the wave equation in spherical coordinates is not presented here but the result is given. Any solution can be written as [50, pp. 186]

$$p(r,\theta,\phi,\omega) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}(A_n^m j_n(kr) + B_n^m y_n(kr)) Y_n^m(\theta,\phi) e^{-\mathrm{i}\omega t} \tag{2.11a}$$

$$p(r,\theta,\phi,\omega) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n}(C_n^m h_n^{(1)}(kr) + D_n^m h_n^{(2)}(kr)) Y_n^m(\theta,\phi) e^{-\mathrm{i}\omega t} \tag{2.11b}$$

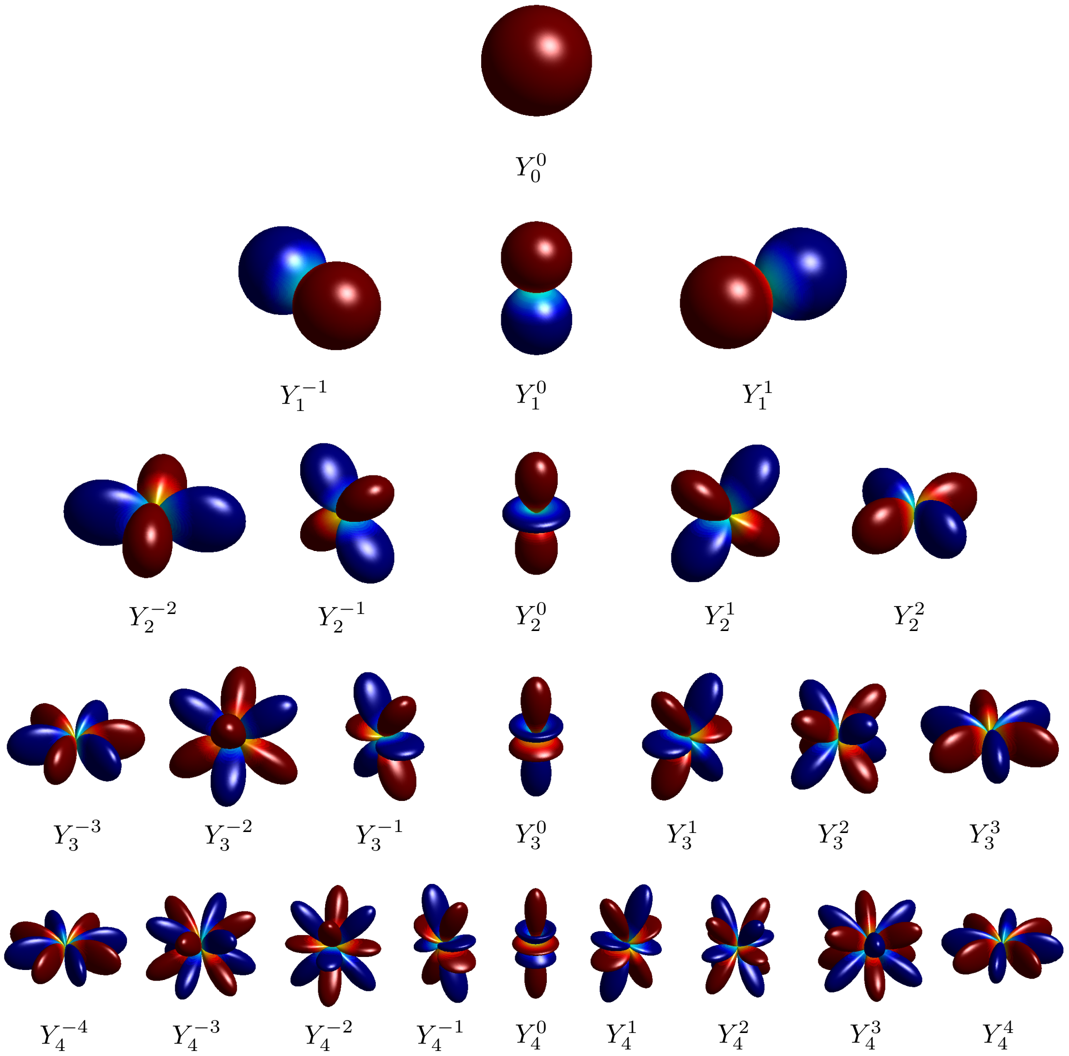

Y_0^0
Y_1^{-1} Y_1^0 Y_1^1
Y_2^{-2} Y_2^{-1} Y_2^0 Y_2^1 Y_2^2
Y_3^{-3} Y_3^{-2} Y_3^{-1} Y_3^0 Y_3^1 Y_3^2 Y_3^3
Y_4^{-4} Y_4^{-3} Y_4^{-2} Y_4^{-1} Y_4^0 Y_4^1 Y_4^2 Y_4^3 Y_4^4

Figure 2.2: Directivity patterns of the real spherical harmonics up to order 4. Red colours are positive amplitudes and blue colours are negative amplitudes.

for standing wave and traveling wave solutions, respectively. In the following, the time-dependent term $e^{-\mathrm{i}\omega t}$ will be omitted for simplicity. The coefficients A_n^m, B_n^m, C_n^m and D_n^m are generally frequency-dependent functions and can be regarded as a kind of spatial Fourier Transform coefficients in the spherical domain. j_n, y_n, $h_n^{(1)}$ and $h_n^{(2)}$ are radial functions, which represent the radial dependency of the wave field. The spherical Hankel functions of the first and second kind are defined as

$$h_n^{(1)}(x) = \sqrt{\frac{\pi}{2x}}\left(J_{n+1/2}(x) + \mathrm{i}Y_{n+1/2}(x)\right) \propto e^{\mathrm{i}x} \tag{2.12a}$$

$$h_n^{(2)}(x) = \sqrt{\frac{\pi}{2x}}\left(J_{n+1/2}(x) - \mathrm{i}Y_{n+1/2}(x)\right) \propto e^{-\mathrm{i}x} \tag{2.12b}$$

that represent outgoing and incoming waves, respectively. Whether we keep one or both of these terms depends on the specific acoustic problem. J_n and Y_n are the

Bessel functions of the first and second kind. The spherical Bessel functions

$$j_n(x) = \sqrt{\frac{\pi}{2x}} J_{n+1/2}(x) \tag{2.13a}$$

$$y_n(x) = \sqrt{\frac{\pi}{2x}} Y_{n+1/2}(x) \tag{2.13b}$$

represent the standing wave solutions, but the first kind is particularly interesting: A general solution for an interior problem (all sources are placed outside a sphere region of validity, see Figure 2.10) can be written as:

$$p(r, \theta, \phi, \omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} A_n^m(\omega) j_n(kr) Y_n^m(\theta, \phi) \tag{2.14}$$

This expression is very useful when capturing a sound field in a limited area with a spherical microphone array, or reconstructing such a sound field with an array of loudspeakers.

2.1.5 Plane wave representation with spherical harmonics

Equation (2.14) must be able to express a unit magnitude plane wave traveling with the direction (θ_i, ϕ_i). Williams [50, pp. 227] gives the solution:

$$p(r, k, \theta, \phi) = 4\pi \sum_{n=0}^{\infty} \mathrm{i}^n j_n(kr) \sum_{m=-n}^{n} Y_n^m(\theta, \phi) Y_n^m(\theta_i, \phi_i)^* \tag{2.15}$$

For the simpler, axisymmetric case, where the wave arrives from the negative z-direction, the equation can be simplified with Equation (2.8) to form:

$$p(r, k, \theta, \phi) = \sum_{n=0}^{\infty} i^n (2n+1) j_n(kr) P_n^0(\cos\theta) \tag{2.16}$$

2.1.6 Scattering from a rigid sphere

A spherical microphone array can be constructed as an array of pressure sensors in free space, or as an array of sensors mounted at a rigid sphere [8–11, 18]. In addition, the introduction of a human listener in the sound field will scatter the sound field. In some cases, the head can be approximated by a rigid spherical scatterer. Teutsch [33, pp. 39] gives an expression for the total sound field for a plane wave traveling with the direction (θ_i, ϕ_i), scattered by a rigid sphere of radius R centred in the origin (see Figure 2.3):

$$p_{tot}(r, \theta, \phi, \omega) = 4\pi \sum_{n=0}^{\infty} \mathrm{i}^n \left[j_n(kr) - \frac{j_n'(kR) h_n(kr)}{h_n'(kR)} \right] \sum_{m=-n}^{n} Y_n^m(\theta, \phi), Y_n^m(\theta_i, \phi_i)^* \tag{2.17}$$

This equation is also derived in Appendix A. Here j_n' and h_n' are the first derivatives of the spherical Bessel and Hankel functions of the first kind. For simplicity, the Wronsikan expression can be used to obtain [32]

$$\left[j_n(kr) - \frac{j_n'(kR) h_n(kr)}{h_n'(kR)} \right] = \frac{\mathrm{i}}{(kR)^2 h_n'(kR)} \tag{2.18}$$

when $r = R$, i.e. a simpler expression for the pressure at the rigid sphere surface.

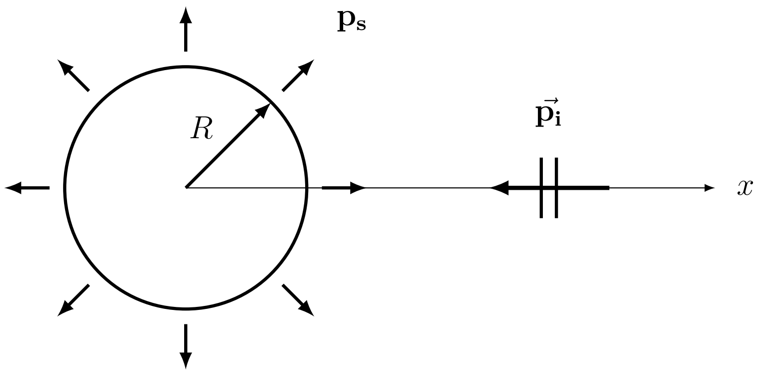


Figure 2.3: A plane wave $\vec{\mathbf{p_i}}$ arriving from the positive x-direction $(\theta_i, \phi_i) = (\pi/2, \pi)$, along with the scattered pressure $\mathbf{p_s}$ from a rigid sphere with radius R.

2.2 Higher Order Ambisonics

Ambisonics was introduced by Gerzon [5,6] in the 1970s. He experimented with placing four cardioid microphones in a tetrahedral formation, to produce a four-channel representation of the spatial audio, consisting of the monopole channel (W) and dipole element channels (X,Y,Z) in Figure 2.2. The WXYZ configuration, called B-format, is Ambisonics in its simplest form. This corresponds to truncating the series in Equation (2.14) to order $N = 1$, and thus a lot of spatial information is lost. Higher Order Ambisonics is the method of representing the sound field with a higher truncation order $N > 1$. This requires $(N + 1)^2$ HOA signals, as each order N contains $2N + 1$ signals, which can be represented in either the frequency- or time-domain, and, if needed, compressed to reduce spatial redundancy [51, 52].

There are two methods for constructing a HOA signal:

1. The sound field from a virtual source placed at a point in space (r, θ, ϕ) is encoded directly with spherical harmonics. Both plane-wave (far field) and spherical-wave (near field) sources can be encoded.

2. The sound field is recorded by a microphone array centred in the origin of the coordinate system. Typically, a spherical microphone array is used, with pressure sensors mounted on a spherical grid or on a rigid sphere. The array output is then encoded to the spherical harmonics domain.

The $(N + 1)^2$ HOA signals can then be transmitted or stored in a format which is very flexible. A sound field can be easily rotated around the origin [23], and scaled down to a lower order N'. This only reduces the spatial information at high frequencies, which may not be needed for some applications.

On the reproduction side, the sound field can be reproduced with a loudspeaker array distributed uniformly on a sphere centered at the listening position. This requires at least $(N+1)^2$ loudspeakers [8, 14]. Ideally, the loudspeakers should reproduce plane waves, but the fact that loudspeakers radiate spherical waves will cause a slight error. This can be corrected for using Daniel's Near Field Compensation (NFC) [12, 13].

HOA can either be formulated in 2D or 3D. The 2D formulation uses cylindrical coordinates [8], and comprises spatial information in the horizontal plane. Thus, spatial audio can be reproduced in the 360° horizontal plane, but with no elevation cues. In practice, 2D Ambisonics is realised with a circular loudspeaker array around the listener.

As the HOA representation order increases, the amount of spatial information increases. The reconstruction error arising from a finite number of loudspeakers (and thus HOA signal components) will increase with the distance from the origin. Near perfect reconstruction is only possible inside a sphere with radius $r \leq N/k$, as will be shown later in this chapter.

2.2.1 Encoding - virtual sources

Sound sources can be encoded in the HOA format when the source position and directivity is known. A plane wave arriving from a direction (θ_S, ϕ_S) can be represented with Equation (2.15). Thus, the complex coefficients in Equation (2.14) are:

$$A_n^m = 4\pi \mathrm{i}^n Y_n^m(\theta_S, \phi_S) \tag{2.19}$$

Daniel's HOA formulation [8] with *real* spherical harmonics uses the coefficient notation B_n^m instead of A_n^m (do not confuse with B_n^m in equation (2.11a)). From here, this notation convention will be used. Thus, the encoding is simplified to

$$\mathbf{B} = S\mathbf{Y} \tag{2.20}$$

where S is the source signal, $\mathbf{Y}$ is the vector of *real* spherical harmonics Υ_n^m and $\mathbf{B}$ is the vector of *real* Ambisonics signals B_n^m. In practice, $\mathbf{B}$ is a matrix of $(N+1)^2$ digital signals of length L. According to Daniel's formulation, we can now describe the sound field with a truncated series

$$\tilde{p}(r,\theta,\phi,\omega) = \sum_{n=0}^{N} \sum_{m=-n}^{n} \mathrm{i}^n j_n(kr) B_n^m(\omega) \Upsilon_n^m(\theta,\phi) \tag{2.21}$$

which can be expressed similarly with complex spherical harmonics as in Equation (2.14). Using the complex definition will result in complex Ambisonic signals, but due to the relation in Equation (2.5) we can easily retrieve the real Ambisonic signals. Only the complex Ambisonic signals with $m \geq 0$ need to be transmitted.

To encode near-field sources, the sound field from a monopole must be expressed with a spherical harmonics expansion. Details are not shown here, but Daniel [12] provides a simple expression to include the distance information in the HOA signals. A source at a distance ρ can be encoded as:

$$B_n^m = S F_n(\rho,\omega) \Upsilon_n^m(\theta_S, \phi_S) \tag{2.22}$$

where the frequency-dependent functions F_n are defined as:

$$F_n(\rho,\omega) = \sum_{m=0}^{n} \frac{(n+m)!}{(n-m)!m!} \left(\frac{-\mathrm{i}c}{2\omega\rho}\right)^m \tag{2.23}$$

Note that the $1/\rho$ distance attenuation and air absorption must be modelled separately. These functions affect mainly the low frequencies and can be implemented as time- or frequency-domain filters. However, the filters cause excessive amplification at low frequencies which is impractical in a filter implementation. This is solved by applying a correction filter taking the loudspeaker distance on the reproduction side into account. Thus, a realisable filter (replacing F_n in Equation (2.22)) can be expressed as

$$H_n(\omega) = \frac{F_n(\rho, \omega)}{F_n(R, \omega)} \tag{2.24}$$

where R is the reproduction loudspeaker distance.

Figure 2.4 shows the correction filters for an example where the source is to be located 1m from the listener, and the loudspeakers are located 3m away. The main effects of the filters are low-frequency amplification or attenuation, for close sources and loudspeakers, respectively. However, it will be shown later that particularly the high-order Ambisonic signals will need to be high-passed anyway, minimising the result of near field effects.

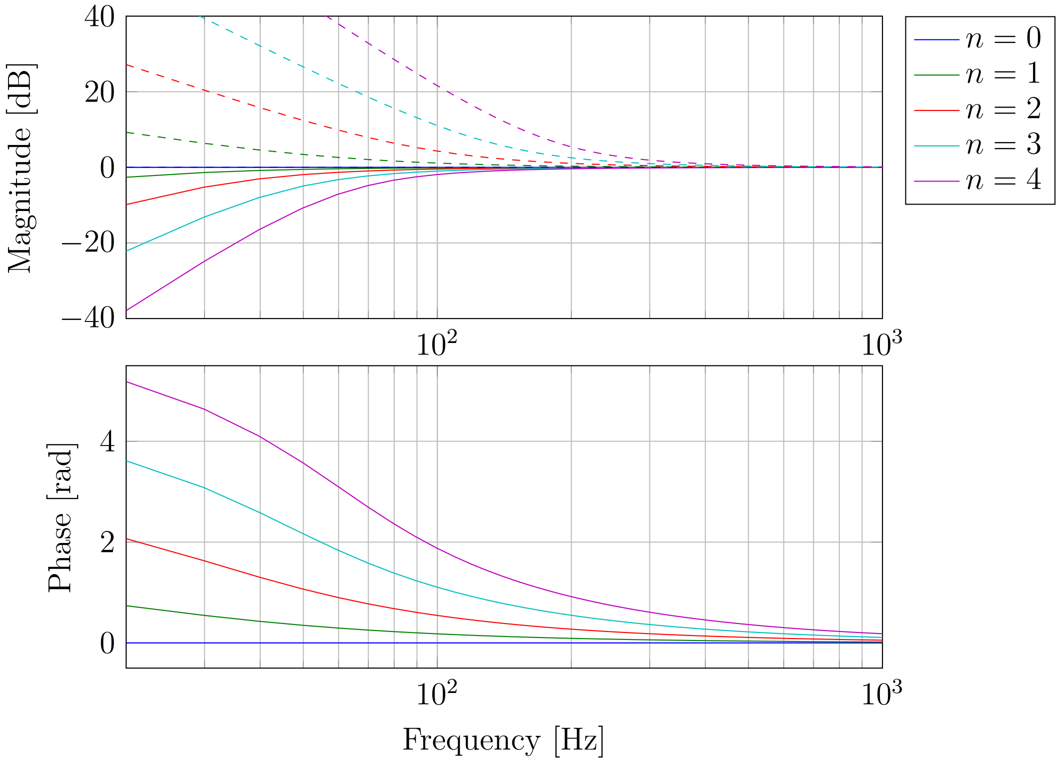


Figure 2.4: Magnitude (top) and phase (bottom) of the NFC filters. Loudspeaker filters $1/F_n(R, \omega)$ (solid lines), $R = 3$m, and source filters $F_n(\rho, \omega)$ (dashed lines), $\rho = 1$m, orders 0 through 4.

2.2.2 Encoding - sound field recording

The other way to encode a HOA signal is to capture a real sound field with a microphone array. Due to the spherical harmonics decomposition of the sound field, a spherical microphone array is the most convenient configuration. With an array that has $Q > (N + 1)^2$ sensors, with radius R, it is possible to estimate the Ambisonics signals up to order N. There are two ways to do this, either by direct integration (DI) or by the least-squares method [39]. With the DI method, one pre-defines a quadrature for the sphere and uses Equation (2.4) to estimate the coefficients such that

$$B_n^m W_n(kR) = \int_S p(R,\theta,\phi)\Upsilon_n^m(\theta,\phi)^* \,\mathrm{d}S \tag{2.25}$$

where $W_n(kR)$ is a radial function, depending on the array geometry. S is the sphere surface. The integral has to be approximated by numerical integration, which introduces approximation errors. With the least squares method discussed in the following, these errors are minimised.

The sound field can be sampled on a spherical surface, which can be regarded as a sampling of the left side of Equation (2.14). The pressure on a microphone q placed on a sphere can generally be described as [10]

$$p_q(\omega) = \sum_{n=0}^{N} \sum_{m=-n}^{n} W_n(kR) B_n^m(\omega) \Upsilon_n^m(\theta_q, \phi_q) \tag{2.26}$$

In matrix form we obtain:

$$\mathbf{p} = \mathbf{YWB} \tag{2.27}$$

where the sensor pressures are defined in the vector

$$\mathbf{p} = [p_1,\, p_2,\, p_3 \mathinner{..} p_Q]^T, \tag{2.28}$$

the spherical harmonics matrix,

$$\mathbf{Y} = \begin{bmatrix} \Upsilon_0^0(\theta_1,\phi_1) & \Upsilon_1^{-1}(\theta_1,\phi_1) & \Upsilon_1^0(\theta_1,\phi_1) & \Upsilon_1^1(\theta_1,\phi_1) & .. & \Upsilon_N^N(\theta_1,\phi_1) \\ \Upsilon_0^0(\theta_2,\phi_2) & \Upsilon_1^{-1}(\theta_2,\phi_2) & \Upsilon_1^0(\theta_2,\phi_2) & \Upsilon_1^1(\theta_2,\phi_2) & .. & \Upsilon_N^N(\theta_2,\phi_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \Upsilon_0^0(\theta_Q,\phi_Q) & \Upsilon_1^{-1}(\theta_Q,\phi_Q) & \Upsilon_1^0(\theta_Q,\phi_Q) & \Upsilon_1^1(\theta_Q,\phi_Q) & .. & \Upsilon_N^N(\theta_Q,\phi_Q) \end{bmatrix}, \tag{2.29}$$

and the radial function matrix is defined as a ”pseudo-diagonal” matrix:

$$\mathbf{W} = \mathrm{pdiag}[W_n(kR)] \equiv \begin{bmatrix} W_0(kR) & 0 & 0 & .. & 0 \\ 0 & W_1(kR) & 0 & .. & 0 \\ 0 & 0 & W_1(kR) & .. & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & W_N(kR) \end{bmatrix} \tag{2.30}$$

Note that each element W_n is repeated $2n+1$ times in the diagonal matrix. Also, note that the real spherical harmonics matrix is denoted $\mathbf{Y}$, but in principle this could be complex spherical harmonics as well, yielding complex Ambisonics signals. Finally, the Ambisonics signals matrix is defined as

$$\mathbf{B} = [B_0^0,\, B_1^{-1},\, B_1^0\, B_1^1 \mathinner{..} B_N^N]^T \tag{2.31}$$

Also, note that the order of spherical harmonics and Ambisonics coefficients runs from small to large values of m, from small to large order. This is a matter of convention, but as long as consistency is maintained, any convention could be used. Different authors may use different conventions, so this is important to keep in mind when working with Ambisonics signals.

We want to determine $\mathbf{B}$. Equation (2.26) implies that the sound field representation is truncated to order N. Thus the number of microphones must be larger than the number of coefficients in $\mathbf{B}$, i.e. $(N+1)^2$. If not, the system is underdetermined and has no unique solution. If $Q=(N+1)^2$, the system can be easily solved by multiplying both sides with the inverse of $\mathbf{WY}$. However, if $Q>(N+1)^2$, the system is overdetermined and the solution must be determined by a least squares solution [10], which results in

$$\tilde{\mathbf{B}} = \mathbf{W}^{-1}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{p} \tag{2.32}$$

The matrix $(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T$ is the Moorse-Penrose matrix pinv($\mathbf{Y}$), also called the pseudoinverse of $\mathbf{Y}$. Thus we obtain an estimation of the HOA signals with two simple operations: Multiplication of the microphone signals with an encoding matrix $\mathbf{E}=\text{pinv}(\mathbf{Y})$ and filtering the HOA components of order n with the inverse radial filter $1/W_n(kR)$:

$$\tilde{\mathbf{B}} = \text{pdiag}[1/W_n(kR)]\mathbf{Ep} \tag{2.33}$$

Note that if the loudspeakers are placed irregularly on the sphere, $\mathbf{Y}$ will be ill-conditioned and the solution is prone to numerical errors (see Section 3.6). The same will happen if one tries to estimate an encoding matrix where the number of microphones is smaller than $(N+1)^2$. Thus, for any new sampling scheme on the sphere, the condition number should be calculated to avoid ill-conditioning.

Figure 2.5 shows the encoding operation with a matrix mixer and inverse radial filters. The filters will depend on the array configuration. If the microphones are omnidirectional and placed in free space, the filters simply reduce to the spherical Bessel functions $i^n j_n(kr)$(Equation (2.13a)). However, these functions have nulls at certain frequencies, as seen in Figure 2.6. This poses problems when designing the inverse filters in Equation (2.33), as they will create infinite amplification at the nulls. This is not possible in practice and noise will also be a problem.

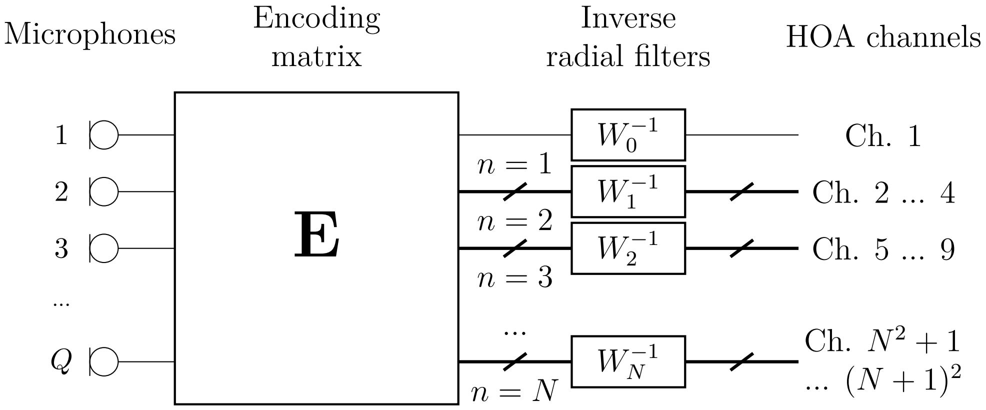


Figure 2.5: Encoding operation for a spherical microphone array with Q sensors, resulting in a HOA representation of order N and $(N+1)^2$ channels. Thick lines are groups of channels running through the same filters W_n^{-1} (one filter for each channel).

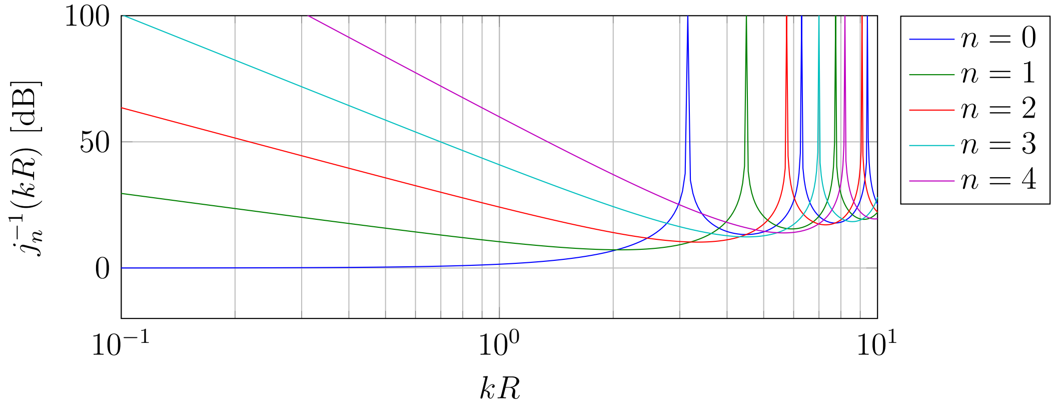


Figure 2.6: Inverse spherical Bessel functions of order 0 to 4. Due to the nulls in the Bessel functions, the inverse goes to infinity at certain points.

By using directional microphones by mounting them on a hard sphere, we avoid this problem. It is also possible to use e.g. cardioid microphones [8]. The weighting filters are then given by Equations (2.17)-(2.18):

$$W_n(kR) = \frac{\mathrm{i}^{n+1}}{(kR)^2 h_n'(kR)} \tag{2.34}$$

The inverse of these filters are easier to realise than the inverse (reciprocal) spherical Bessel functions because the absence of nulls in the radial filters, and the magnitude is plotted in Figure 2.7. However, for orders $N \geq 1$, they lead to high amplification at low frequencies that results in amplification of self-noise and microphone position errors [10]. This is because the wavelength is much larger than the array dimensions, and trying to capture the spatial information is analogous to finding the derivative of the wave field. The derivative is very small at low frequencies, so a noise blow-up is expected. At high frequencies, the inverse filters follow the kR line asymptotically due to screening from the sphere. This is because the large argument limit of $h_n'(x)$ is proportional to $e^{\mathrm{i}x}/x$, and thus $W_n(kR) \propto 1/kR$.

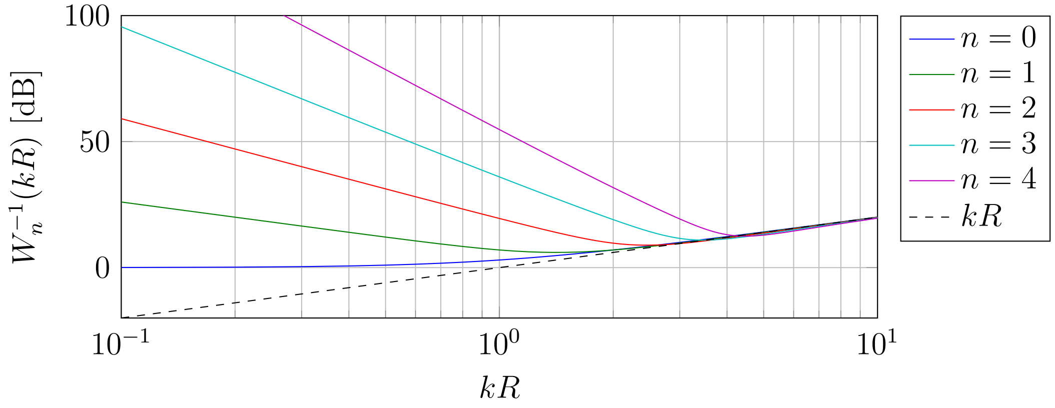


Figure 2.7: Inverse radial filters, along with the high frequency asymptote kR.

To avoid this noise blow-up at low frequencies, several methods are suggested in the literature: In [10] a Tikhonov regularisation filter is applied, from a maximal sensor noise amplification criterion. These are in practice high-pass filters at low frequencies. In [12], the compensation for a finite loudspeaker (Equation (2.24)) distance is shown to counteract the excessive amplification. In [13], high-pass filters with at least a slope of 6 dB/octave/order are applied to compensate for the slope. Figure 2.8 shows the realisable encoder with high-pass filters.

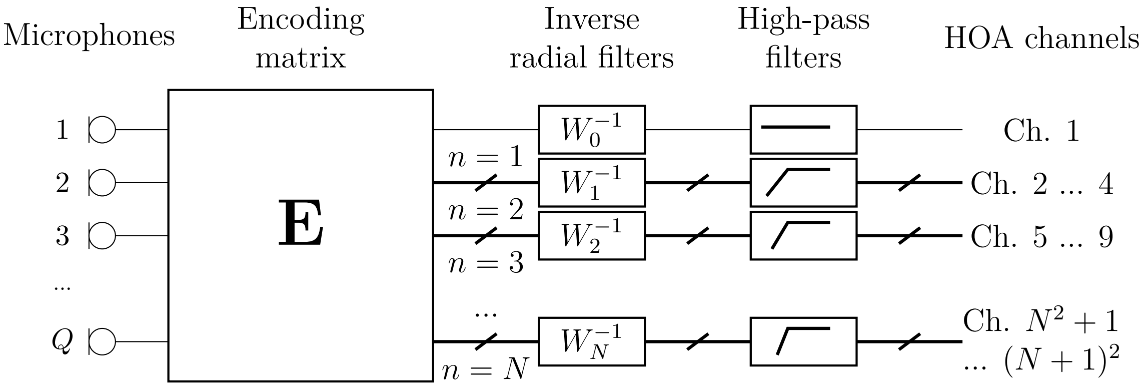


Figure 2.8: Encoding operation for a spherical microphone array, with high-pass filters to avoid noise blow-up.

The requirements for the high-pass filters depends on the microphone Signal-to-Noise Ratio and the desired reproduction accuracy at low frequencies. Requirements for the filters will be discussed in Section 2.3.1.

2.2.3 Sound field rotation

In many cases, it may be necessary to rotate the sound field with an arbitrary rotation operation. Such cases can be: Combining (mixing) several HOA sound fields, repositioning already encoded virtual sources, moving sources, compensating for a moving reproduction system (e.g. binaural reproduction with head-tracking). There are several ways to mathematically describe a rotation, but a common property is that the rotation of an object has three degrees of freedom. Thus it can be represented by three scalar values, much like a translation operation.

The common way to rotate a spherical harmonics representation such as a HOA signal, is with a rotation matrix **R** defined from three *Euler angles* (α, β, γ):

$$\mathbf{B}_{rot} = \mathbf{R}\mathbf{B} \tag{2.35}$$

We now define the 3-2-3 convention Euler angles as 1) rotate α radians around the z-axis, 2) rotate β radians around the *new* y-axis and 3) rotate γ radians around the *new* z-axis. The operation is shown in Figure 2.9.

The rotation matrix is a block-diagonal matrix consisting of $2n{+}1 \times 2n{+}1$-matrices for each order $n = 0..N$:

$$\mathbf{R} = \begin{bmatrix} \mathbf{D_0} & .. & .. & .. \\ : & \mathbf{D_1} & : & : \\ : & : & : & : \\ : & : & : & \mathbf{D_N} \end{bmatrix} \tag{2.36}$$

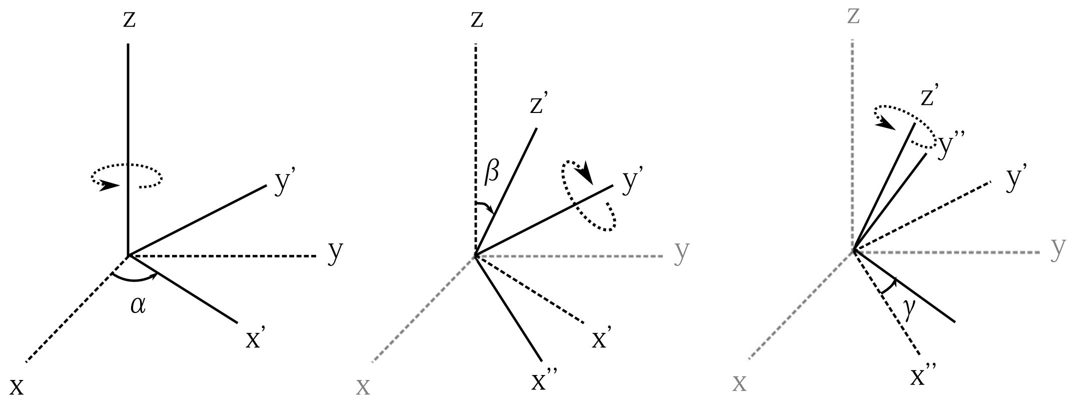


Figure 2.9: Rotation of an object by the 3-2-3 Euler angles (α, β, γ)

For the complex spherical harmonics, each matrix consist of elements:

$$D^n_{mm'} = e^{-\mathrm{i}m\phi} d^n_{mm'}(\beta) e^{-\mathrm{i}m'\gamma}, \quad -n \leq m \leq n \tag{2.37}$$

The expressions for $d^n_{mm'}(\beta)$ can be found in [23]. How to calculate the matrices will not be described in detail here, but the reader is referred to [53] for an effective way of calculating the rotation matrices. Similar matrices for real spherical harmonics and how to calculate them are shown in [54].

2.2.4 Decoding and reproduction of sound fields

Now, assume that a HOA signal with $(N+1)^2$ channels shall be decoded to reproduce the sound field with an array of loudspeakers at a listener's position. It is assumed that the loudspeakers are placed far enough away such that the wavefront is sufficiently plane at the origin. Figure 2.10 shows a set of plane waves arriving from a spherical surface centred at the listener. The region of validity (ROV) for the reproduction is inside this sphere.

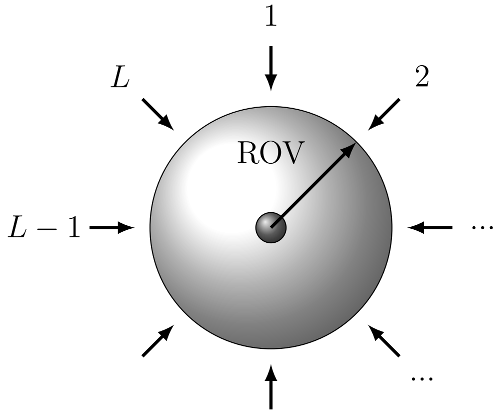


Figure 2.10: Illustration of the reconstruction procedure with plane waves arriving from L loudspeakers, and the region of validity for the reconstruction (see Equation (2.14)). The listener is illustrated as a small sphere in the centre.

Equation (2.21) expresses the truncated sound field as a sum of spherical harmonic components, and is valid when sources are placed outside the ROV. By equating this equation to a sum of plane waves (Equation (2.15), but with real spherical harmonics) from a set of L loudspeakers, we obtain

$$\sum_{n=0}^{N}\sum_{m=-n}^{n} \mathrm{i}^n j_n(kr) B_n^m(\omega) \Upsilon_n^m(\theta,\phi) = \sum_{l=1}^{L} S_l(\omega) \sum_{n=0}^{\infty} \mathrm{i}^n j_n(kr) \sum_{m=-n}^{n} \Upsilon_n^m(\theta,\phi)\Upsilon_n^m(\theta_l,\phi_l)^* \tag{2.38}$$

where S_l is the reproduction amplitude of loudspeaker l radiating from (θ_l,ϕ_l). Expressed in a matrix format

$$\mathbf{JYB} = \mathbf{JYY_L S} \tag{2.39}$$

where $\mathbf{J}$ is the pseudo-diagonal matrix

$$\mathbf{J} = \mathrm{diag}[\mathrm{i}^n j_n(kr)] \tag{2.40}$$

similarly to $\mathbf{W}$ in section 2.2.2. The solution [8] can be found by the equation

$$\mathbf{S} = \mathbf{DB} \tag{2.41}$$

where $\mathbf{D}$ is a decoding matrix consisting of elements $D_{n,l}^m$. The matrix can be found directly if the number of loudspeakers equals the number of HOA components ($\mathbf{D} = \mathbf{Y_L^{-1}}$). As on the encoding side, a minimum of $(N+1)^2$ loudspeakers is required for reproduction, and if the number of loudspeakers is larger, the solution is found with the pseudo-inverse $\mathbf{D} = \mathrm{pinv}(\mathbf{Y_L})$. The loudspeakers must also be distributed regularly on the sphere to avoid ill-conditioning of the decoding matrix. Figure 2.11 shows the simple decoder.

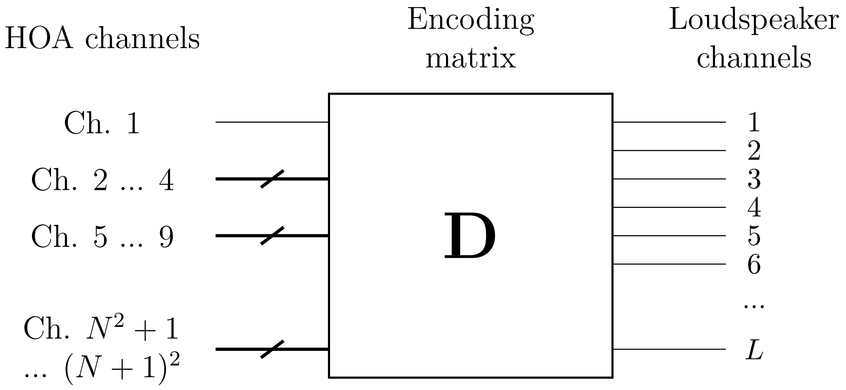


Figure 2.11: Decoding $(N+1)^2$ HOA signals to L loudspeakers.

The decoding method described above has frequently been named the *mode-matching* method [32], because one aims to reconstruct the Ambisonic modes as well as possible. It is though possible to use other decoding techniques at high frequencies, such as Max-r_E and In-phase [3]. These methods use gain factors g_n on the different Ambisonics signals before the mode-matching decoding. It has recently been experimentally shown that the Max-r_E method performs better at higher frequencies [55].

2.3 Sound field capture and reproduction error

Ideally, we want to capture a 3D sound field and reproduce it exactly, with no error. Perfect sound field capture would only be possible if a microphone had a continuous microphone distribution. With a finite number of microphones, the sound field must be truncated to an order N. This gives errors both because of lack of higher modes (truncation error) and the existence of higher order modes in the real sound field, which causes spatial aliasing. Also, a finite number of loudspeakers limits the representation order of a virtual source.

2.3.1 Truncation error

Truncation error is a result of dropping the higher order terms in Equation (2.14). An important part of HOA performance analysis is the error introduced by this truncation. Considering a sound field represented with spherical harmonics, the *normalised truncation error* is defined as

$$\epsilon_N(kr) = \frac{\displaystyle\int_S |p_\infty(r,\theta,\phi,k) - p_N(r,\theta,\phi,k)|^2 \, \mathrm{d}S}{\displaystyle\int_S |p_\infty(r,\theta,\phi,k)|^2 \, \mathrm{d}S} \tag{2.42}$$

where the subscript refers to the truncation order. The integral covers the sphere surface S. p_∞ is the real non-truncated sound field. It is shown [14] that an expression for the truncation error, resulting from truncating the plane wave in Equation (2.15), is

$$\epsilon_N(kr) = 1 - \sum_{n=0}^{N} (2n+1)(j_n(kr))^2 \tag{2.43}$$

in free field conditions, i.e. no scattering objects are present in the ROV (Figure 2.10). The ”rule of thumb” is that for a desired reproduction wavenumber-distance product kr, the HOA order must satisfy

$$N = \lceil kr \rceil \tag{2.44}$$

that is, to limit the normalised truncation error to 4 %. However, when introducing a scattering object in the sound field, such as a human listener, one will have a different truncation error. A simple way to model this is to introduce a spherical scatterer in the centre of the sound field, and calculate the normalised truncation error. The result is

$$\epsilon_{N,s}(kR) = 1 - \frac{\displaystyle\sum_{n=0}^{N} |h'_n(kR)|^{-2}(2n+1)}{\displaystyle\sum_{n=0}^{\infty} |h'_n(kR)|^{-2}(2n+1)} \tag{2.45}$$

which is derived in Appendix A. Here, R is the radius of the rigid sphere.

Figure 2.12 shows the normalised truncation error plotted for the free field and spherical scatterer case. For instance, a 4th order reproduction system with a desired error of 4% may need to be increased to a 5th order system when the scatterer is introduced.

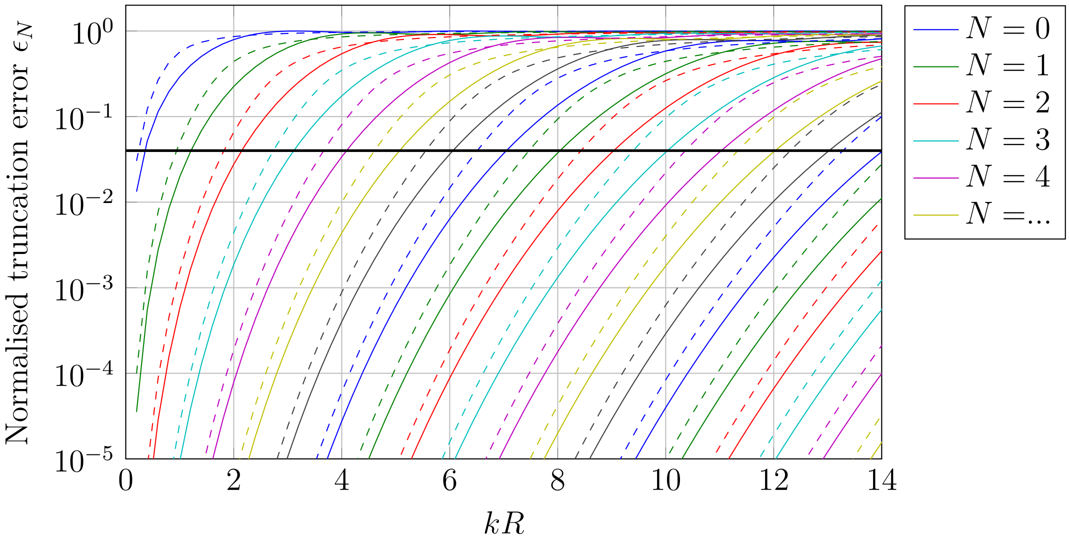


Figure 2.12: Normalised truncation error for free field (solid lines) and free field with a spherical scatterer of size R (dashed lines). The horizontal line represents an error of 4 % (-14 dB).

The equations above are also convenient when calculating the cut-off frequencies needed for the practical implementation of microphone EQ filters. For example when expanding from a truncation order N to $N + 1$, one can calculate the value of kR for the lower order and desired error level, to determine how low in frequency the signals in the highest order needs to be represented. This is further discussed in Section 3.4.2.

2.3.2 Spatial aliasing

Another source of error is the spatial aliasing occurring when capturing the sound field with a microphone array. The Nyquist-Shannon sampling theorem is applicable in the spatial domain. It states that the upper limiting frequency for which spatial aliasing does not occur is

$$f_l = \frac{c}{2d} \tag{2.46}$$

where d is the largest distance between two microphones. The theorem is analogous to the time domain sampling theorem, but this is a more complicated case. A spherical sampling differs from a linear sampling of a sound field, and the signals are also affectted by the rigid sphere geometry. Thus, a comprehensive discussion of spherical microphone aliasing is not included here.

A main point is the fact that a spherical microphone array can achieve a near perfect sampling of a Nth order truncated sound field. However, in the actual sound field, higher order modes will be present, and these ”bleed” into the recorded lower order modes. One can regard this as an under-sampling of the modal sound field. Figure 2.13 shows an example where a 32-capsule Eigenmike® array is used to capture a sound field containing single modes only. The resulting 25 Ambisonic signals suffer from aliasing from the higher order modes. In particular, the 6th order modes (index 37-49 on the x-axis) will show up in the 4th order signals (index 17-25 on the y-axis). Note that the first 25 modes are captured with no aliasing.

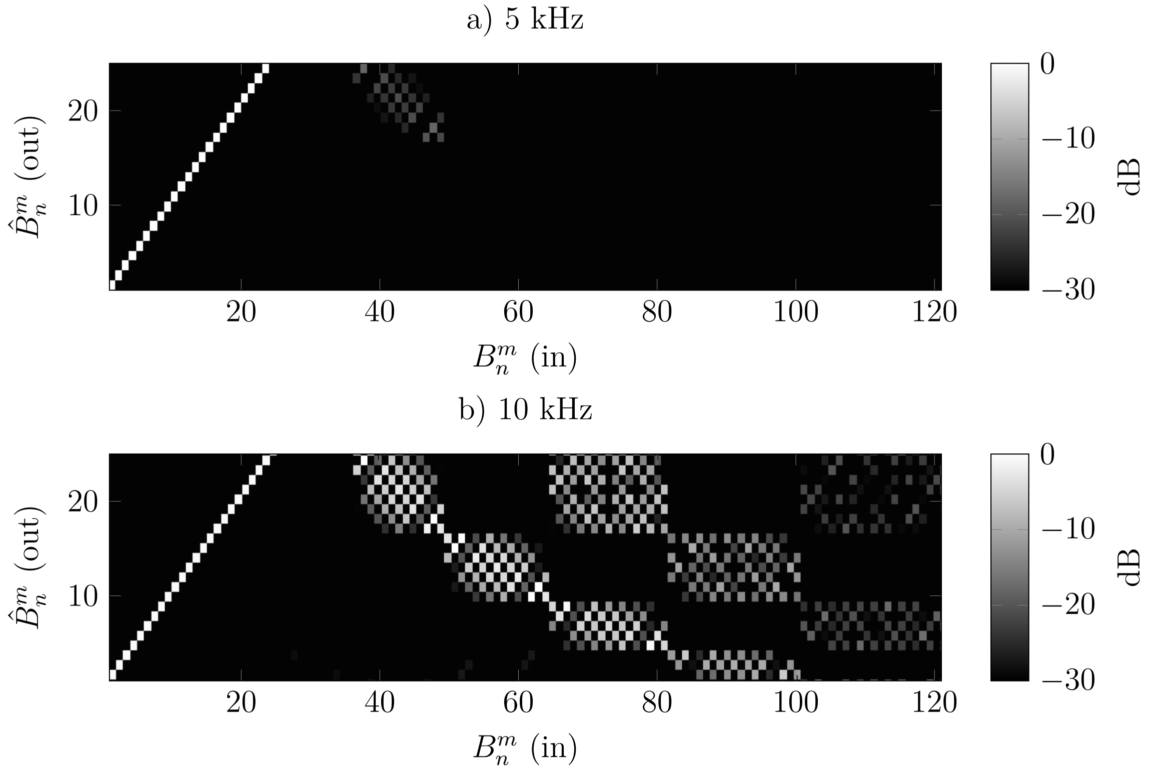


Figure 2.13: Amplitude of encoded Ambisonic signals $\hat{B}_n^m$ when exposed to a sound field containing only the mode B_n^m, from simulated Eigenmike® signals at 5 and 10 kHz. The axis indices represent $(n+1)^2 - n + m$. The inverse radial filters have been applied. Inspired by Meyer and Elko [56].

For a more thorough study on spherical microphone aliasing, sensor noise and positioning errors, the reader is referred to Rafaely et al. [20, 22].

2.4 Basics of binaural hearing

To be able to analyse and optimise the binaural reproduction system, the fundamentals of binaural hearing are reviewed. The human auditory system is an extremely complex system that can localise sound sources with a very good accuracy in some directions. Mills [57] found that the minimum audible angle difference one can perceive is about 1° when a sound source is located straight ahead.

Localisation of sources is mainly facilitated by three mechanisms [58, 59]:

- Interaural Level Differences (ILD) - The sound intensity difference between the ear signals
- Interaural Time/Phase Differences (ITD/IPD) - The time delay/phase shift between the ear signals
- Spectral cues - Localisation based on recognition of patterns in the frequency spectrum

In addition, *head movements* are important to improve front/back localisation.

Interaural level difference
Diffraction and screening from the head will cause the two ear signals to differ in amplitude when the source is not located in the median plane, which is the plane that cuts between the eyes. This mechanism is most prominent at high frequencies where the wavelengths are short compared to the head dimensions. Our threshold for detecting ILD is about 1 dB [60], while the maximum difference is in the order of 20 dB, at 6 kHz [61].

Interaural time difference
In addition to the ILD, the sound will arrive at the ears at different times. At low to mid frequencies, the auditory system can sense this time difference by comparing the phase of the two ear signals. At high frequencies, phase difference estimation collapses due to two reasons: The short wavelengths makes pure tone phase difference estimation ambiguous, and the auditory system senses an amplitude envelope rather than the actual sine signal. Thus, the auditory system uses ITD for localisation only at low frequencies. The threshold of ITD detection is about 10 μs [62], which is astonishingly low and corresponds to a frequency of 100 kHz, far above the audible frequency range. This is the reason for the $\sim 1^\circ$ localisation accuracy when the source is located directly in front of the head.

Spectral cues
For sources placed in the median plane, the ITD and ILD will tend to be zero. However, the shape of the frequency spectra will differ, notably with peaks and notches in the frequency response, called *spectral cues*. This is due to reflections from the torso, a non-spherical head shape and the pinna shape. The literature is not conclusive on which frequency spectral cues determine which directions, but most of the spectral cues are situated in the 4-16 kHz frequency range [63].

Head movements
An important property of sound localisation is the ability to move the head to improve localisation, especially by reducing the front/back confusion [64,65], commonly named *dynamic* localisation. Imagine that a person tries to determine whether a sound source is located ahead or behind. By rotating the head, it is possible to determine the front/back location by detecting changes in ITD and ILD.

The human auditory system uses all four mechanisms to judge the position of a source, or more accurately, the arrival direction of the wave. At low frequencies, the ITD is non-ambiguous and is primarily used to locate the source azimuth. Between 1.5 and 2 kHz, the ITD is ambiguous and the localisation accuracy is at its lowest. Above 2 kHz, the ILD becomes dominating for azimuth localisation. Localisation in the median plane and front/back separation is facilitated by the spectral cues and thus requires a relatively broadband sound to be present.

Distance perception is mostly related to the overall sound level, amount of reverberation and semantic cues [59]. However, for close sources, Brungart and Rabinowitz [66] found that the ILD at low frequencies may be significant for the distance perception.

2.4.1 Head-Related Transfer Functions

The information in ILD, ITD, spectral cues as well as other localisation cues our auditory system use can be represented with a set of transfer functions between points in space and the pressure at the ear drum. In practice these Head-Related Transfer Functions (HRTFs) are often measured from a loudspeaker at a point in space to a microphone at the ear canal entrance. Headphones or earplugs can then be used to replicate the sound pressures at these microphones. With a set of HRTFs covering a sufficient angular domain with a sufficient resolution, it is possible to study the input signal to the auditory system in detail and, even better, synthesise virtual sources with headphones or earplugs.

A major challenge in synthesising 3D audio with HRTFs is the influence of individual head geometry. The auditory system of an individual is very adapted to the exact geometry of the person's head. Thus, for the most accurate representation, the HRTFs must be individualised, either by individual measurements or simulations from a 3D scan [67]. A common assumption is that individual HRTFs is important to realistic binaural synthesis (especially front/back-confusion [68]), and Batke et al. showed that this also applies to HOA rendering [44].

Figure 2.14 shows examples of HRTF magnitude and Head-Related Impulse Response (HRIR) shape for a few azimuth angles. The HRTFs are taken from the Neumann KU-100 measurements by Benjamin Bernschütz [69], further discussed in Section 3.6. The level and time dependency on the incident angle is clearly visible. At low frequencies, the levels remain similar due to the long wavelength compared to the head dimensions, but the time differences are clearly identifiable. At high frequencies, both time and level differences are visible. Peaks and notches in the region above 2-4 kHz will also contribute to elevation and front/back sensation.

2.4.2 ITD estimation

The ITD is defined as the difference in travel time for a wave that reaches the ears. Normally, it must be estimated from measured HRTFs of individuals or manikins. Several methods exist to estimate the ITD from measured HRTFs [70]. However, to obtain the ITD as a function of frequency, one must either consider the HRTF phase, or divide the HRTF into frequency bands (e.g. critical bands [71]). From the phase information, the *phase* and *group* delay can be found. The group delay is defined as

$$\Delta t_g = -\frac{d\varphi}{d\omega} \tag{2.47}$$

and the phase delay

$$\Delta t_p = -\frac{\varphi}{\omega} \tag{2.48}$$

where φ is the unwrapped phase angle of the HRTF. If the HRTF has linear phase, the group delay equals the phase delay $-\varphi/\omega$, and the ITD will be constant for all frequencies. However, one will normally observe that the ITD changes with frequency [72]. In particular, the ITD is somewhat higher at lower frequencies for a given angle of incidence.

The correct way to determine ITD is by considering the group delay, which can be regarded as the envelope delay of a certain frequency component. However, small

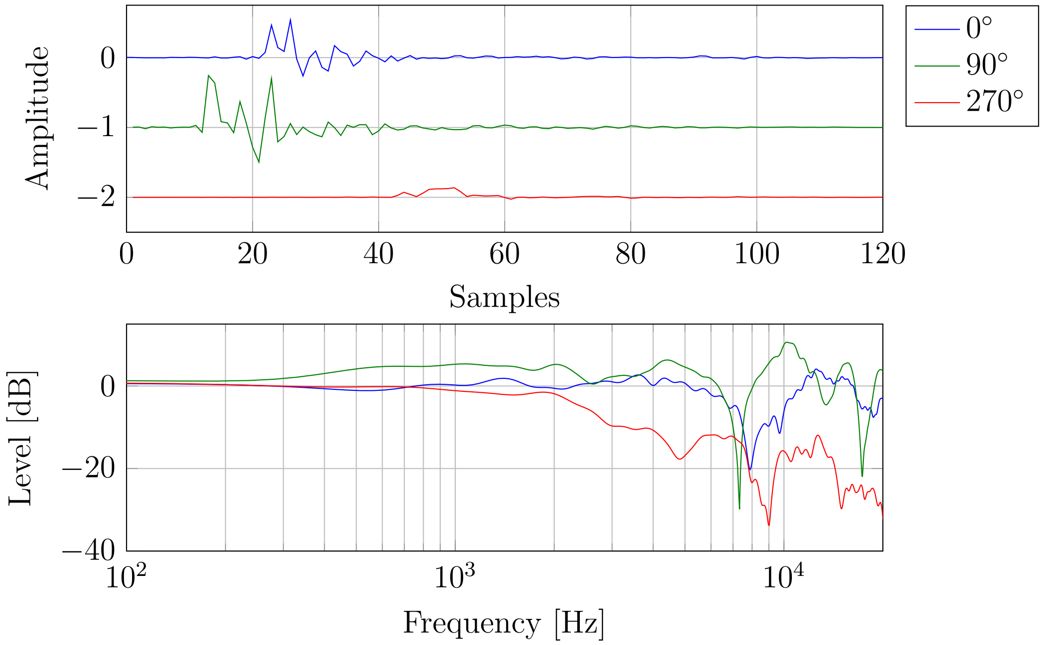


Figure 2.14: Example of HRIRs (above) and HRTFs (below) from Bernschütz' Neumann KU-100 dummy head HRTF library (see Section 3.6). $\theta = 90°$, left ear. The HRIRs are shifted vertically for visual purposes.

variations in the phase angle will result in large ITD variations because of the differential relation. By considering the phase delay, one obtains a much smoother ITD, with the assumption that the phase angle is a linear function of frequency. This would be the case if the ITD was constant with frequency.

Figure 2.15 shows some estimated ITDs for a few incident angles in the horizontal plane. When calculated from the group delay, the ITD estimate shows both that it is highly frequency-dependent, and possibly incorrectly estimated at higher frequencies. This shows that narrow-band ITD estimation is somewhat difficult, and the physically incorrect way of using the phase delay may seem to yield a more reasonable result at high frequencies, since the phase delay is effectively a smoothing of the group delay in the frequency range $[0, f]$. Thus, one should be careful with using the phase delay for ITD calculation without taking the necessary assumptions. Note that the high frequency behaviour of the ITD is less important in this context, because it is mainly used for localisation at lower frequencies.

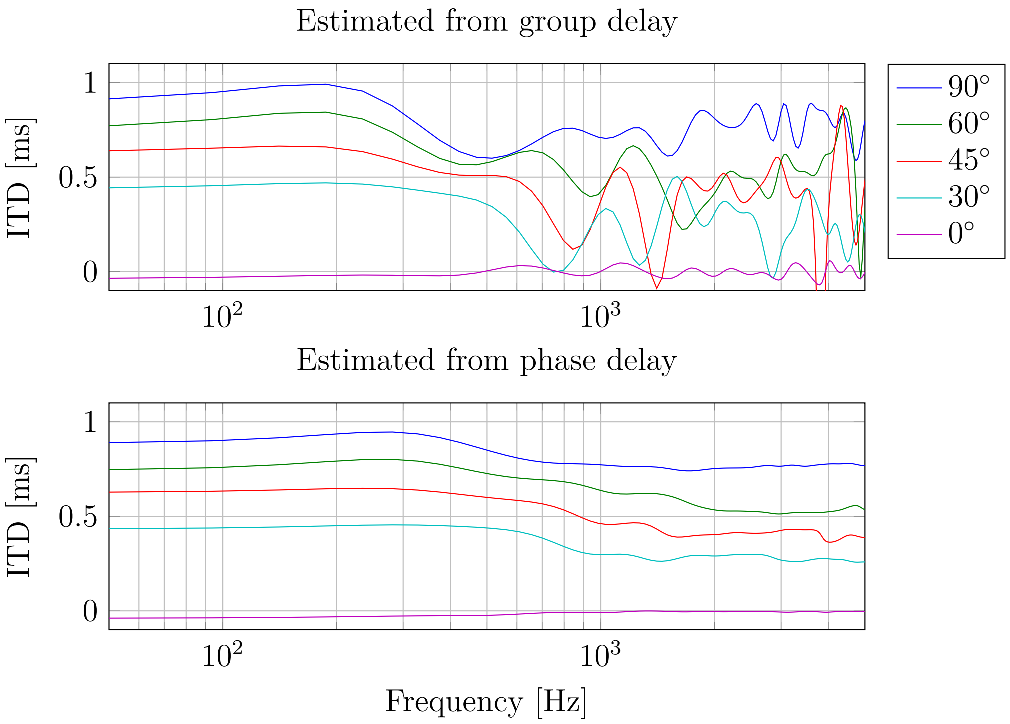


Figure 2.15: Estimated ITDs at certain azimuths angles, from the Neumann KU-100 HRTF library. 0° azimuth is straight ahead.

2.5 Binaural rendering of HOA

A 3D sound field is now represented in the HOA format that can be reproduced with a spherical loudspeaker array (or more precisely, a set of plane wave sources distributed on a sphere, and radiating towards the origin). Since a large loudspeaker array is very impractical in most cases, it can be desirable to reproduce the sound field through headphones. There are two approaches to do this, which yield exactly the same result. The most intuitive way is to decode the HOA signals to a virtual loudspeaker array, and assign two HRTFs to each loudspeaker, one for each ear. Thus, the signal at each ear is the sum of L loudspeaker signals $S_l(\omega)$ filtered with the corresponding HRTFs $H_{l,Left}(\omega)$ and $H_{l,Right}(\omega)$:

$$S_{ear}(\omega) = \sum_{l=1}^{L} H_l(\omega) S_l(\omega) \tag{2.49}$$

Since each loudspeaker signal is a mix of the HOA signals (Equation (2.41)), the sum can be written as

$$S_{ear}(\omega) = \sum_{l=1}^{L} H_l(\omega) \left(\sum_{n=0}^{N} \sum_{m=-n}^{n} B_n^m(\omega) D_{n,l}^m \right) \tag{2.50}$$

which can be rearranged to

$$S_{ear}(\omega) = \sum_{n=0}^{N} \sum_{m=-n}^{n} B_n^m(\omega) H_n^m(\omega), \tag{2.51}$$

$$H_n^m(\omega) = \sum_{l=1}^{L} D_{n,l}^m H_l(\omega) \tag{2.52}$$

i.e. we need to pre-compute a set of spherical harmonics-based HRTFs $H_n^m(\omega)$ for each ear.

The second approach to binaural HOA reproduction is exactly this operation – the HRTF set is converted to a spherical harmonics representation [39] by solving the equation

$$H(\theta, \phi, \omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} H_n^m(\omega) Y_n^m(\theta, \phi) \tag{2.53}$$

where the spherical harmonics coefficients (in practice, truncated to an order N) must be determined either by direct integration or by the least squares solution. Thus, the problem of ill-conditioned matrices must be considered, which requires a uniform (or dense enough) sampling of the sphere.

Care must be taken when choosing the HRTF set, and in particular, the number of HRTF measurement positions. Solvang [43] showed that the number of positions is a trade-off between the reproduction error at lower frequencies ($kr < N$) and "spectral impairments" at higher frequencies ($kr > N$), which will reduce the high frequency levels due to loss of high-frequency energy present in the higher order modes.

One can regard each spherical harmonics-based HRTF H_n^mas the transfer function between a wave field that contains only that mode, and the ear. For example, a spherical wave traveling inwards to the origin will yield the transfer function H_0^0.

2.5.1 Phase correction at high frequencies

Now, a new method for improving the reproduction at high frequencies is presented. The method was originally suggested by A. Solvang[2], and further developed by the author.

From the previous discussion and theory, it is evident that the high frequency limit for decent reproduction is somewhere around

$$f_{lim} = Nc/(2\pi r). \tag{2.54}$$

However, the reproduction would still be near perfect for smaller radii, and this can be further exploited. At low frequencies, the ITD is essential for localisation, and thus phase preservation is important. The system described in Chapter 3 is a 4th order system that gives good reproduction up to 2.2 kHz inside a sphere of radius 0.1 m. In practice, the ears are placed closer to the origin than 0.1m for most humans, so this is a conservative estimate. Consequently, the ITD is mainly preserved in the desired frequency range. However, at high frequencies, ILD becomes increasingly important and the ITD becomes ambiguous. It then seems reasonable to pursue an improvement

[2] Research scientist at SINTEF ICT, Audun.Solvang@sintef.no

in ILD at high frequencies, contrary to improving both the ILD and ITD, as shown in Figure 2.16. In addition, improvement in spectral cues and timbre at high frequencies is desirable. This motivates for an overall improvement of the magnitude spectrum above f_{lim}, but not the phase spectrum.

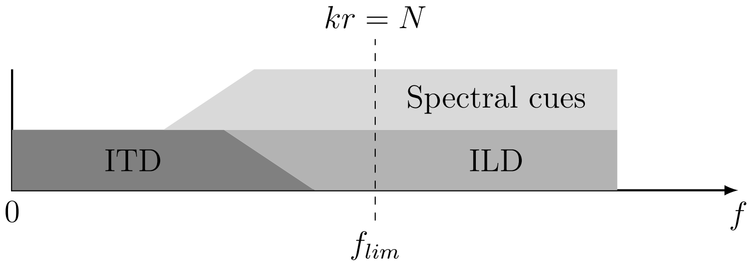


Figure 2.16: The different localisation methods and their relation to the limiting reproduction frequency. At low frequencies, ITD is the primary input. At mid frequencies (1-2 kHz), there is a transition from ITD to ILD. For the method to work, f_{lim} should be above this mid-range. At mid to high frequencies, spectral elevation cues are detected.

To obtain a more accurate high frequency reproduction, the observation points can be moved closer to the origin, to maintain a constant value of kr. This can be done by moving the ears' observation points towards the origin as the frequency increases. As shown in the spherical head model in Figure 2.17, the time delay at each ear will then converge to some value, T_0, as the frequency increases. Time delay, and consequently phase angle, can be calculated by considering the travel distance from a point in space $S(r, \theta, \phi)$ (the measurement loudspeaker) to the ear points on the sphere. The simple model assumes that the loudspeaker is far enough away to assume plane wave incidence, and that the ears are placed at azimuths ±$\pi/2$.

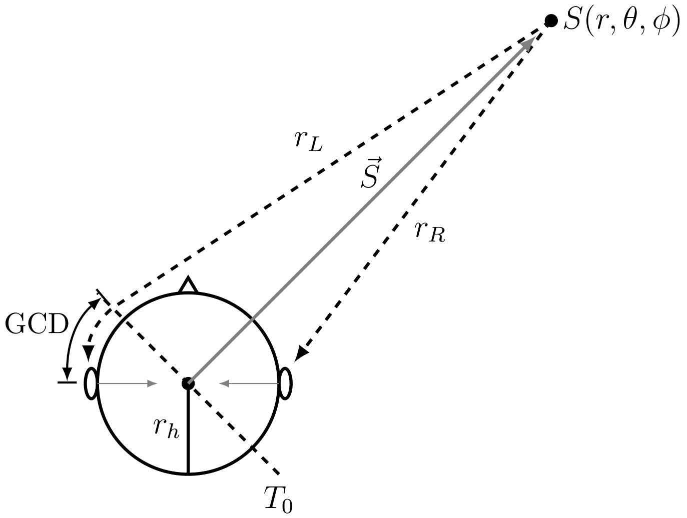


Figure 2.17: The geometry of a spherical head and the operation of reducing the effective head radius. The small, gray arrows illustrates the process of reducing the head radius at high frequencies that result in different travel distances r_L and r_R.

The calculation of the difference in travel time in Figure 2.17 can then be done as follows (for each HRTF):

1. For the closest ear, find the direct distance (r_R on the figure).

2. Find the intersection between the sphere and the plane that is normal to the vector $\vec{S}$ and runs through the origin. This intersection is a circle in space $\mathbf{S}$. Then, find the distance from S to this circle.

3. Find the closest point on the circle $\mathbf{S}$ to the other ear (in this case the left ear), and find the great circle distance (GCD) from this point to the ear. The travel distance r_L is then the distance from 2) plus the GCD.

4. Above f_{lim}, reduce the sphere radius such that the value of kr_h is constant, i. e. equal to N.

5. Repeat steps 1-3 for the sphere with reduced radius, calculate the travel time differences $\Delta t_L, \Delta t_R$ and the resulting phase angle difference.

6. Add the differences in phase angle to each HRTF in question, by multiplying with $e^{-i\omega\Delta t}$

A second, more simple approach is to assume that the head radius is zero at frequencies above f_{lim}, and thus forcing the time delay to equal T_0 in this region. This is done by linearising the phase angle with a slope of $\frac{d\phi}{d\omega} = -T_0$.

To illustrate the effect of these phase corrections, an example of a phase corrected HRTF set is shown in Figure 2.18. The plots show the HRIR amplitude as function of azimuth angle and time, in different frequency bands. Clearly, the HRIRs are nearly unchanged in the 1 and 2 kHz bands as expected. At high frequencies, the amplitude envelope is more independent of azimuth when the phase corrections methods are applied. In the 8 kHz band, the radius reduction method seems to preserve more of the fine details in the HRIRs, while the linearised phase method seems to concentrate all the energy between 0.5 and 1 ms, removing much of the phase details. It is difficult to determine how this will affect the reproduction, but this will be further investigated in Chapter 4.

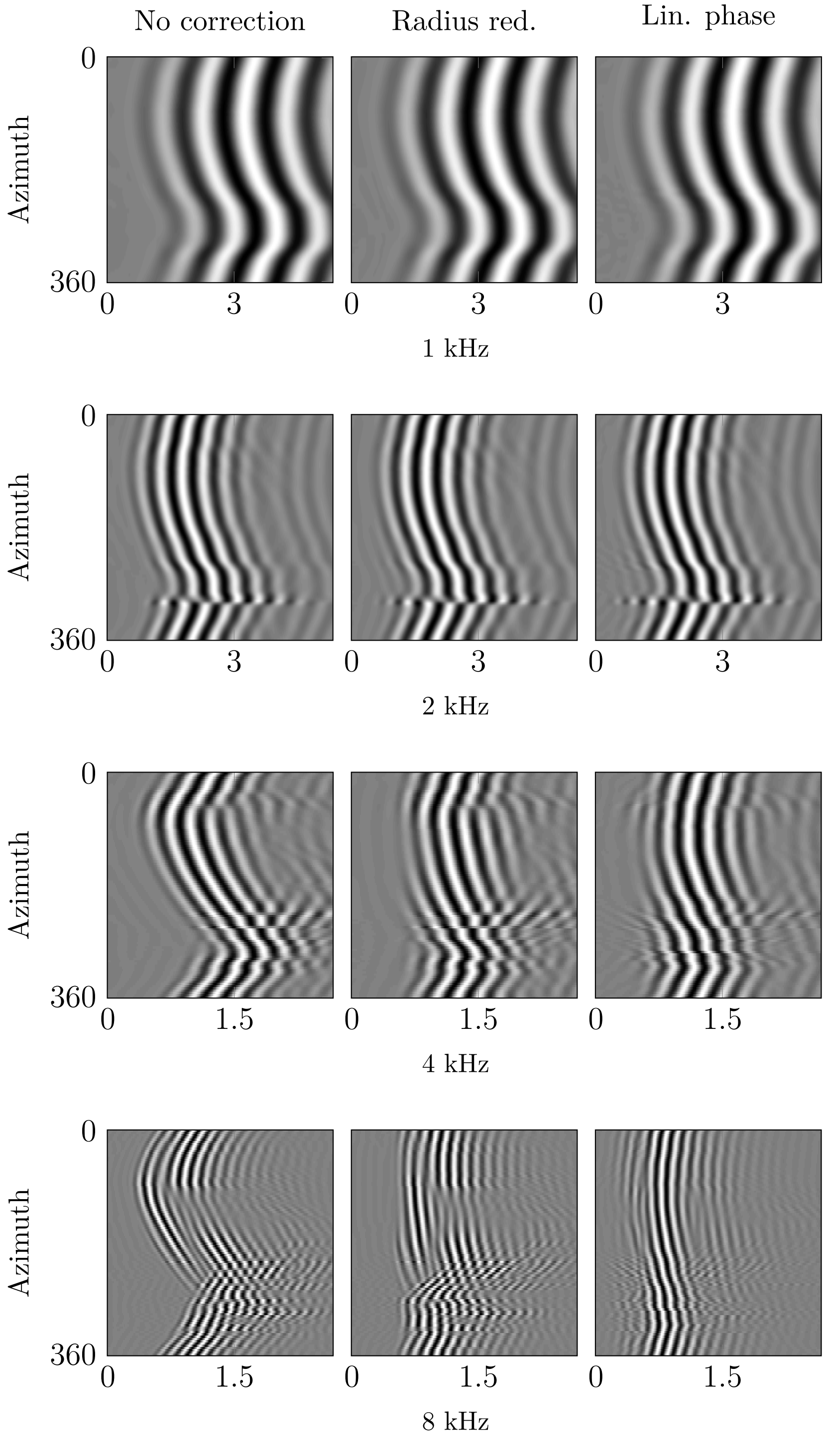


Figure 2.18: Example of the phase correction methods (Radius reduction and Linearised phase). The resulting HRIRs are filtered with 1/3-octave band filters with the given centre frequencies. Brightness represents amplitude, where black is negative and white is positive values. Note the smaller time scale on the 4 and 8 kHz plots. Neumann KU-100 HRIRs.

CHAPTER 3

Implementation of a real-time system

In this chapter, a signal processing system for binaural Higher Order Ambisonics reproduction is presented. The system can either use a microphone array as an audio source, or virtual sources may be created in the space around the listener. Then, the system is implemented as a real-time system where a microphone array or virtual source input is captured, processed and simultaneously reproduced over headphones. A commercial head-tracking system is used to compensate for head motion. Finally, the system performance is evaluated in terms of resource use and latency.

3.1 Hardware

The real-time auralization system needs a limited amount of hardware to operate. Five main components constitute the hardware:

- A spherical microphone array
- A data acquisition device (sound card)
- A signal processor (e.g. personal computer)
- A head-tracking device
- Headphones

The em32 Eigenmike® from mh acoustics[1] (Fig. 3.1) was chosen as the spherical microphone array to be used in the system. The Eigenmike is a hard sphere with of radius 4.2 cm which has 32 microphone capsules (1/2”) mounted nearly uniformly on the surface. With 32 sensors, it is possible to create a 4th order system with 25 HOA channels. Inside the sphere there are microphone preamplifiers and AD converters, so only one digital cable connects the array to the separate power supply and FireWire unit. The audio is then fed to a computer via a low-latency FireWire bus, and the microphone signals are available as a 32-channel sound card on the computer. Thus, separate sound card hardware is not required, except for the headphone output.

[1] http://www.mhacoustics.com/products (28.04.2014)

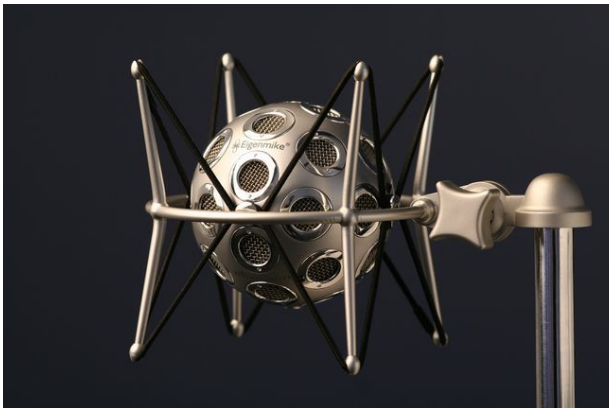


Figure 3.1: The em32 Eigenmike® spherical microphone array. *Photo used with permission from mh acoustics.*

Any PC can be used to perform the signal processing, but the code was implemented and optimised on Mac OS X. It can easily be ported to run on Windows or Linux, though one may have to use different toolboxes for audio I/O and motion input. An Intel Core 2 Duo-based machine was sufficient to run the processing without glitches. The built-in sound card was used to feed headphone audio.

Several motion tracking sensor systems are commercially available, and the Freespace® FSM-9[2] (Fig. 3.2) was chosen in this specific system. The sensor communicates via the USB Human Interface Devices (HID) protocol, and provides inertial, angular and directional information, with a sample period of minimum 2 ms. It is small in size and can easily be mounted on a pair of headphones. A combination of latency, connectivity, size and price was used to decide the choice of sensor.

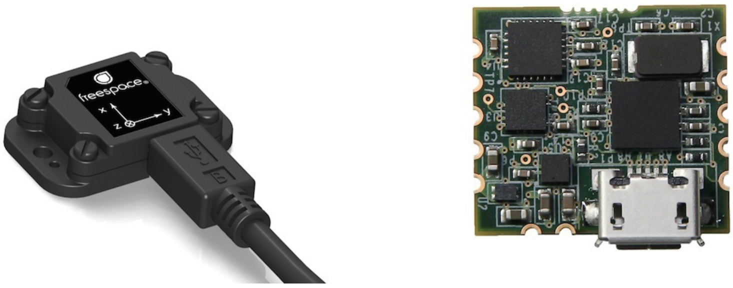


Figure 3.2: Hillcrest Labs Freespace® FSM-9 motion sensor, with (left) and without (right) casing. *Photos used with permission from Hillcrest Labs.*

The listener can freely choose which headphones to use, as this should not influence the reproduction quality considerably. It is recommended to use a good quality pair of headphones to avoid any possible quality degradation.

[2] http://hillcrestlabs.com/products/sensor-modules/fsm-9/ (28.04.2014)

3.2 Practical HOA processing

The strength of HOA is that a complex sound field can be coded in a multi-channel format that contains full 3D spatial information. Any number of sound field recordings and virtual sources can be coded and mixed together into a data stream of finite size. For simple sound fields such as a single virtual source, the audio may be stored or transmitted in a single-channel format along with meta data describing the spatial information of the source. The HOA encoding/decoding operation can then be performed with simple loudspeaker weights at the playback device. However, in a complex environment, e.g. with many reflections, the number of virtual sources quickly exceeds the number of HOA channels. Then it is more sensible to transmit the audio on the HOA format.

In general, a HOA system is composed of an encoder, a transmission or storage medium, and a decoder at the receiver side. Figure 3.3 shows conceptually how the system works. The encoder receives virtual sources or microphone array signals and encodes them to the HOA format. The $(N+1)^2$ HOA channels are then, if needed, summed with other HOA signals of any order, and the decoder creates audio signals that will be fed to a listener through loudspeakers or headphones. Conventional mono, stereo or surround sound system signals may also be decoded from the HOA signal (see e.g. [73]).

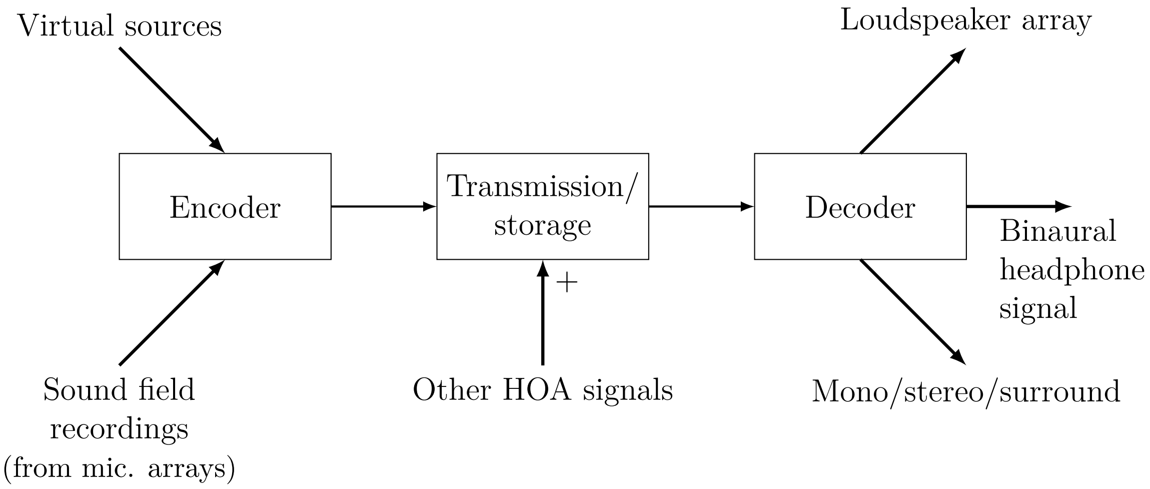


Figure 3.3: Conceptual block diagram of a HOA system, based on the transmission model of communication.

For virtual sources, the encoder is a very simple operation only consisting of the gain multiplication in Equation 2.20. Distance coding filters [12] must be implemented if near-field sources are to be considered, otherwise only a $1/r$ gain factor and possibly high frequency air absorption attenuation has to be considered. The distance coding was not implemented in this version due to the main focus on spherical microphone arrays as sources, which does not need distance coding. In addition, the HRTFs were regarded as far-field sources, as is common in auralization.

When a spherical microphone array is used, the encoding process consists of two operations. First, the signals are mixed with an encoding matrix $\mathbf{E}$ that creates 25

HOA signals (for the 4th order system). These signals have to be equalised to compensate for the microphone geometry. The equalisation filters in Figure 2.7 must therefore be implemented in a practical manner. The filters are defined in the frequency domain, so either the filtering must be done by multiplication in the frequency domain, or by constructing a time domain filter, e.g. by frequency domain sampling. Due to the infinite low frequency amplification, the filter responses must be constrained. Out of the methods suggested in Section 2.2.2, the high-pass filter approach was chosen. Design of the finite impulse response (FIR) filters are more thoroughly covered in Section 3.4.2.

In the HOA system in question, transmission or storage is not an issue. Audio is captured, processed and reproduced on the same device, so the data is only stored in the device's memory. However, in a more practical system, the data rate is quite large (25 channels of uncompressed audio), so a compression algorithm would be needed in many cases. Such algorithms are discussed by Hellerud et al. [51, 52].

The decoder consist of a simple decoding matrix $\mathbf{D}$ and HRTF filters. These are combined to form decoding filters $H_n^m(\omega)$ for the left and right ear. Thus, a total of 50 spherical harmonics-based HRTF filters are needed in a 4th order system. Finally, the 25 filtered channels per ear is summed to form the binaural signal. Further details on the HRTF database is discussed in Section 3.6.

3.3 Real-time audio signal processing

Applied signal processing can roughly be divided into two fields; real-time or non real-time. The latter concerns all signal processing operations where the outcome does not need to be ready at a specific point in time. Such operations can be media coding, post-acquisition data analysis or simulations. Real-time systems require that the processed signal is ready for the user at a certain time, such as in teleconferencing, TV streaming, or even aircraft system controls. There exist many definitions and levels of real-time programming, however, details on the more critical parts of real-time systems are not discussed here.

In real-time audio processing, two common demands are that the processed audio must be free of glitches/drop-outs, and the latency from the input to the output must be sufficiently small. For example, in a mobile phone conversation, frequent dropouts are usually unacceptable and the latency must be small enough to perform a conversation. Normally, real-time audio processing is performed on dedicated hardware or software solutions, so the demands can be satisfied as long as the code is efficient and no interruptions occur. However, in certain systems such as on a PC, one cannot always guarantee that the desired system resources are available. This is because most operating systems will prioritise other tasks and applications. The way around this is normally to accept some interruptions and try to make the processing as efficient and close to the hardware level as possible.

In the most common operating systems, audio input/output (I/O) is usually controlled by the audio Application Programming Interface (API), such as ASIO, Core

Audio or PortAudio[3]. The audio API allows a programmer to easily communicate with the audio hardware within the code, without needing to think about low-level system operations needed to perform the data transfer.

Real-time audio processing is performed in *frames* or *blocks*, or sample-by-sample. Frames are needed when e.g. FFT operations are done, or when the time to communicate with the audio API is too long to allow receiving only one sample at each iteration. Sample-by-sample processing reduces the latency because there is no need for I/O buffering of the audio signals. On PCs, frames are normally needed because of the API, and DSPs can provide both methods[4].

3.3.1 MATLAB as real-time software

Even though MATLAB was primarily designed for research and algorithm prototyping and thus not suited for most real-time applications, there are a few options available to process audio in real-time. Fundamentally, the programming language is *interpretive*, which means that each code line is interpreted "on the fly" and executed by the interpreter. Thus, the code is not complied to machine code, and will often be executed inefficiently. This particularly affects loops and the lack of multithreading abilities. However, many built-in functions such as the `fft` algorithm is pre-compiled and thus quite efficient. In addition, MATLAB can execute `.mex` files that are pre-compiled with C code. This also enables 3rd party developers to program interfaces to new devices, such as USB HID (see Section 3.4.3).

For audio I/O, MATLAB has a built-in framework that supports real-time recording and playback, *DSP System Toolbox*. One can create `dsp.AudioRecorder` and `dsp.AudioPlayer` objects that support any number of I/O channels supplied by the sound card. Receiving and sending data is done with the `step` command. A good 3rd party audio API is *Psychophysics Toolbox*[5], which offers the `PsychPortAudio.mex` program based on the low-latency PortAudio API. The latter alternative is a bit more difficult to implement, but turned out to be more efficient resulting in lower latencies without dropouts. However, both toolboxes facilitate real-time audio processing with latency of a few hundreds of a second, and both were implemented in the real-time system.

The most time-consuming operations in the HOA-processing algorithm are the EQ and HRTF filter banks. These can be implemented as FIR filters, with the `filter` function. A more efficient implementation is the `fftfilt` function that uses the overlap-add method to perform FIR filtering [74]. However, since the `fft` and `ifft` functions are efficiently implemented for 2D arrays (filtering along only one of the dimensions), they were found to be the most efficient filtering method. Since they are implemented with multithreading capabilities, multicore computers benefit from this. A single call to these functions can perform the whole filter bank operation without any need of loops.

MATLAB was found to be a suitable tool for the task with only the need for a few toolboxes. Since it is well suited for prototyping, the system can easily be modified without the need of low-level coding. However, it is not suitable for creating a final

[3] www.portaudio.com

[4] Such as the Analog Devices ADAU1452

[5] http://psychtoolbox.org/

product, which would need stable, efficient and low-latency real-time processing, as well as low cost.

3.4 System overview

Figure 3.4 shows a block diagram of the real-time processing system. A frame-processing structure is used, which means the audio is processed in frames of e.g. 1024 samples. This is done for two reasons: sound card I/O is rather time consuming, and the filtering is speeded up because efficient FFT filtering can be used.

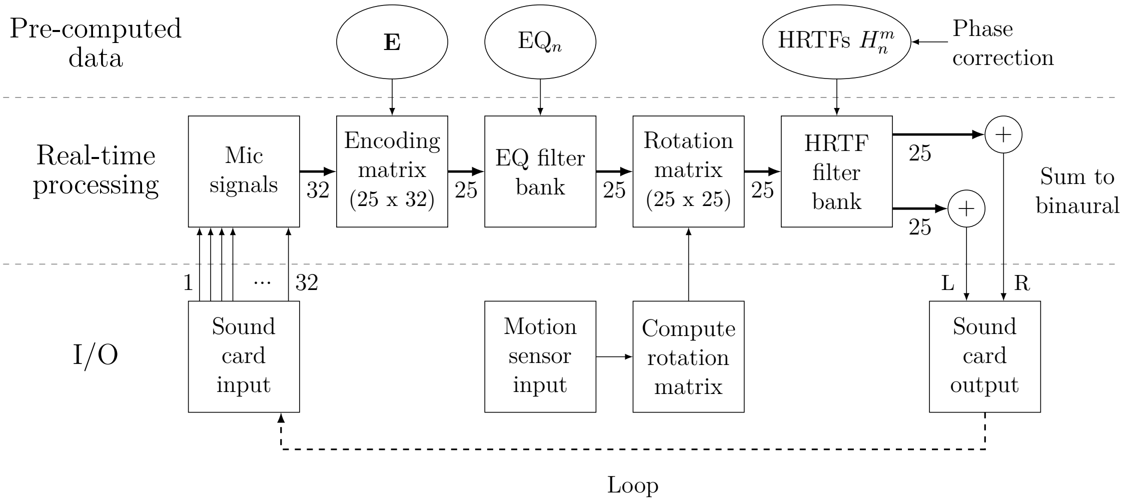


Figure 3.4: Block diagram of the real-time HOA capture and reproduction system. The FFT/IFFT operations are not shown for simplicity. Thick lines represent groups of signals.

First, the filter data and encoding/decoding matrices are pre-computed (or loaded from a file). Then, the audio I/O and motion sensor input is initialised. A loop is then entered, which sequentially processes each audio frame. The looping code will first wait for new audio to be available from the audio API. Once new data is available, the audio frame is multiplied with the HOA encoding matrix, converted into the frequency domain and FFT filtered with the microphone equalisation filter bank EQ$_n$. Note that the data is now compliant with Daniel's HOA format in the frequency domain. Subsequently, the motion sensor data from the head-tracker is read and a rotation matrix $\mathbf{R}$ is constructed, that is used to rotate the sound field in the opposite direction of the head motion. The rotation matrix is then multiplied with the HOA signal. Finally, the signals are FFT filtered with the spherical harmonics based HRTFs H_n^m, converted back to the time domain with an IFFT and summed to form a binaural signal. This two-channel frame of samples is then fed back to the audio API.

3.4.1 Latency and buffer size

Latency can be divided into three categories: System latency (related to hardware and APIs), computational latency (Central Processing Unit, CPU, time) and algorithmic latency (filter delays). There are two relevant latencies in the real-time system, as shown in Figure 3.5. The most important one is the *end-to-end* latency [75] between head motion and the resulting sound field rotation ($T_2 - T_1$), which is a combination of sensor latency, USB communication latency, sound field rotation, HRTF filtering, and sound card output latency. It is a somewhat challenging task to measure all of these separately, but some of them can be calculated or estimated (see Section 3.5). The sensor and USB communication latency is in general unknown, but can be measured. For this specific system, computational latency is the CPU processing time needed to compute the rotated HOA signals and perform the HRTF filtering, and can be estimated with MATLAB's profiler. Algorithmic delay consists only of the HRTF filter delay. The sound card output latency is mainly determined by the buffer size, but possibly by other mechanisms in the audio API and hardware as well.

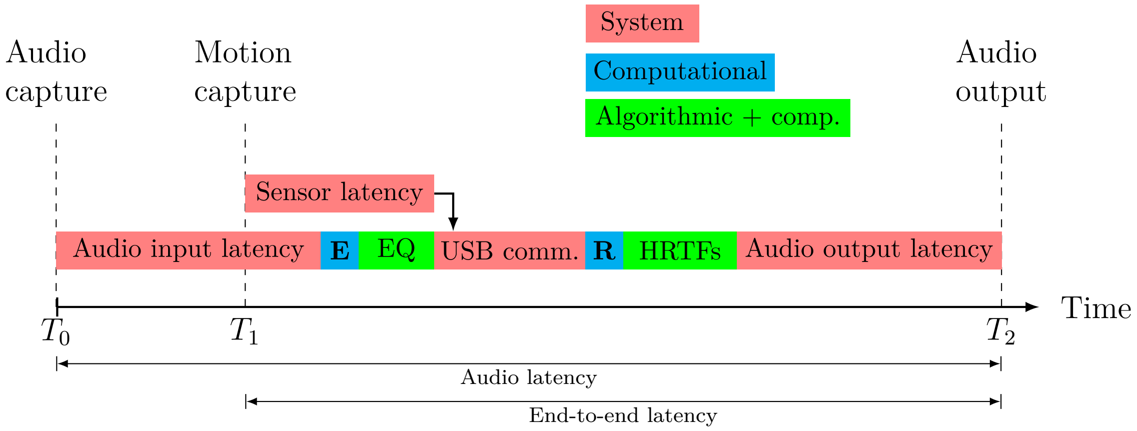


Figure 3.5: The main components of the system latency, and their categories. Here, the algorithmic delays are merged with the corresponding FFT/IFFT computation time.

A second but less important latency is the *audio latency* between the audio input and output (T_2 - T_0), which excludes the sensor latency, but includes audio input latency, the multiplication with **E** and the EQ filter delay. These must be estimated in the same way as the end-to-end latency components. Since the audio capture will often happen on a different computer, the measured audio I/O latency is not representative for a practical system. Such a system could be e.g. a teleconferencing system.

DSP System Toolbox requires a specified buffer size to operate. PsychPortAudio operates in a slightly different way, and determines the buffer size from a user specified latency requirement. Buffer latency can be calculated as the buffer size divided by the sample rate, e.g. a typical buffer size of 1024 samples will cause 23.2 ms of latency at a sample rate of 44.1 kHz. Assuming the input and output buffer sizes are equal and create equal amounts of delay, the total audio latency will be *at least* 46.4 ms. Thus, a main limitation on both end-to-end latency and audio latency is the audio I/O buffers.

3.4.2 Microphone EQ filtering

As discussed in Section 2.2.2, the radial equalisation filters $1/W_n(kR)$ must be modified to avoid excessive low frequency amplification. Moreau and Daniel's approach was followed by introducing high-pass filters to counteract the steep slope [9]. A requirement for these filters is that they preserve the phase in the pass-band to avoid phase mismatch.

The required reproduction order can be estimated from Equations (2.43)-(2.45), by selecting a desired error level, reproduction radius and frequency limit. Consequently, for a given reproduction order, one can estimate how high in frequency one will have a sufficiently small error. Thus, higher order HOA signals will not contribute much to the lower frequency bands, and the low frequency components can be filtered out. Figure 3.6 shows the lower limit of the ”useful frequency bands” as defined by Daniel, estimated with an error level of 4% (-14 dB) . This figure yields the cut-off frequencies for the high-pass filters. For example, the 4th order modes only need to be included above 1.5 kHz (1.7 kHz with a rigid sphere).

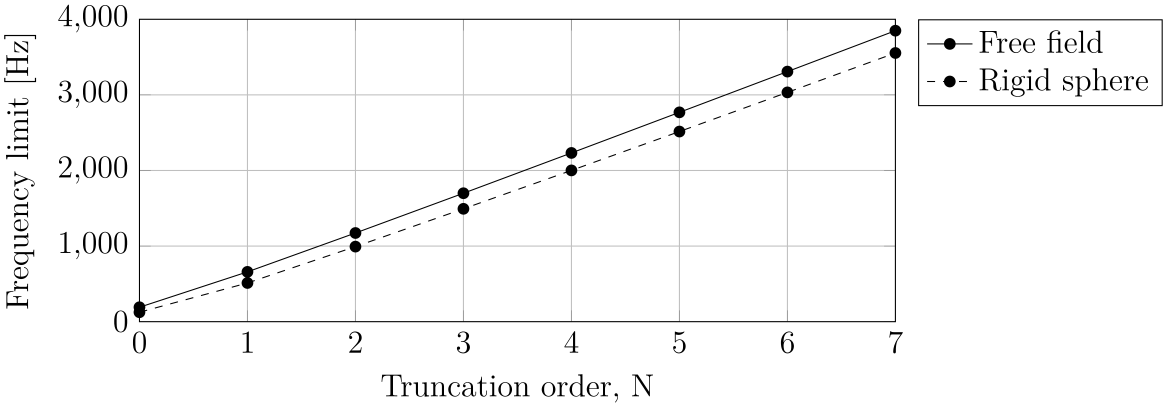


Figure 3.6: Frequency limits for 4 % normalised truncation error as a function of truncation order. The reproduction radius is 0.1 m, both free-field and with a 0.1 m rigid sphere.

Figure 3.7a) shows the radial filters multiplied with the magnitude of nth order Butterworth filters, which counteracts the asymptotic low-frequency behaviour. Discarding the phase of the high-pass filters preserves the phase of the radial filters. An error criterion of 4 % results in just below 40 dB of gain up to 2 kHz for the 4th order filter. Depending on the amount of microphone noise, this limit should be adapted to avoid noise annoyance, at the cost of reproduction accuracy and thus directivity.

By using frequency-sampling of the desired frequency domain response, FIR filters are obtained by multiplying with a time delay factor $e^{-\mathrm{i}\omega\Delta t}$ to obtain causality and performing an IFFT. The filter length was constrained by multiplying the impulse response by a Hanning window of 256 samples and removing the resulting zeros outside the window. This will result in slight magnitude and phase errors, as shown in Figure 3.7b)-c). However, the magnitude errors are only large at orders 1 and 3, at low frequencies, while phase errors are below 0.005 radians above 150 Hz.

It can be argued that further high-pass filtering should be implemented, to limit the amount of low frequency noise in the higher order modes. As high-pass filters with a low cut-off frequency cannot be implemented with short FIR filters, such filters

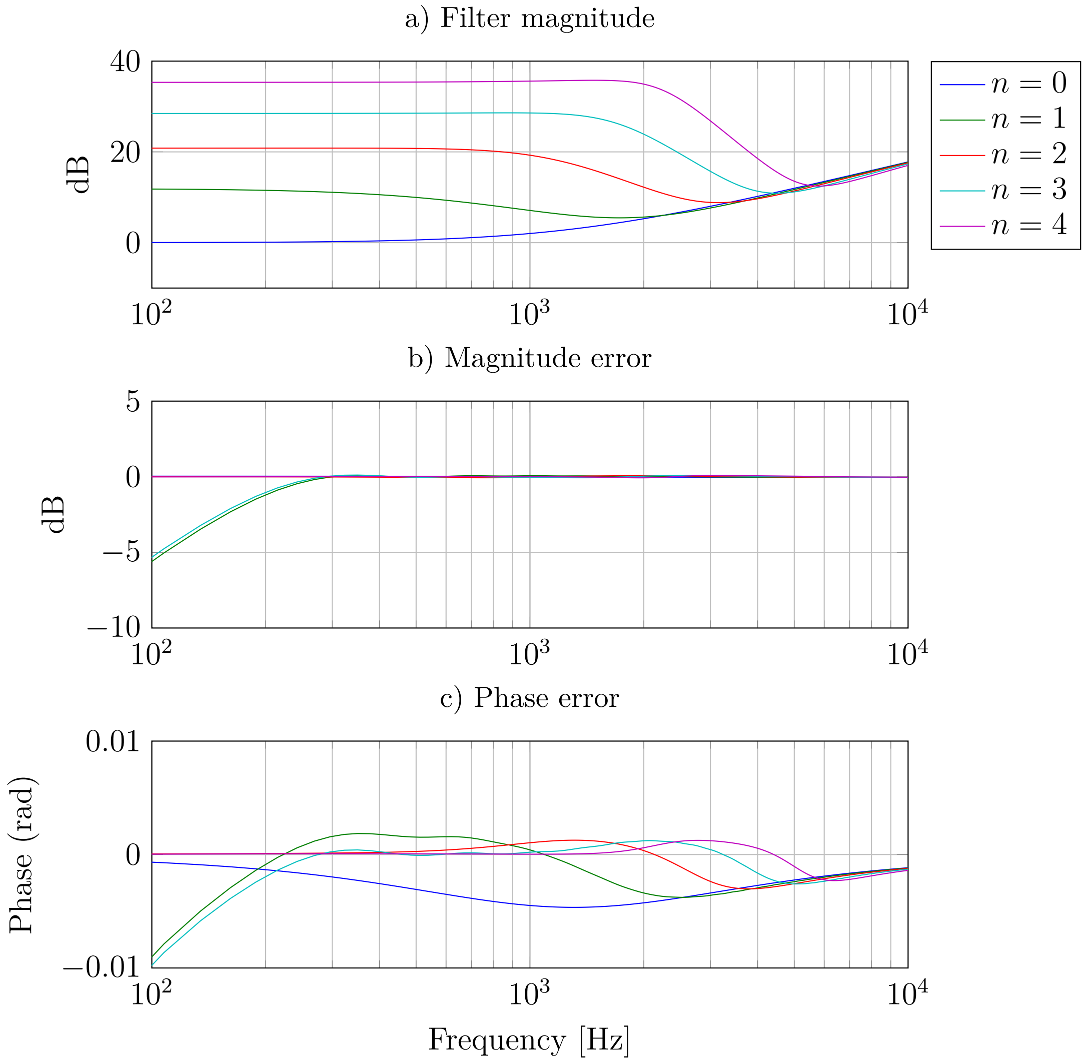


Figure 3.7: Design of the spherical microphone array FIR filters. a) Ideal filter magnitude (radial filters combined with high-pass filters). b) Magnitude and c) phase phase error when reducing to a 256-tap FIR filter with a Hanning window.

should be designed as IIR filters to minimise processing requirements. Phase effects must then be considered, possibly by reversing the high-pass filter phase response with the FIR filters. However, the observed noise levels were moderate so further high-pass filtering was not a prioritised task.

3.4.3 Motion input

To obtain a realistic perception of sound sources, including dynamic HRTF effects that results from head movements, it is essential to use head-tracking in binaural reproduction of 3D audio [64, 76, 77]. Head-tracking also enables the listener to try and localising the source by moving the head.

The Freespace® FSM-9 uses USB HID as communication protocol, but MATLAB does not have built-in support for USB HID. However, Psychophysics Toolbox provides a `.mex`-file, `PsychHID`, that can communicate via USB HID. Communication is mainly done with *reports*, packages of data, both for configuring the sensor and receiving motion data. Specification of the report protocol is thoroughly documented in the Hillcrest Labs HCOMM manual [78]. Basically, the sensor is initialised by determining its device ID and setting the operating mode (Full Motion On) and sampling period. In addition, a `PsychHID` timing parameter must be set set to minimise the motion data reception time consumption in the real-time loop.

Motion data is supplied from the sensor as *quaternion* data, which is simply a unit vector $x\vec{i} + y\vec{j} + z\vec{k}$ rotated by an angle α. It is represented in the following way:

$$\begin{aligned} q_w &= \cos\frac{\alpha}{2} \\ q_x &= \vec{i}x\sin\frac{\alpha}{2} \\ q_y &= \vec{j}y\sin\frac{\alpha}{2} \\ q_z &= \vec{k}z\sin\frac{\alpha}{2} \end{aligned} \tag{3.1}$$

Quaternions are popular in 3D virtual reality, robotics and navigation because they are simple to use and avoid gimbal lock[6]. To obtain the 3-2-3 Euler angles for the rotation operation (Section 2.2.3), the following relations can be used [79]:

$$\begin{aligned} \alpha &= \arctan_2(2q_yq_z + q_wq_x, 2q_w^2 + 2q_z^2 - 1) \\ \beta &= \arcsin(2q_wq_y - q_wq_z) \\ \gamma &= \arctan_2(2q_xq_y + 2q_wq_z, 2wq^2 + 2q_x^2 - 1) \end{aligned} \tag{3.2}$$

To calibrate the sensor, a quaternion reading is done during the initialisation. This quaternion rotation is then subtracted from the subsequent readings using conjugate quaternion multiplication.

3.4.4 Time-variant input variables

A challenge with block processing is that each block assumes a stationary system (in signal processing terms, a *time-invariant* system). If there are time-varying signal processing operations in the system, these must either be implemented as sample-by-sample operations, or the time-variance must be sampled at each new block input event. In the system in question, the sound field rotation angles are time-varying variables that must be considered.

The simplest approach would be to obtain the angular information once per block of input samples, resulting in a stepwise angular function as input to the HOA decoder. This may be interpreted as instantaneous movements of the sound field around the listener, again resulting in discontinuities in the binaural signal perceived as clicks. Consequently, the amount and loudness of these clicks must be evaluated to determine if they are audible.

Another approach would be to cross-fade between two successive angular positions. This could be done by e.g. multiplying the block of samples with a triangular time

[6] Gimbal lock occurs when two rotational axes are pointing in the same direction, restricting rotation of the object.

window, and adding it to the next block, which is rotated by a new angle. An overlap-add scheme must be used, which will eventually result in an increase of processing requirements. Figure 3.8 shows the cross-fading method. Possibly, the fades may be shorter as well, as indicated in the figure. Less processing requirements will then be required at the cost of sharper transitions between the rotation angles.

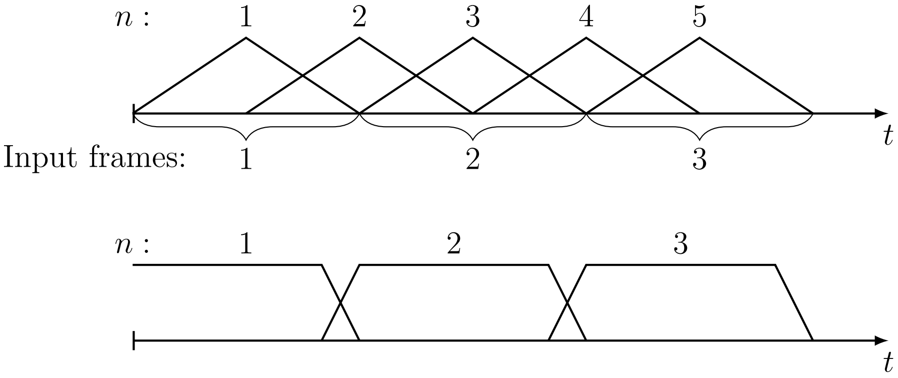


Figure 3.8: Smoothing of angular position by cross-fading. Triangular windows (top) along with a more efficient version with shorter fades (bottom). n represents frames that must be processed.

A less usable approach would be to interpolate the values between two successive rotation angles, and compute the rotation matrix sample-by-sample with the interpolated value. This requires a lot of calculation time and is considered unfeasible.

Since the effects of using the simplest approach was not particularly disturbing with slow head movements, implementing an overlap-add method was not prioritised. However, in a future version, e.g. for subjective listening tests, this issue should be resolved.

3.5 Performance measures

To have an idea of the amount of system resources the HOA system consumes, some simple performance measurements and calculations have been done. It is important to identify the system performance to:

1. Determine a suitable platform for further implementation and system development
2. Determine bottlenecks and where system optimisation is critical.
3. Investigate whether the system performs in order to satisfy psychoacoustic demands (head-tracking latency) and overall performance demands.

The CPU usage will affect power consumption (relevant for DSPs and laptop computers), and the ability to run smoothly on a personal computer. On a dual-core 2010 Macbook Pro (Intel Core 2 Duo CPU), the CPU time usage is about 35% with the current real-time code. This acceptable for prototyping purposes, but one should seek to reduce this significantly if the system needs to run simultaneously with e.g. a video feed. It is expected that modern DSPs will manage this load easily, as they are more optimised for FFT and filtering operations.

Assuming that all encoding, decoding and filtering matrices are pre-computed, only 2.5 MB of memory is required using 1024-sample frames of 32 bit single precision data. Any modern PC can handle this, but care must be chosen if the processing shall be done on a DSP chip. Such processors may only have a few kB of memory, and thus the filter length and/or buffer sizes must be adapted. At 44.1 kHz sampling rate, the data rate will be at least $32 \times 44100 \times 16 \approx 22$ Mbit/s at the audio input. Thus, a DSP or sound card must provide such a bandwidth, and a transmission system must provide 17 Mbit/s (25 HOA channels) if no compression is used.

Latency measurements of the real-time system were performed. The end-to-end latency was performed in a very simple way: The system was modified to produce an audio pulse at the headphone output when motion is sensed. Then, the motion sensor was knocked physically with a pen to produce both an acoustic impulse that was recorded with a microphone, and an electric pulse that was simultaneously recorded. The end-to-end latency was then calculated as the time difference between these two impulses. However, it is only a rough estimate of the actual latency due to the simplicity of the method. Audio latency (from the spherical microphone array to the headphone output) was measured with an external measurement system (WinMLS 2004[7]) with a loudspeaker providing a signal for the spherical microphone array. Several consequent latency measurements were performed to obtain a standard deviation.

Table 3.1 sums up the key performance numbers for the real-time system.

Property	Value	Unit
CPU load	35	%
Memory usage	2.5	MBytes
Data rate (microphone)	22	Mbit/s
Data rate (HOA signals)	17	Mbit/s
Audio I/O latency	$\mu = 92$, $\sigma = 3$	ms
End-to-end latency	$\mu = 96$, $\sigma = 14$	ms

Table 3.1: Key performance numbers for the real-time HOA system. μ and σ is the mean and standard deviation of the latency measurements.

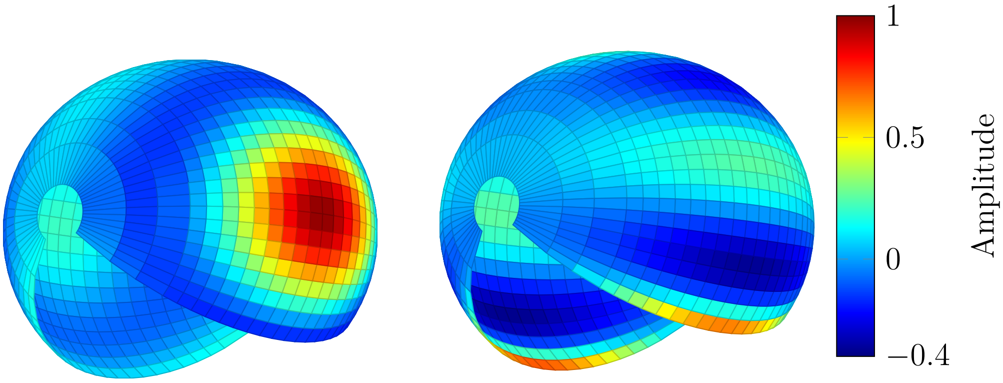
[7] http://www.winmls.com

3.6 HRTF database

The choice of HRTF database is crucial for the reproduction quality. As discussed in Section 2.4.1, individualised HRTFs are needed to provide the best results. However, in most cases measuring each person's HRTF is impractical and a database of HRTFs from different persons must be used. One must then choose the most appropriate HRTF set from the database, matching the head geometry as far as possible.

Several HRTF databases are publicly available. The CIPIC database[8] [80] contains HRTFs for 45 subjects at 1250 directions, but unfortunately sources in the lower quarter-sphere have not been measured. The LISTEN HRTF database[9] and the ARI HRTF database[10] also contains many subjects but lacks measurements from low elevation angles. MIT Media Lab[11] has measured the KEMAR dummy head response and Bernschütz [69][12] measured the Neumann KU-100 at a very high angular resolution over a full sphere of source locations.

If there is a large gap on the measurement sphere, such as lack of below-torso measurements, the condition number of the $\mathbf{Y}$-matrix will grow quickly with the representation order (see Rafaely and Avni [81]). This reflects the fact that sources placed in the gap cannot be accurately reproduced, due to the lack of spatial information. Figure 3.9 shows this, comparing a plane wave virtual source radiating from ahead and below the listener. When the source is located in the HRTF gap, the loudspeaker amplitudes are distributed on the edges of the gap, and the Gibbs' phenomena resulting from order truncation are stronger. This will most likely increase the truncation error and thus localisation blur in this area. One way to cope with this issue would be to insert imaginary loudspeakers with no output, effectively hiding sources at low elevations, or interpolate the HRTFs in a way to approximate the non-existing HRTFs.



(a) Ahead, $\theta = \pi/2$, $\phi = 0$.

(b) Below, $\theta = \pi$, $\phi = 0$.

Figure 3.9: Reproduction of a 4th order virtual source ahead and below the listener, through the CIPIC HRTF database with 1250 HRTF positions. The colours represent virtual loudspeaker amplitudes, one sphere segment for each HRTF position.

[8] http://interface.cipic.ucdavis.edu/sound/hrtf.html (09.06.2014)

[9] http://recherche.ircam.fr/equipes/salles/listen/ (09.06.2014)

[10] http://www.kfs.oeaw.ac.at/ (09.06.2014)

[11] http://sound.media.mit.edu/resources/KEMAR.html (09.06.2014)

[12] http://www.audiogroup.web.fh-koeln.de/ku100hrir.html (09.06.2014)

Figure 3.9 is also a good illustration of how signals are reproduced with HOA, as a 4th order system ”smears out” the loudspeaker energy at adjacent angular positions. Note also that the phase difference between the loudspeakers are zero, except those who run 180° out of phase.

Due to the complications mentioned above, the Neumann KU-100 database by Bernschütz was used. The measurements contain 2354 angular positions on the sphere, distributed as a Lebedev grid, well suited for HOA use. A robotic arm was used to align the dummy head, so high positioning accuracy is expected. Below 200 Hz, the HRTFs are extended with an analytical model, and excess phase components removed to obtain 128-tap HRIR filters. Also, note that the KU-100 dummy head does not include a torso, which has to be taken into account in the analysis of the system. The HRTF database is freely available under a Creative Commons CC BY-SA 3.0 license.

CHAPTER 4

Results

In this chapter, the reproduction quality of the binaural HOA system is evaluated with objective measures. The capture and encoding of HOA signals with a spherical microphone array introduces aliasing errors and noise, which will be investigated. Truncation of the spherical harmonics representation, limited by the microphone array configuration, data rate, number and distribution of HRTF measurement points, limits the reconstruction accuracy. This will again result in spectral coloration and errors in the ILD and ITD. These errors are analysed and discussed in this chapter.

Unless otherwise stated, all results are obtained with a truncation order of $\mathbf{N} = \mathbf{4}$. This is due to two reasons:

- The real-time system was designed as a 4th order system due to the limitations of the spherical microphone array.
- A 4th order system will have near perfect reconstruction below around 2 kHz. Thus, the most important ITD cues are already well preserved.

In addition, all the binaural cue results are calculated from the Neumann KU-100 HRTFs, as described in Section 3.6. For reference, the speed of sound is set to $c = 343$ m/s, and the head radius was estimated to be 9 cm.

4.1 Spherical microphone array aliasing

The aliasing effects discussed in Section 2.3.2 affects the encoding of HOA signals from a real, infinite order sound field mainly at higher frequencies. This is due to the discrete sampling on the sphere, limiting the spherical harmonics order that can be captured. Higher order spherical harmonics will then show up in the lower order harmonics as spatial aliasing components, especially at high frequencies where the high order components have a larger amplitude.

One way to visualise the aliasing problem is to create a beamformer from the microphone and consider the directivity at different frequencies. A simple beamformer can be realised as [21]

$$y(\omega, \theta_L, \phi_L) = \sum_{n=0}^{N} \sum_{m=-n}^{n} B_n^m(\omega) \Upsilon_n^m(\theta_L, \phi_L) \tag{4.1}$$

where $y(\omega, \theta_L, \phi_L)$ is the beamformer output at a look direction (θ_L, ϕ_L). Figure 4.1 shows the simulated and measured beam patterns obtained with a plane wave impinging on the Eigenmike®. The simulated pressures were obtained from 32 sensor points on a rigid sphere with a radius of 4.2 cm, thus, the effects of microphone element size are not included.

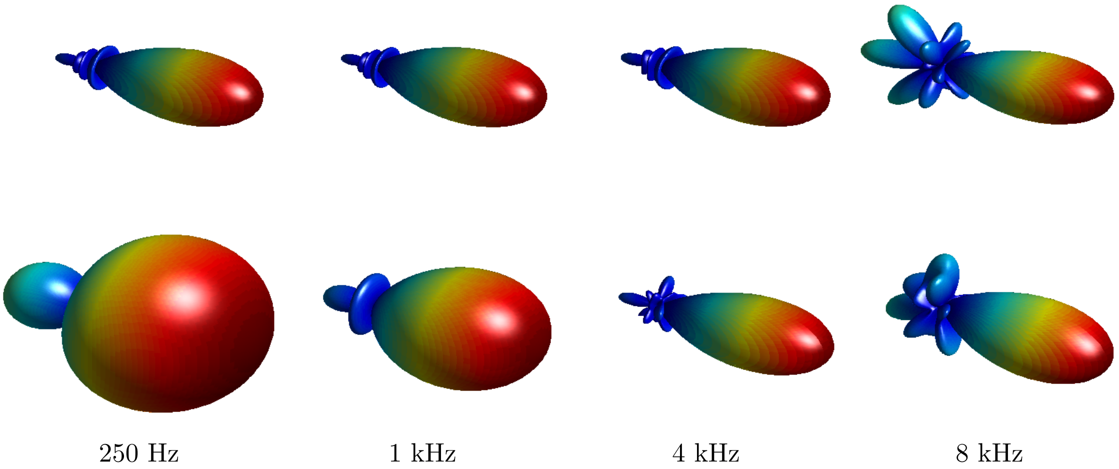


Figure 4.1: Simulated (top) and measured (bottom) beam patterns obtained with an Eigenmike® 32-capsule spherical microphone. In the measurements, the higher-order components were limited at low frequencies to suppress noise.

The plots show that the beam pattern starts to break up at around 4 kHz (measured) and 8 kHz (simulated), where aliasing phenomena occur. These are observed as large side lobes in the beam pattern. Consequently, high-frequency sources may appear to be located in more than one direction, possibly decreasing localisation accuracy, although the main lobe is still pointing in the correct direction.

Note that for low frequencies, the higher order components must be limited to suppress noise. This applies only to the measured beam patterns. Here, the amplitude of the EQ filters is limited as shown in Figure 3.7a). Consequently, the beam patterns become less directive at low frequencies. One must take care to not generalise this to a loss of localisation, since localisation is a combination of features in different frequency bands, and phase information at low frequencies.

At 8 kHz, one can actually observe somewhat less aliasing in the measured beam pattern compared to the simulated. This is most likely due to the microphone capsule size compared to wavelength, which results in a spatial smoothing of the sampling scheme. Epain and Daniel showed that using large capsules decreases the amount of aliasing at high frequencies [11].

4.2 Truncation error

As shown in Section 2.3.1, the truncation order will determine the radius of sufficient reproduction accuracy, defined by an error tolerance. However, the definition in Equation (2.45) concerns the integrated square error over the whole sphere, resulting in a single error value. It is also interesting, though, to investigate the spatial distribution of the truncation error. Figure 4.2 shows the normalised truncation error

$$\epsilon_N(kR,\theta,\phi) = \frac{|p_\infty(kR,\theta,\phi) - p_N(kR,\theta,\phi)|^2}{|p_\infty(kR,\theta,\phi)|^2} \tag{4.2}$$

where the denominator represents the mean squared pressure at the sphere surface. The pressure was calculated with Equation (2.17), where p_∞ is approximated with a high order $N = 50$. As shown in Figure 2.12, the error increases substantially for $kR/N > 0.5$, and quite large errors are observed when $kR/N \geq 1$. It is important to notice that the error is not equal everywhere, and thus the perceived error will differ according to the wave arrival direction. As can be seen in the plots, where the wave arrives from the right, the error is highest at the front and back of the sphere, and smallest at the sides. Thus, any wave arriving from the median plane will give the smallest errors at the ears, assuming that the ears are placed at each side of the sphere ($\phi = \pm\pi/2$), although this will differ from person to person. Waves arriving from the side of the head will give larger errors. The resulting differences in ITD, ILD and spectral cues, and their dependence on arrival direction, will be presented in the next section.

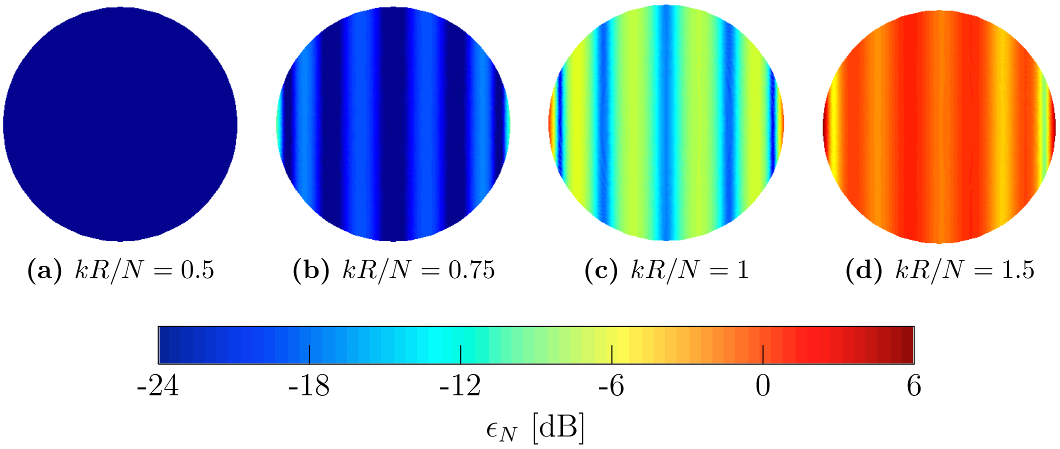

(a) $kR/N = 0.5$ **(b)** $kR/N = 0.75$ **(c)** $kR/N = 1$ **(d)** $kR/N = 1.5$

Figure 4.2: Normalised truncation error $\epsilon_N(kR,\theta,\phi)$ on the surface of a rigid sphere, for different wavenumber-radius products. The plane wave is impinging on the right side of the sphere. Truncation order $N = 4$.

4.3 Binaural representation error

In this section, the error resulting from truncating the sound field representation and reproducing it binaurally is studied. All the results in this section is the errors that arise when a plane wave is represented with Higher Order Ambisonics, and reproduced with virtual loudspeakers filtered with the corresponding HRTFs. Thus, the error can be seen as the difference between

- using the original HRTFs from a particular angle, and
- using the HOA-based Nth order truncated HRTFs, mainly with $N = 4$.

It is important to keep in mind that errors resulting from inaccurate HRTF measurements or non-individualised HRTFs are not included. Such errors will not be investigated in this thesis.

The normalised truncation error is a general quantitative measure on the reconstruction error. However, it does not give any information on the auditory system's ability to locate a source, and certainly not the perceived sound quality and spatiality of the source. As explained in Section 2.4, the ILD, ITD and spectral cues are the commonly used measures for localisation. In addition, *timbre* – the spectral coloration of a sound – is important to sound quality and possibly externalisation, distance perception and localisation [49]. The normalised truncation error does not depend on the number of loudspeakers, which, as shown later, affects the reconstruction.

ILDs and ITDs are mainly used to localise the azimuth of a source, and thus it makes the most sense to investigate errors for these measures in the horizontal plane. Spectral cues are used for determining elevation, and consequently errors in the median plane will be presented. The contour plots that follow are convenient for visualising how the error changes with frequency and source direction. Then, median error values from all incidence angles are presented, which give an impression of how much the error reduces with the new phase correction methods. Finally, sphere plots with the ILD error in octave bands show how the error depends on the source direction and correction method.

Errors affecting spectral cues and ILD are plotted in a frequency range from 0 to 15 kHz, because localisation cues based on levels can mainly be found in this range. The ITD is primarily used below 1.5 - 2 kHz, although the 4th order HOA reproduction is quite accurate up to 2 kHz, so an extended frequency range of 0 to 5 kHz is used in the plots to show some of the resulting errors.

4.3.1 Magnitude error - timbre and spectral cues

In the following, the impact of a HOA-representation on the reconstructed sound field magnitude is investigated. Both the timbre and the spectral cues will be affected by magnitude errors. Timbre is a measure on the frequency content of the reproduction, and a significant amplification or reduction at certain frequencies will cause a coloration of the sound. Spectral cues are fluctuations in the HRTF frequency response that vary with elevation angle, and HRTF magnitude errors will disrupt these cues.

The magnitude error can be calculated with

$$\epsilon_m = 20 \log_{10} \left| \frac{H_N(f)}{H_\infty(f)} \right|, \tag{4.3}$$

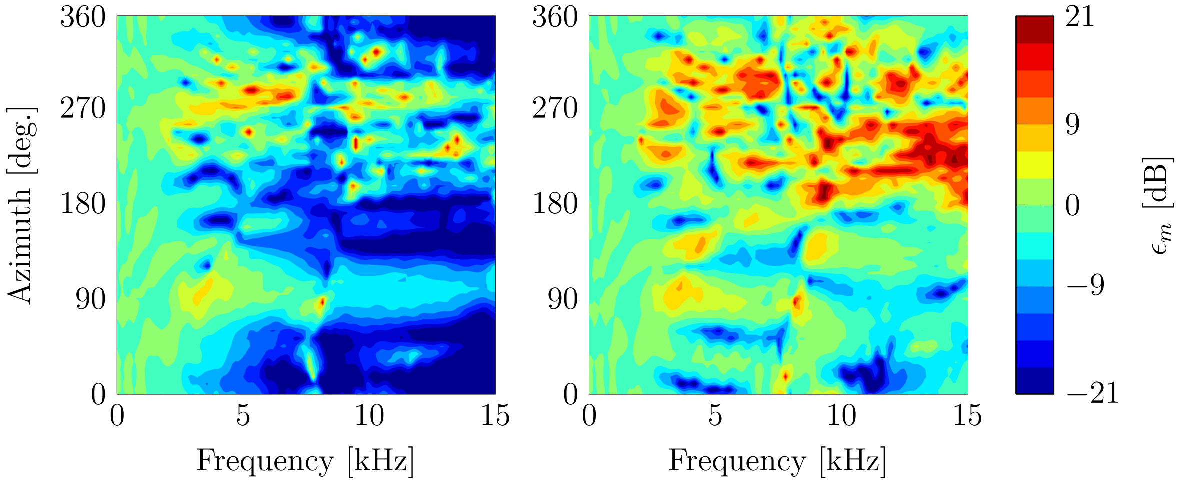
where $H_N(f)$ is the HOA-truncated HRTF, and $H_\infty(f)$ is the original HRTF.

Influence of the number of loudspeakers

As mentioned in Section 2.5, a large number of loudspeakers, or in this case, HRTF measurement positions, will result in spectral impairment at high frequencies. It is interesting to see how this number affects the magnitude response around and above $kr = N$. Two cases are investigated; one with 2354 virtual loudspeakers as in the HRTF database, and one with 32 virtual loudspeakers distributed on an icosahedron surface like the EigenMike®.

The 32 new HRTFs were calculated by using a very high order ($N = 40$) HOA representation of the original HRTFs, and synthesising plane wave sources from the 32 directions, resulting in a new set with HRTFs. These will be nearly equal to hypothetically measured HRTFs from the 32 directions, only limited by the very high order, yielding accurate ($kr = N$) reconstruction up to around 24 kHz.

Figure 4.3 shows that a very high number of HRTFs will cause a loss of high frequencies. In particular, the loss is large in front of and behind the listener. This behaviour can of course be seen when using real loudspeakers as well. With only 32 HRTFs, the loss is not that clear, and in particular the reproduced levels are too high at the hidden side of the head, indicated by large positive errors in the plot.



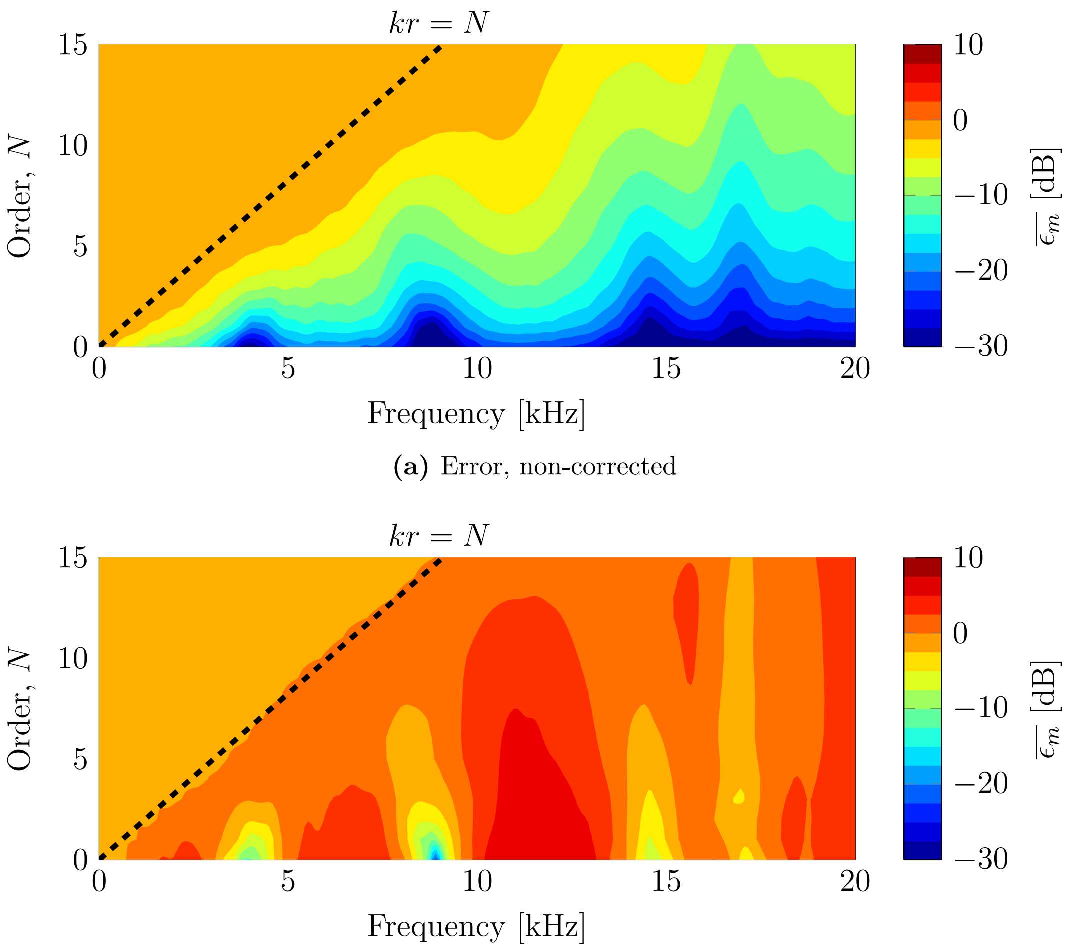
(a) $L = 2354$

(b) $L = 32$

Figure 4.3: Magnitude error ϵ_m as function of frequency and azimuth angle, for two different numbers of virtual loudspeakers, L. $N = 4$. Left ear HRTFs. The plot is limited to ±21 dB.

Dependence on N

A higher truncation order will give good reconstruction up to a higher frequency. This is investigated by averaging the energy from all directions, in practice calculated at the 2354 angles that constitute the original HRTF set, and comparing the truncated and original sound fields. This approximates diffuse field conditions, which is merely a composition of uncorrelated plane waves arriving from all directions.



(a) Error, non-corrected

(b) Error, with Villeval's timbre correction

Figure 4.4: Diffuse field energy error as function of frequency and truncation order $\overline{\epsilon_m} = 10\log_{10}\left(|p_{diff,N}|^2/|p_{diff,\infty}|^2\right)$. Left ear HRTFs. Note how the error contour lines follow $kr = N$ at small error levels. $L = 2354$ virtual loudspeakers.

Figure 4.4a) shows the average error in received energy received, when truncating the sound field to an order N. Note that in this case, the original HRTF set with 2354 positions is used. The magnitude error increases with frequency but decreases with a higher representation order. In particular, the error is small up to $kr = N$, as expected. Above this frequency, the loss of high frequencies in the HOA representation is quite clear. This is due to the absence of higher order modes that contain the high frequency content. Since a diffuse field can be decomposed into a sum of uncorrelated HOA modes, each mode contributes to a portion of energy in the sound field. Removing the higher modes will effectively reduce the total energy level at high frequencies.

This general trend could be corrected with a timbre correction filter as suggested by Villeval [49]:

$$C(f)\Big|_{N\to N_h} = \frac{\overline{p(f,R)}_{N_h}}{\overline{p(f,R)}_{N}} \tag{4.4}$$

The filter corrects the average frequency response such that it resembles a higher order N_h. $\overline{p(f,R)}_N$ is the average magnitude response over a sphere for truncation order N. However, it does only affect the total magnitude response, not the differences between the ears (e.g. ILD and ITD). In Figure 4.4b), the correction filter is applied for a desired order $N_h = 50$. The error is now much smaller, and in some areas it is even positive, i.e. above 0 dB. This is most likely due to over-estimation of the actual error in Equation (4.4), because a rigid sphere is assumed rather than a human or artificial head.

Errors affecting spectral cues and timbre

Now, we move on to study the frequency response in the median plane and how magnitude errors will affect spectral cues. In addition, the phase correction methods are included in the study from now.

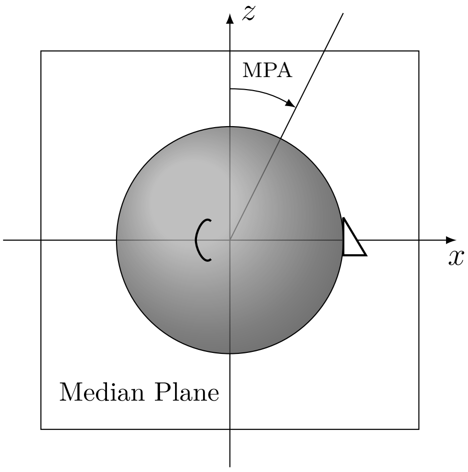
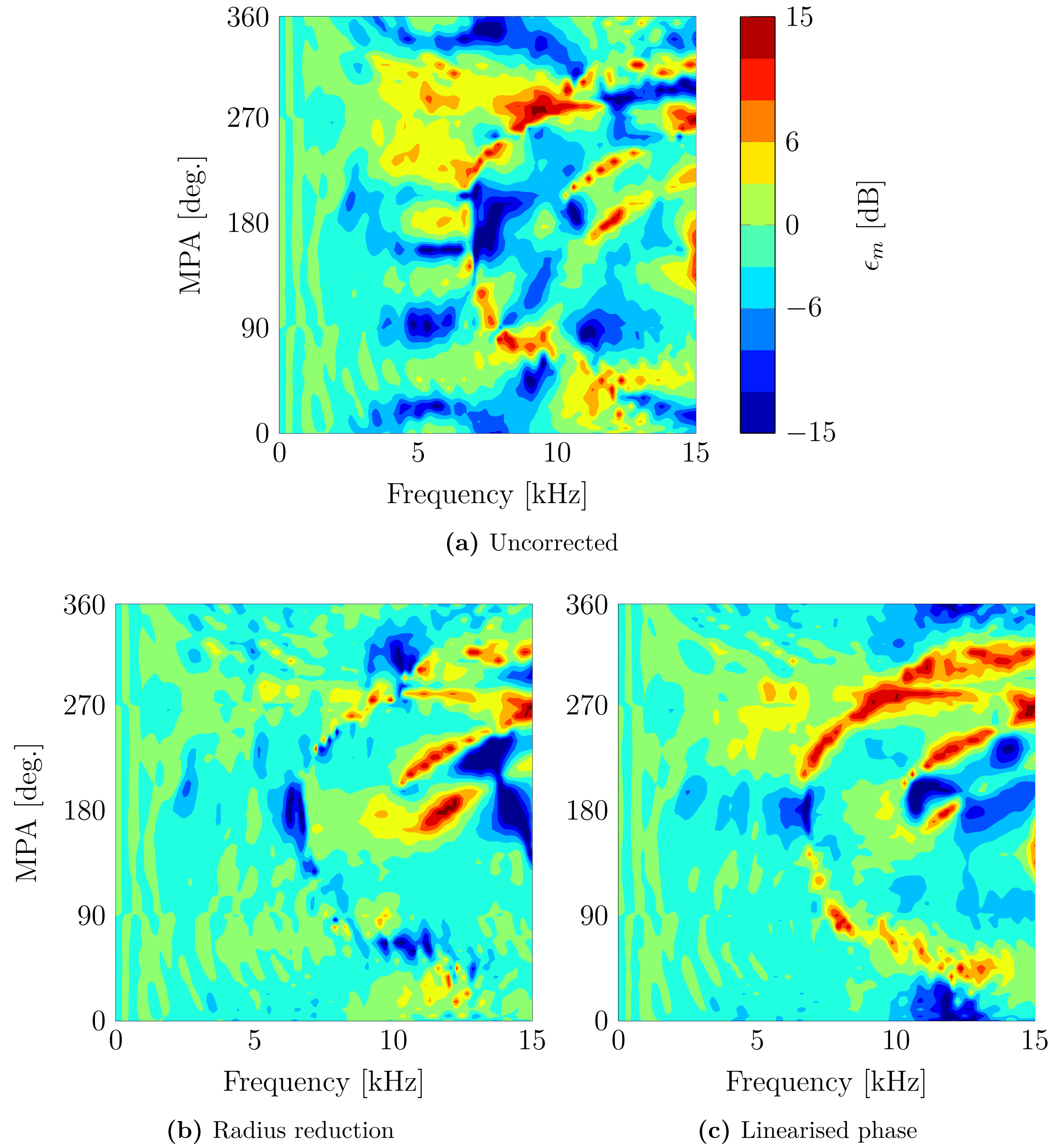


Figure 4.5: Definition of the Median Plane Angle (MPA).

The Median Plane Angle (MPA) is now defined as the angle from the positive z-axis to a point in the median plane, and with a range MPA $\in [0, 2\pi]$ radians such that MPA equals 0 rad. above, $\pi/2$ rad. in front of, π rad. below and $3\pi/2$ rad. behind the head, see Figure 4.5. Figure 4.6a) shows the magnitude error with a fixed order $N = 4$ and as function of MPA. The error is very small below approximately 2.5 kHz, as expected. Above this frequency, the errors are largely dependent on angle, which is a clear indication of that the spectral cues will be modified, increasing lateral localisation errors. Timbre is also likely to be affected, because sounds arriving from different directions in the median plane will be spectrally coloured.

There are some areas where the magnitude error is largely positive. This is when the HOA representation is unable to reproduce clear notches in the HRTFs, as shown in Figure 4.7. Thus, these areas are not extrema in the reproduced signals, but rather



(a) Uncorrected

(b) Radius reduction

(c) Linearised phase

Figure 4.6: Magnitude error ϵ_m as function of frequency and MPA, $N = 4$. Left ear HRTFs. The plot is limited to ±15 dB. $L = 32$ virtual loudspeakers.

lack of the attenuation that exists in the original HRTFs. Particularly the notch that changes systematically with MPA, resembling the letter ”C”, causes errors.

Figure 4.7 is also a good example on how the spectral cues change with elevation angle.

Following the discussion in Section 2.5.1, two methods for improving high frequency reproduction are assessed. The first method is to reduce the radius of the head above f_{lim} by considering the sphere model, and thereby changing the phase. The second method is to assume linear phase above f_{lim}, corresponding to a delay T_0. These two methods are named ”radius reduction” and ”linearised phase” in the following. In our case, $f_{lim} \approx 2.2$ kHz.

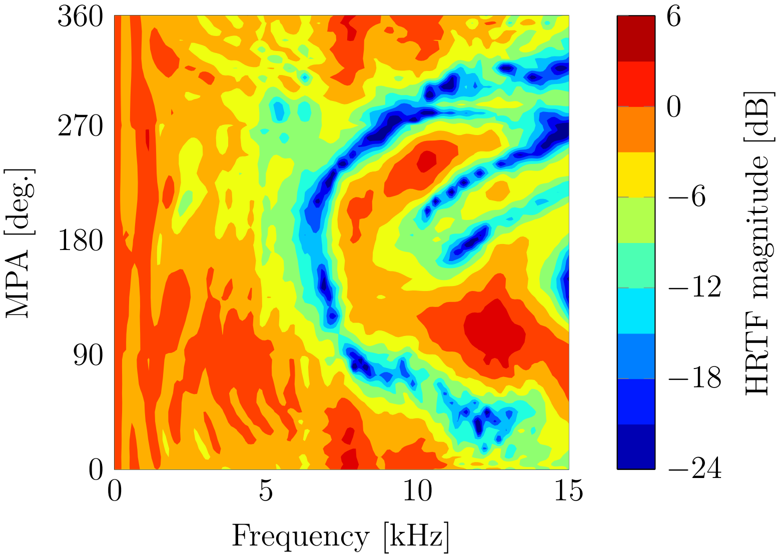


Figure 4.7: Magnitude of the original left ear HRTFs in the median plane. Notches in the original HRTFs result in positive errors when reproduced with HOA, as seen in Figure 4.6. Note the deep notch that changes systematically with MPA, that forms a "C" letter.

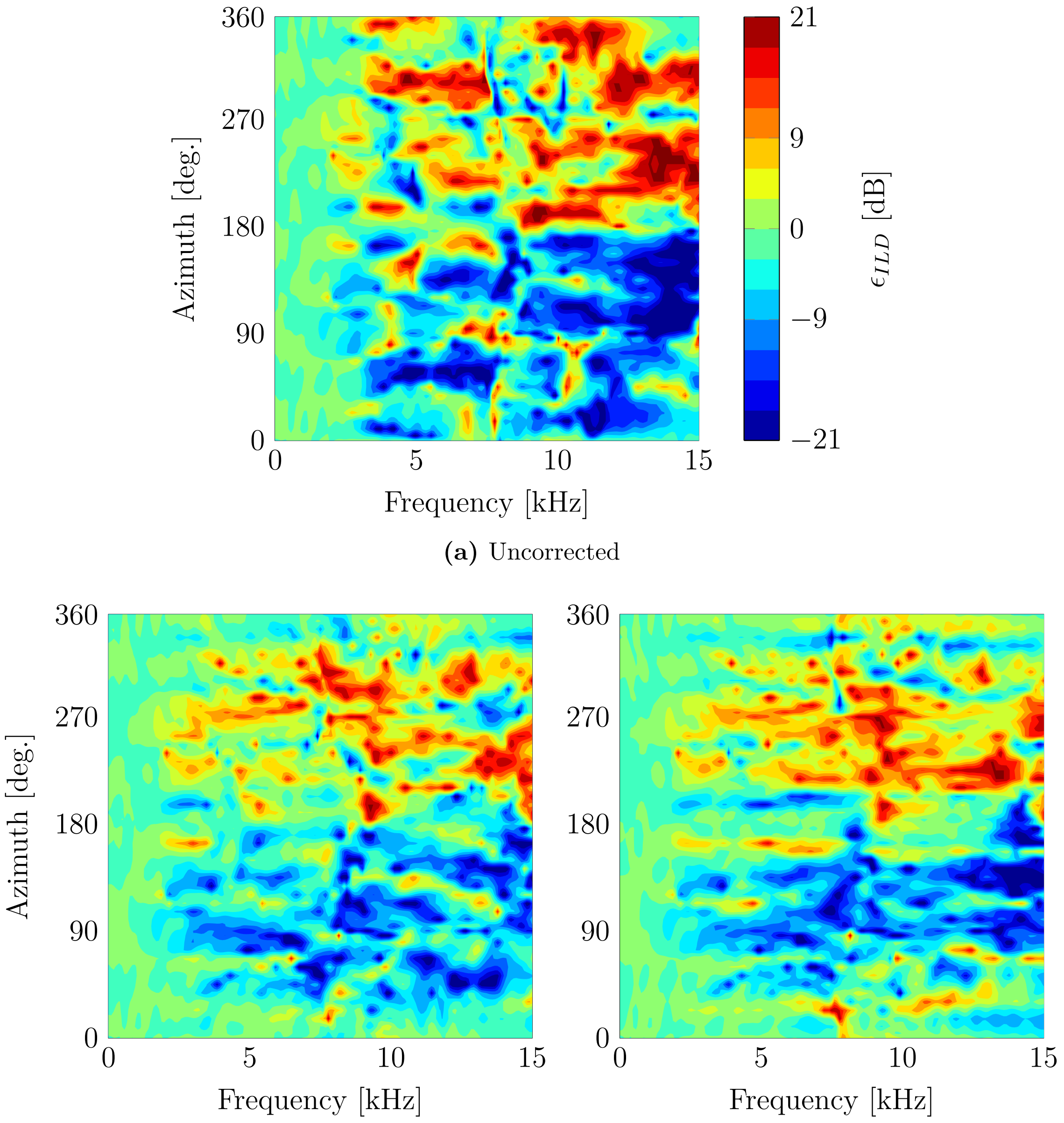
Both methods seem to improve the spectral coloration in the median plane, as seen in Figures 4.6b)-c). The frequency limit for decent reproduction is increased, and large errors are now moved to above 5 kHz. At very high frequencies, there are still large errors, but mainly due to the notches in the HRTFs, as mentioned earlier. Interestingly, the radius reduction method seems to avoid the notch-related errors quite well. Both the spectral cues and timbre will be improved in the median plane.

4.3.2 ILD error

Now, we move on to studying the error in ILD and possible improvements obtained with the phase correction methods. The ILD error can be calculated from the difference between a truncated and original HRTF,

$$\epsilon_{ILD} = 20\log_{10}\left|\frac{H_L(f)}{H_R(f)}\right|_N - 20\log_{10}\left|\frac{H_L(f)}{H_R(f)}\right|_\infty$$

where H_L and H_R are the HRTFs for the left and right ears, truncated to order N or non-truncated (subscript ∞). The ILD is used for localisation at frequencies above 1-2 kHz, so a good high frequency ILD reproduction is important. Figure 4.8a) shows that this is not the case with a reproduction order $N = 4$, above around 3 kHz. Large notches and peaks in the error plot indicate that the ILD will change a lot with source location and thus make high frequency localisation very difficult. An important observation is that, in general, the ILD is too small, resulting in an area of predominantly negative error for the left ear half-space (an azimuth between 0 and 180 degrees) and a positive area for the right ear half-space (180-360 degrees), since the ILD here is defined as the left divided by the right ear magnitude. Too small ILDs will most likely result in a sensation that the source is closer to the centre, either in front of or behind the listener.



(a) Uncorrected

(b) Radius reduction

(c) Linearised phase

Figure 4.8: ILD error ϵ_{ILD} as function of frequency and azimuth angle. The plot is limited to ±21 dB, $N = 4$. $L = 32$ virtual loudspeakers.

The main expected improvement from the phase correction approach is the more accurate ILD at high frequencies, as a consequence of improving the reproduction magnitude. Figures 4.8b)-c) show that the ILD is improved significantly with both methods, though there is still quite distinct errors when the sound source is placed at either side of the head, where the ILD is supposed to be large. It is difficult to determine whether one method is superior simply from the figure, but it is clear that the ILD is still too small when both methods are used.

4.3.3 ITD error

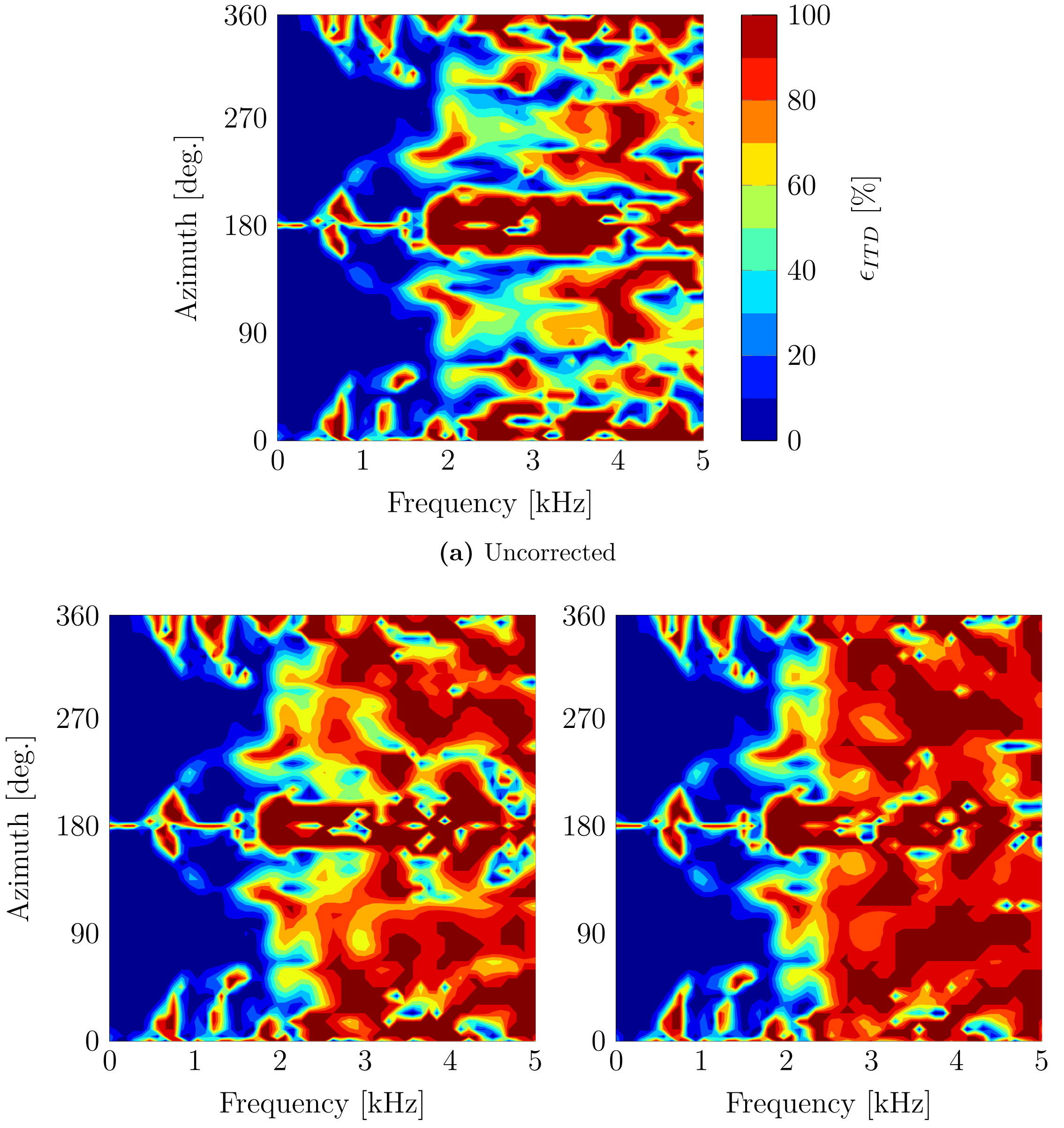
Contrary to the ILD, the ITD is used for localisation mainly at low frequencies. Thus, high frequency preservation of phase and ITD in the HRTFs is not equally important as the ILD and spectral cues. At azimuths close to 0 and π, the ITD will be very small, but localisation accuracy is at its best. Small errors at these azimuth angles will result in large localisation errors. Consequently, a relative error percentage measure is introduced to compensate for this,

$$\epsilon_{ITD} = 100 \left| \frac{ITD_N - ITD_\infty}{ITD_\infty} \right| \quad [\%], \tag{4.5}$$

where the ITD is estimated from the group delay. Note that, if the original ITD is very small, small errors become very magnified. Thus, relatively large ITD errors may appear at azimuths close to 0 and π.

Figure 4.9a) shows the ITD errors, calculated from the group delay. The HOA-truncated ITD has quite low errors below 2 kHz, except at the incident angles where the original ITD is very small. Two explanations are possible: Either, the error introduced by truncating the sound field is relatively much larger than the original ITD, or the ITD estimation procedure is inaccurate. This is quite possible, judging from Figure 2.15, but probably it is a combination of both factors. Although some large errors occur below 2 kHz, the ITD behaves well in general and low-frequency localisation is thus expected to work well with a truncation order of $N = 4$, except perhaps straight in front of and behind the head. Below 500 Hz, the ITD behaves very well for all incidence angles in the horizontal plane.

Since the phase is changed at high frequencies, the ITD is expected to be unchanged only for frequencies below f_{lim}. The ITD errors with both correction methods shown in Figure 4.9b)-c) confirm this. Above 2 kHz, the ITD totally breaks down due to the phase modifications, as expected. However, below 2 kHz the ITDs are mainly preserved with errors smaller than 10 %, and the error does not seem to behave differently than the case where phase correction is not applied. Thus, the corrections add no further errors in the ITD below 2 kHz.



(a) Uncorrected

(b) Radius reduction

(c) Linearised phase

Figure 4.9: ITD error ϵ_{ITD} as function of frequency and azimuth angle. The plot is limited to 100 %, $N = 4$. $L = 32$ virtual loudspeakers. Note the narrower frequency range than in previous plots.

4.3.4 Median error in magnitude, ILD and ITD

To get an overview over the total effect of truncating the HRTF representation, and applying the phase corrections, some percentile results are presented in the following. Percentile values are the values below which a certain percentage of observations fall. In the following, 50% percentile values are presented, meaning the value below which 50% of absolute dB errors fall. This is the same as the median value.

Although calculation of the median value is just a matter of sorting the data and picking the central value, it can be beneficial to investigate how the data is distributed. Percentiles are calculated from the empirical cumulative distribution function[1] (CDF), found with MATLAB's `ecdf`. Figure 4.10 shows an example of such a calculation.

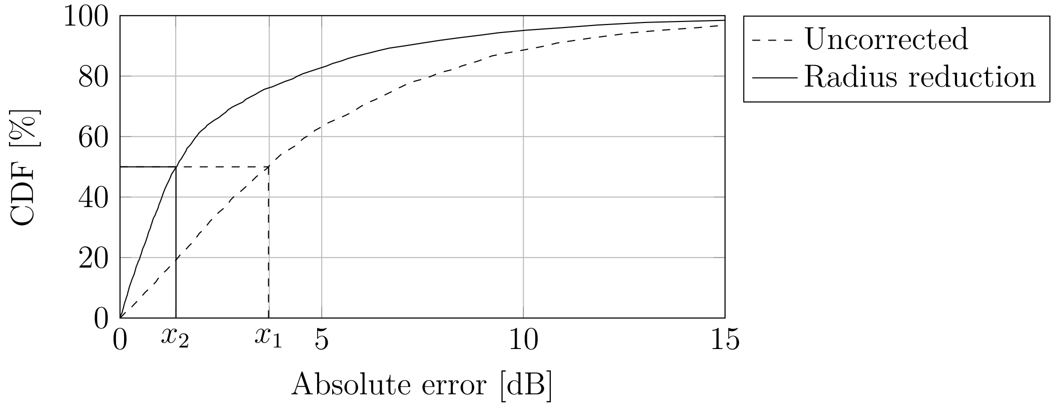


Figure 4.10: Example of a typical cumulative distribution function. The example data is absolute magnitude error in dB, from all incidence directions, at the frequency bin of 5 kHz. In this example, 50% of the absolute error values (in dB) are smaller than $x_1 = 3.8$ dB (non-corrected) and $x_2 = 1.4$ dB (radius reduction). Thus, these values can also be found in Figure 4.11 at 5 kHz.

From the CDF one can conclude that the radius reduction shifts the curve to the left, decreasing the number of errors for a certain percentile, at least at a frequency of 5 kHz. Plotting CDFs for each case and frequency is not an option as the amount of plots will be huge, so it is more convenient to present median values for each correction case as function of frequency. Note that in this section, a logarithmic frequency axis is used, to get a more perceptually realistic impression of how the errors contribute to the total picture. In the previous section, a linear frequency axis was used to show the details more clearly at high frequencies.

Figure 4.11 and 4.12 shows the median value of the absolute magnitude and ILD error in dB, respectively. The dB-errors were calculated from all incidence directions (i.e. a uniform spherical distribution). It is quite clear that for both the magnitude and ILD, the median error value is reduced. The limit frequency where the curves start to differ is around 2.5 kHz. Above this frequency, the magnitude error is reduced with close to 50% with both proposed methods, except at very high frequencies. Thus, the perceived timbre of the sound should be improved, and spectral cues are better

[1] The cumulative distribution function $f(x)$ describes the probability that an observation will be smaller or equal to x.

preserved. The median ILD error is also quite effectively reduced, and stays below 4 dB below 9 kHz. It cannot be easily determined which correction method is superior, as both have roughly the same error levels.

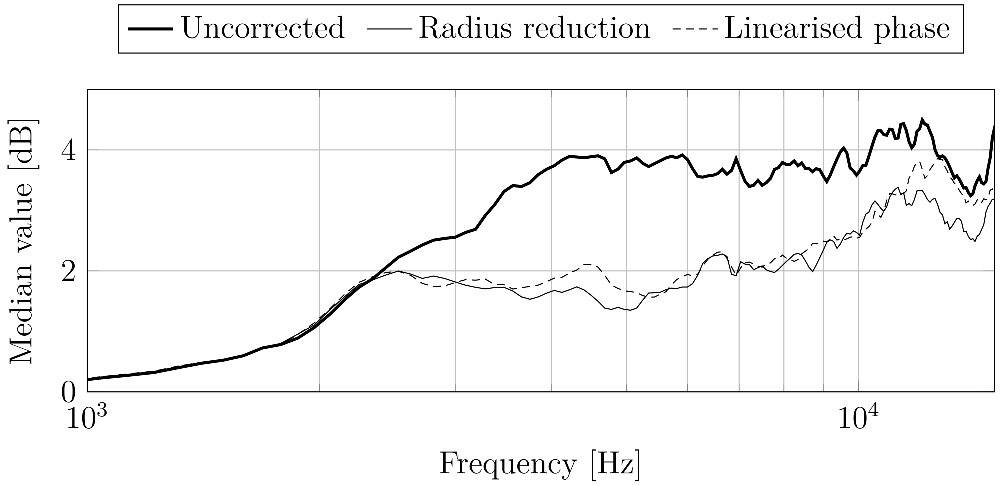


Figure 4.11: Absolute magnitude error from all incidence directions, median value. $N = 4$, $L = 32$ virtual loudspeakers.

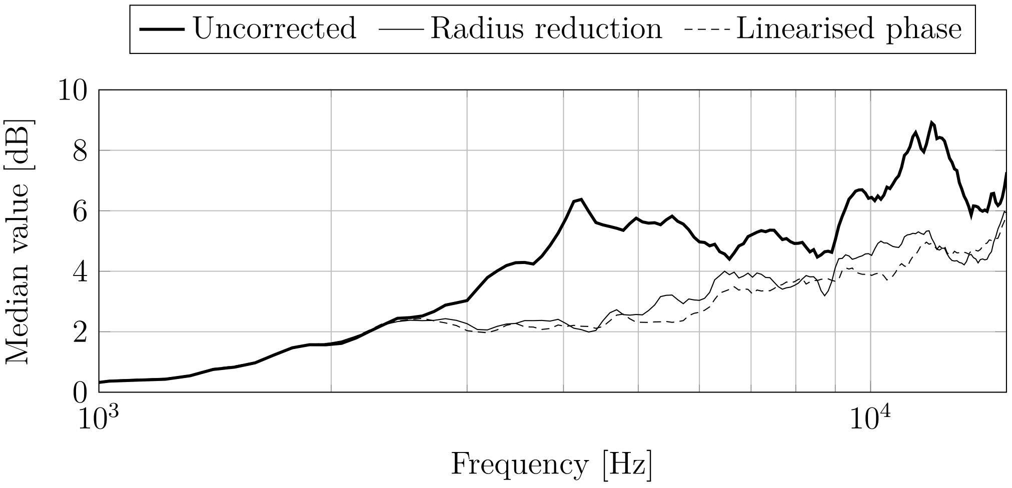


Figure 4.12: Absolute ILD error from all incidence directions, median value. $N = 4$, $L = 32$ virtual loudspeakers.

Finally, Figure 4.13 shows that the ITD error, as defined in equation (4.5) remains mostly unchanged below 2 kHz. Although the error is increased between 2 and 4 kHz, and somewhat at higher frequencies, the ITD behaves similarly in the desired frequency range. From 600 to 1200 Hz, there is a peak, resulting from the blow-up of small ITDs at azimuths angles close to 0 and π, as seen in Figure 4.9. Below 600 Hz, the median values are close to 0%.

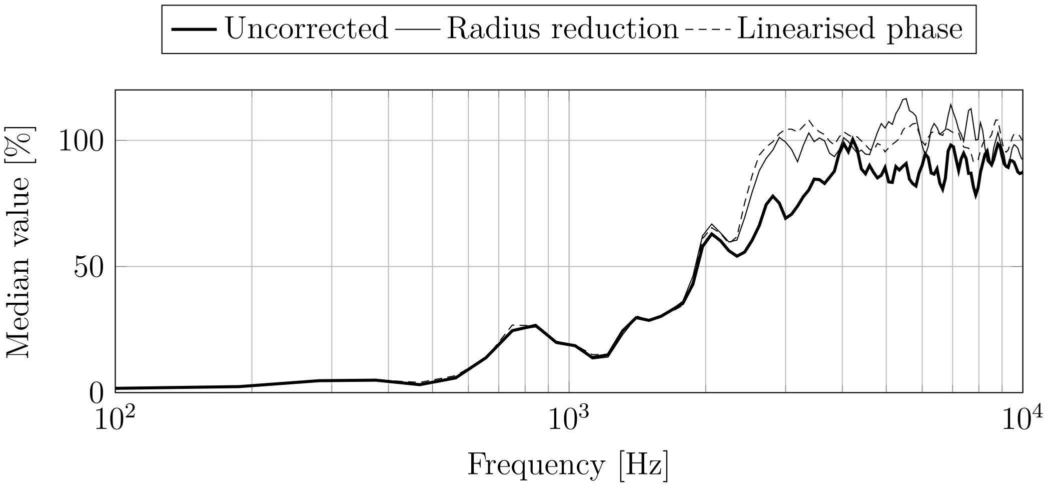


Figure 4.13: Absolute ITD error from all incidence directions, median value. $N = 4$, $L = 32$ virtual loudspeakers. Note the different frequency range compared to the previous two plots.

4.3.5 ILD error in octave bands

To get a further impression of how the proposed methods perform, the ILD error is investigated in octave bands, and from all incidence directions. Figure 4.14 shows spherical plots of the error, with a "camera" view from -30° azimuth and 60° elevation. In the 2 kHz octave band, the error is relatively small (a median value of 1.2 dB), and the correction methods have a little effect. In addition, the largest errors arise when the source is placed at the left or right side of the head. This tendency is less prominent at higher frequencies, but there is a tendency that the error is small for sources located in the median plane. In the 4-16 kHz octave bands, the error increases with frequency, but the increase of the median value is about 50% less when the correction methods are applied. The error is also quite unevenly distributed, which may result in localisation confusion when the head or source is moving. On the other hand, the listener may also move the head towards the source, often resulting in smaller errors judging by the sphere plots.

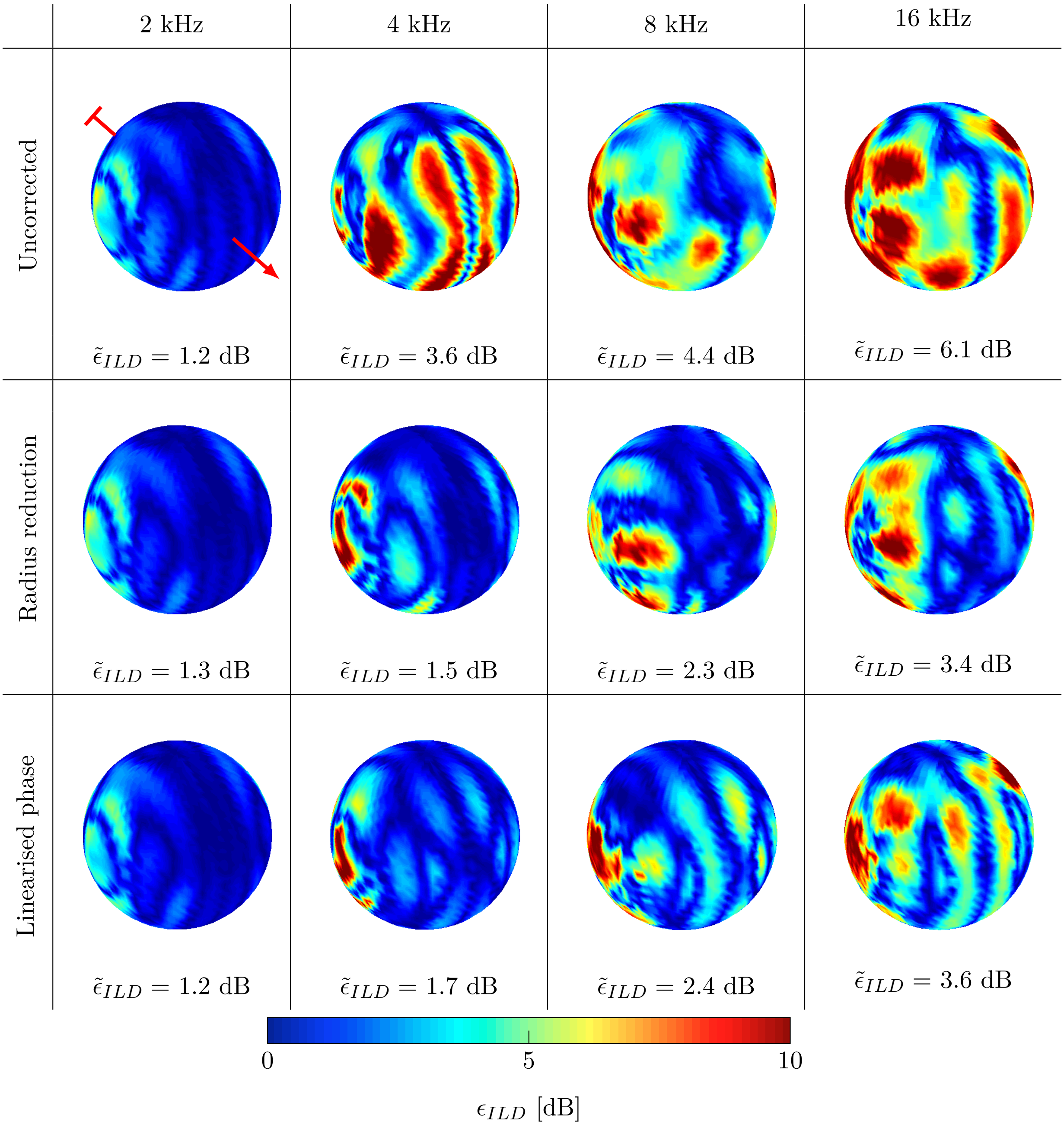


Figure 4.14: Absolute ILD error, in octave bands (kHz), from all incidence directions (spherical distribution). The red arrow indicates the x-axis (nose direction). $\tilde{\epsilon}_{ILD}$ is the median of the absolute ILD error over the sphere. $N = 4$, $L = 32$ loudspeakers. The plots are limited to maximum 10 dB.

CHAPTER 5

Discussion

This chapter will go through and discuss the work presented in this thesis. First, some general comments on the binaural HOA system and the possible improvements are included. Then, the real-time implementation and simulation results are reviewed, along with the implications of the phase corrections. Finally, some suggestions for future work are presented.

A main purpose of the work was to implement a working real-time system. When this was accomplished, most of the work went into the second purpose, to study the reproduction accuracy and possible improvements. Thus, this last part receives the most attention in this chapter.

5.1 General comments on Binaural HOA

Higher Order Ambisonics is a very convenient audio format for 3D audio. The independence from signal acquisition and reproduction is a clear advantage over more common systems such as consumer surround sound (e.g. 5.1 systems). Also, it can easily be rotated around the listener without changing loudspeaker positions, which is a requirement in many applications. It is also scalable, meaning that only a few, or even just one channel can be transmitted if necessary. Thus, it fits right into the world of streaming media, where service providers normally operate with scalable video and audio formats for different customer bandwidths.

The main drawbacks with HOA are the number of required channels and the limited ”sweet spot” of decent reproduction. A high number of channels can be avoided by down-scaling (truncating) when necessary, but this shrinks the sweet spot and consequently decreases high-frequency performance. In a loudspeaker reproduction system, the sweet spot, limited by $kr = N$, is an issue when multiple listeners are present or a single listener wants to move. Thus, the enjoyment of high-quality spatial audio with HOA is highly dependent on many loudspeakers and restrictions on movement.

Using headphones instead of loudspeakers elegantly solves the issues mentioned above. The number of loudspeakers is now only limited by the spatial resolution of the HRTF database. Multiple listeners and freedom of movement is also possible. Reproduction accuracy is now limited by the source material, e.g. the HOA-signals, and to what degree the HRTF database is a good fit to the listener. However, a low-order truncated signal may still be adequate if the phase correction methods investigated in this study are applied. As a consequence, one can use spherical microphone arrays with a relatively low number of sensors and still obtain good high frequency reproduction.

Even though the transition from loudspeakers to binaural reproduction solves some fundamental issues, there are still certain challenges due to the limitations with spherical microphone arrays. Such arrays suffer from noise at low frequencies, due to the usually small sphere radius, and aliasing at high frequencies, limited by the number of sensors. Optimisation and improvements of spherical microphone arrays was not the main focus in this study, as there is extensive research on the topic in the literature. It cannot be neglected though, that one will never have a better reproduction than the source material allows. Therefore, further development on such arrays, with low signal-to-noise ratios and minimal aliasing problems, is important if the goal is to listen to recorded sound fields.

It is widely recognised that individual HRTF sets are required to achieve a very realistic reproduction, especially when it comes to externalisation, i.e. perceiving the sound source outside the head. Several approaches are possible to face this challenge, one may accept that non-individual HRTFs must be used, one may try to find the most suitable available set, or one may seek to use individual HRTFs. If the last option is chosen, there are two options, acoustic measurements or 3D scanning. Acoustic measurements is the traditional method, and is time-consuming and often expensive. 3D scanning with consequent Boundary Element Method calculation of the HRTFs has become more popular during the last decade [67], and may be an efficient way to individualise HRTFs in the future.

Another challenge with HRTFs is the dynamic behaviour when one rotates the head. Since the shoulders and torso is not moving, reflection patterns will differ and consequently the HRTFs will change. One possible solution is to model such reflections with simple geometric models.

Figure 5.1 shows a hypothetical sketch of how the perceived quality of the system may be, that is, how close to reality it is experienced. The best achievable quality with headphones is obtained with individual HRTFs and a high truncation order. Non-individual HRTFs limit the maximum achievable quality, even with a high truncation order. Using an order-limited source such as a spherical microphone array, adds noise and aliasing, so the maximum achievable quality is further reduced, in addition to a constraint on the reproduction order. The red arrow shows how the HRTF phase correction could improve the quality at low truncation orders, although there is always an upper quality limit, indicated by horizontal lines in the figure. Also, loudspeaker reproduction may be the best alternative quality-wise (indicated by the dotted line at the top), but at the cost of very many loudspeakers.

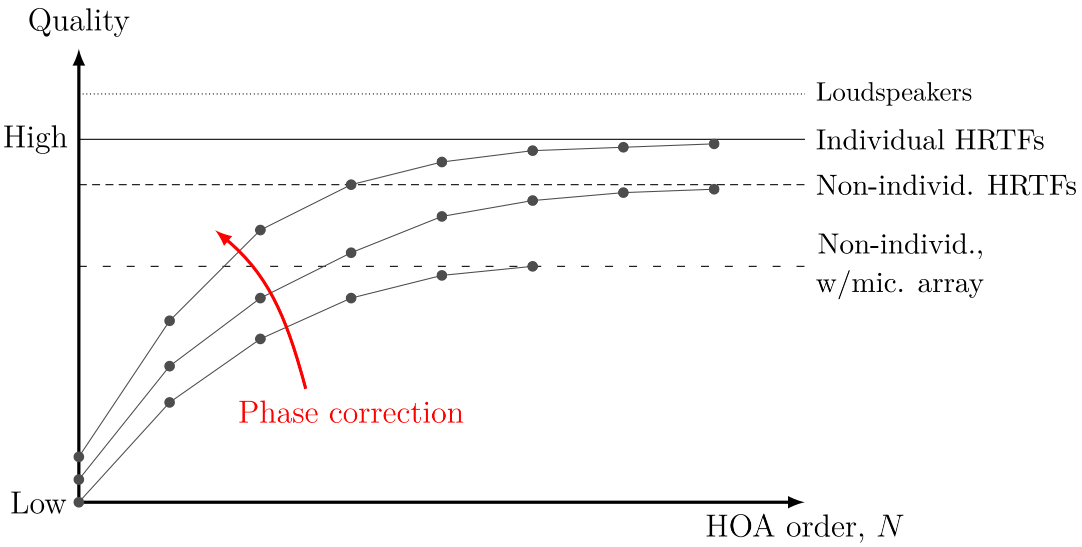


Figure 5.1: Conceptual sketch of how the perceived quality increases with truncation order. The horizontal lines indicate maximum achievable quality depending on whether individual HRTFs are used and the source material. Possible improvements by the phase correction method is indicated with the red arrow.

5.2 Real-time implementation

An important part of the thesis work was the successful implementation of a real-time system to perform the sound field recording and binaural reproduction simultaneously. Even though the developed system was not extensively used and evaluated for practical purposes, it is an important component for future research and development. A central part of further evaluation of the system is listening tests, which will need a working real-time implementation.

Since it is very easy to acquire the audio signals from the Eigenmike® to a personal computer, a lot of work was saved. Programming an interface between the FSM-9 motion sensor and MATLAB was also a relatively small task, due to the existence of Psychophysics Toolbox. There is however a large step from this type of implementation to a finished product. For all practical purposes, the system should be independent of MATLAB, and preferably implemented on a low-cost DSP instead of a PC. This may require a different microphone array, as most DSPs do not support FireWire. DSP implementation is also important to minimise the latency, power draw and risk of dropouts. It is also important to take these steps to make the system easy to operate.

Initial, simple measurements showed that the end-to-end latency between sensor motion and sound field rotation in the headphones was around 95 ms. Brungart et al. [82] suggest that most listeners are not able to detect latencies smaller than 60 ms, based on localisation experiments. Consequently, the system performs just above the borderline of what would be acceptable in a virtual environment. It must be noted, though, that more accurate latency measurements should be performed, due to the simplicity of the performed measurement.

5.3 Quantitative results

This section discusses the main findings in Chapter 4, which are quantitative (numeric) results. However, the reader should keep in mind that it is the qualitative improvement we perceive, so it is important to approach the numeric results, and in particular visually convincing figures, with a critical mind.

The analysis of the spherical microphone array performance in Section 4.1 confirms previous studies (see e.g. [10, 11]), and the simulated and measured beam patterns are quite similar, except at very low frequencies where the noise becomes an issue. This issue is actually not that problematic, because decent reproduction can still be achieved at low frequencies, as kr is low and high order modes are not needed (Fig. 3.6). In fact, the beam patterns may not give a complete impression on how the array performs in the binaural HOA context – as phase (ITD) information is equally relevant for localisation.

It was also shown that the normalised truncation error as previously defined by Ward and Abhayapala [14] needs to be expanded to account for scattering from the head, and that the spatial distribution of the error is quite uneven. This leads to the need for new ways to analyse the truncation error, preferably with psychoacoustic measures such as ILD, ITD and spectral cues. Evaluating HOA with such measures has become more popular in the recent years, and was further studied in this thesis.

At the base of the study on the binaural representation is the HRTF database. The comparisons are made between the original HRTFs and the HOA-reconstructed HRTFs. Thus, the effects of using non-individual HRTFs will not show in the results, although the results will depend on which database is used because head geometry is an important factor. For example, listeners with small heads will naturally experience better reproduction than listeners with large heads, because the $kr = N$ rule of thumb yields a higher frequency limit.

Figure 4.3 shows that the number of loudspeakers in a HOA system, and consequently the number of HRTFs, is very important at frequencies above the frequency limit of 2.2 kHz (assuming a 4th order system). This was pointed out by Solvang [43], but has received little attention elsewhere in the literature. Most authors seem to use as few loudspeakers as possible in loudspeaker systems, but as many HRTFs as possible in headphone systems. The reason for the rather large differences is related to aliasing and modal truncation. As shown in Section 2.5, the virtual loudspeaker approach is identical to estimating the spherical harmonic spectrum of the HRTF set. By truncating the spherical harmonic order, a loss of high frequencies is expected. This is observed in Figure 4.3a) where the spatial sampling is very dense, which suppresses aliasing. However, when the sampling is sparse, the aliasing from higher order modes that are present in the HRTF set will clearly raise the high frequency levels. This actually helps to reduce the magnitude error. Thus, we get a positive spatial aliasing effect.

Also, note that the correction filter in Equation (4.4) is only applicable with a dense HRTF sampling, such as in Figure 4.4. If used on a HRTF set with e.g. 32 measurement points, the filter will over-compensate, and the levels will be too high due to the spatial aliasing.

5.3.1 Binaural cues

The main focus in the evaluation was how HOA can reproduce binaural cues and how the phase correction methods possibly can improve these cues, and the main findings in this study are the results in Section 4.3. Previous work has shown that HOA works well up to a certain frequency limit, so the binaural cue errors (particularly the ITD errors) should be small below this frequency. This was confirmed, except for the ITD at azimuths close to 0 or π. Here, large errors occur because the original ITD is small, and the relative error is thus blown up. Most likely, this is a weakness of the analysis, as there exist more perceptually accurate models than narrowband ITD estimation. Such models were not addressed because the main focus was on improvement of the ILD.

In the median plane, errors in the reproduced magnitude will disturb spectral cues that are essential for determining elevation and front/back separation. Figure 4.6 shows that these cues are significantly better reproduced at high frequencies with both phase correction methods. In particular, the radius reduction method performs well. Both phase corrected and uncorrected HOA reproduction seems to have problems where the original HRTFs have deep notches, although the radius reduction copes quite well. The systematic change of notch position with angle in Figure 4.7, particularly the “C”-notch, is probably important for elevation cues. As a consequence, the radius reduction method is the preferred method to improve spectral cues.

In the horizontal plane, the ILD and ITD is used for localisation. Figure 4.8 shows that the high frequency ILD error decreases with both methods, which seem to perform equally well. This should improve localisation, although it may be that the errors are still too large to make a significant difference. Errors greater than ±9 dB are still observed above 2 kHz, and the tendency is still too small ILDs, possibly rotating the perceived source towards the median plane. Figure 4.9 also shows that ITD error above 2 kHz is somewhat smaller without phase correction. Thus, there may be some loss in envelope-related ITD cues at high frequencies when the phase corrections are applied.

When considering the median of the absolute error values from all directions, the magnitude reproduction is improved considerably above the 2.2 kHz, as shown in Figure 4.11. This confirms that the reproduction will improve in general, for the timbre and spectral cues. The same behaviour is observed for the ILD, though the effect is smaller at frequencies 6-9 kHz. An important observation for both magnitude and ILD is that the median values seem to increase more linearly with frequency when the correction methods are applied. Also, the median value plots do not go in favour of either of the correction methods.

Finally, the spherical plots in Figure 4.14 give an impression on how the ILD error is distributed on the sphere. A general trend is that the largest errors occur when the ILD is largest. Thus, sources at the left or right side of the head may be perceived to be closer to the front or back of the head, and moving sources may seem to traverse this region faster than intended, although this must be confirmed by localisation experiments.

There is little doubt that the simulation results show a better performance in binaural level cues with phase correction. A further analysis should include a more perceptually correct binaural model, taking the auditory system’s critical bands into account, but such models are very likely to prove the same tendencies. Different HRTF sets should also be analysed – the Neumann KU-100 dummy head does not include the torso, which also contributes to localisation. However, since no other full-sphere HRTF sets including torso was found, this could not be investigated in this study.

5.3.2 Further implications of the corrections

Since the proposed corrections do not improve the ITD – high frequency cues are rather impaired – it is important to use a truncation order high enough to maintain the low-frequency ITD cues. An order of $N = 3$ or higher will ensure this, preserving the ITD up to about 1.7 kHz. The number of HRTFs must also be chosen wisely, to balance loss of high frequencies versus spatial aliasing. Phase correction or timbre correction may counteract the high frequency loss, but might not provide the same improvements as with a sparsely sampled set due to the positive aliasing effect. This should be investigated more before making a final choice.

To what degree the HRTF database needs to be individualised will also depend on how well individual details are actually reproduced. For example, if a low-order system without phase correction is used, much of the high frequency information in the HRTFs is useless. One can to a certain degree restrict to measuring or modelling low-frequency behaviour. Then, if phase correction is applied, HRTF magnitude accuracy at high frequencies may play a more important role.

5.3.3 Subjective assessment

The next important step in the evaluation process should be subjective assessment of the system with listening experiments. Here, localisation, externalisation and audio quality should be investigated. Preferably, individualised HRTFs should be used to obtain maximum effect of the correction methods. A low-latency real-time implementation is also needed, but it is assumed that the real-time system designed and presented this thesis will be sufficient if all the parameters are optimised to achieve low latency.

5.4 Sources of error

The main sources of error in this study are errors propagating from the underlying data material, measurement errors and calculation errors.

There is always a risk for having errors in data provided by others, and this cannot be avoided in most cases. Only the HRTF database and the Eigenmike® microphone positions are used as external data in this study. Thus, the risk of such errors is quite low because the data is publicly available.

Measurement errors only apply to the directivity plots in Figure 4.1. Such errors include noise and microphone calibration. However, the measurements were conducted in a low-noise anechoic environment and the microphone is factory calibrated, and no indication of large errors was found in the results.

Since much of the results are based on simulations, there is always a risk of calculation errors, both in the process of converting theory to program code, and pure programming errors. The possibility of such errors was continuously assessed by comparing the results with theoretical considerations (i.e. good performance for $kr < N$), and practical considerations (i.e. error magnitude).

It is especially important to be critical to the results since they cannot be compared with measurements at this point, and should be confirmed by subjective assessment.

5.5 Suggestions for future work

Several aspects of the work in this thesis need to be confirmed with more experiments, and some parts need to be developed further, in particular the real-time implementation. Therefore, some suggestions for future work are provided in the following.

- The effect of the phase correction methods needs to be confirmed by listening experiments. Localisation experiments will be crucial to confirm the improvement in ILD. Externalisation and audio quality is also important, as there may be unwanted side effects of the corrections.

- The real-time system must be developed to reduce latency, computational cost and possibly hardware requirements. It is natural to continue with a C implementation, and finally use a DSP chip to perform the processing.

- If a spherical microphone array is still to be used as a source, it is important to quantify how the suggested corrections work in conjunction with the array performance. Sparsely sampled arrays with much aliasing may cause the corrections to be useless, because high-frequency behaviour is already deteriorated. A study should include the whole chain from microphone to headphones.

- The suggested correction methods could be compared to traditional methods, such as Max-r_E and In-Phase [3].

- There may be alternative methods to do the phase corrections, e.g. in the spherical harmonics domain. Future research should seek to develop such methods, for example with the *a priori* knowledge about the HRTFs and ILD truncation error. This will most likely require an optimisation procedure.

- Even though few HRTF positions seem to yield better results in this study, this was due to positive aliasing effects. How this is perceived, is not known, and thus the sampling density of the HRTF set should be further investigated.

- In addition, different HRTF sets should be investigated, particularly with focus on whether HRTFs with a torso yield the same improvements, and whether individual differences are more important than the errors seen in this study.

- An interesting approach would be to compare the original and phase corrected HRTFs with a listening experiment, without using HOA. Thus, one can determine whether the high frequency phase information in the HRTFs are significant for localisation.

CHAPTER 6

Concluding remarks

This thesis has concerned several aspects of Higher Order Ambisonics, with most attention to real-time implementation and binaural reproduction. Those were the main objectives, but other parts such as spherical microphone array processing and motion handling had to be included to finalise the real-time system.

A thorough review of the theoretical framework was provided to aid the implementation, both for the author and the reader. The theory also serves as reference for future work to be done by SINTEF ICT or university students. To further ease the future work with real-time implementation, a description of the current implementation was provided, with comments and suggestions for improvements. It was clearly established that the current system, developed in MATLAB, is not suitable for a final product, and must be optimised with respect to latency if listening tests shall be performed.

Evaluation of the system focused on the microphone array performance, and the quality of binaural reproduction. Two novel methods for improving the binaural reproduction have been investigated. They use psychoacoustic models to optimise the HRTFs, and exploit the auditory system's insensitivity to phase at high frequencies.

Simulations showed that the binaural reproduction fails above the frequency limit defined by $kr = N$, but both phase correction methods will yield an improved reproduction above this frequency. In particular, the spectral cues and ILD is improved. It was also shown that the number of HRTFs used in the HOA-decoding has a large impact on the reconstruction accuracy. Simulations and measurements confirm the microphone performance compared to theory and previous studies.

These results make way for higher quality reproduction with binaural HOA, or possibly lower truncation orders, although lower orders than $N = 3$ will also impair the ITD. Such impairments cannot be avoided with the proposed methods, so Also, listening tests should be performed to confirm or disprove the findings, and whether other audible artefacts will occur.

Conclusively, the study has shown that binaural HOA is a convenient 3D sound format, with certain limitations that were addressed and possibly improved. The future will show whether this technology is viable, most likely depending on advancements in HRTF measurement and modelling.

Bibliography

[1] A. J. Berkhout, D. de Vries, and P. Vogel, "Acoustic control by wave field synthesis," *The Journal of the Acoustical Society of America*, vol. 93, no. 5, pp. 2764–2778, 1993.

[2] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

[3] J. Daniel, J.-B. Rault, and J.-D. Polack, "Ambisonics encoding of other audio formats for multiple listening conditions," in *AES 105th Convention*, Audio Engineering Society, 1998.

[4] M. Kleiner, B.-I. Dalenbäck, and P. Svensson, "Auralization - an overview," *Journal of the Audio Engineering Society*, vol. 41, no. 11, pp. 861–875, 1993.

[5] M. A. Gerzon, "Periphony: Width-height sound reproduction," *Journal of the Audio Engineering Society*, vol. 21, no. 1, pp. 2–10, 1973.

[6] M. A. Gerzon, "Ambisonics in multichannel broadcasting and video," *Journal of the Audio Engineering Society*, vol. 33, no. 11, pp. 859–871, 1985.

[7] J. Daniel, *Représentation de champs acoustiques, application à la transmission et à la reproduction de scènes sonores complexes dans un contexte multimédia.* PhD thesis, University of Paris VI, France, 2000.

[8] J. Daniel, S. Moreau, and R. Nicol, "Further investigations of high-order Ambisonics and wavefield synthesis for holophonic sound imaging," in *AES 114th Convention*, Audio Engineering Society, 2003.

[9] S. Moreau and J. Daniel, "Study of higher order ambisonic microphone," in *7ème Congrès Français d'Acoustique (Joint congress CFA-DAGA'04)*, 2004.

[10] S. Bertet, J. Daniel, and S. Moreau, "3D sound field recording with higher order ambisonics-objective measurements and validation of spherical microphone," in *AES 120th Convention*, Audio Engineering Society, 2006.

[11] J. Daniel and N. Epain, "Improving spherical microphone arrays," in *AES 124th Convention*, Audio Engineering Society, 2008.

[12] J. Daniel, "Spatial sound encoding including near field effect: Introducing distance coding filters and a viable, new Ambisonic format," in *AES 23rd International Conference on Signal Processing in Audio Recording and Reproduction*, Audio Engineering Society, 2003.

[13] J. Daniel and S. Moreau, "Further study of sound field coding with higher order Ambisonics," in *AES 116th Convention*, Audio Engineering Society, 2004.

[14] D. B. Ward and T. D. Abhayapala, "Reproduction of a plane-wave sound field using an array of loudspeakers," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 6, pp. 697–707, 2001.

[15] S. Spors and J. Ahrens, "A comparison of wave field synthesis and higher-order Ambisonics with respect to physical properties and spatial sampling," in *AES 125th Convention*, Audio Engineering Society, 2008.

[16] J. Ahrens and S. Spors, "An analytical approach to sound field reproduction using circular and spherical loudspeaker distributions," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 988–999, 2008.

[17] S. Spors, V. Kuscher, and J. Ahrens, "Efficient realization of model-based rendering for 2.5-dimensional near-field compensated higher order Ambisonics," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 61–64, IEEE, 2011.

[18] J. Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2002 (ICASSP'02)*, vol. 2, pp. II–1781, IEEE, 2002.

[19] B. Rafaely, "Plane-wave decomposition of the sound field on a sphere by spherical convolution," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2149–2157, 2004.

[20] B. Rafaely, "Analysis and design of spherical microphone arrays," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 135–143, 2005.

[21] M. Park and B. Rafaely, "Sound-field analysis by plane-wave decomposition using spherical microphone array," *The Journal of the Acoustical Society of America*, vol. 118, no. 5, pp. 3094–3103, 2005.

[22] B. Rafaely, B. Weiss, and E. Bachmat, "Spatial aliasing in spherical microphone arrays," *IEEE Transactions on Signal Processing*, vol. 55, no. 3, pp. 1003–1010, 2007.

[23] B. Rafaely and M. Kleider, "Spherical microphone array beam steering using Wigner-D weighting," *IEEE Signal Processing Letters*, vol. 15, pp. 417–420, 2008.

[24] B. Rafaely, Y. Peled, M. Agmon, D. Khaykin, and E. Fisher, "Spherical microphone array beamforming," in *Speech Processing in Modern Communication*, pp. 281–305, Springer, 2010.

[25] E. Fisher and B. Rafaely, "Near-field spherical microphone array processing with radial filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 256–265, 2011.

[26] H. Sun, *Optimal modal signal processing using spherical microphone arrays*. PhD thesis, Norwegian University of Science and Technology (NTNU), Norway, 2011.

[27] Z. Li, R. Duraiswami, E. Grassi, and L. S. Davis, "Flexible layout and optimal cancellation of the orthonormality error for spherical microphone arrays," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004 (ICASSP'04)*, vol. 4, pp. iv–41, IEEE, 2004.

[28] R. Duraiswami, Z. Li, D. N. Zotkin, E. Grassi, and N. A. Gumerov, "Plane-wave decomposition analysis for spherical microphone arrays," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 150–153, IEEE, 2005.

[29] Z. Li and R. Duraiswami, "A robust and self-reconfigurable design of spherical microphone array for multi-resolution beamforming," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005 (ICASSP'05)*, vol. 4, pp. iv–1137, IEEE, 2005.

[30] Z. Li and R. Duraiswami, "Flexible and optimal design of spherical microphone arrays for beamforming," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 702–714, 2007.

[31] Z. Li and R. Ruraiswami, "Hemispherical microphone arrays for sound capture and beamforming," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 106–109, IEEE, 2005.

[32] M. A. Poletti, "Three-dimensional surround sound systems based on spherical harmonics," *Journal of the Audio Engineering Society*, vol. 53, no. 11, pp. 1004–1025, 2005.

[33] H. Teutsch, *Modal array signal processing: Principles and applications of acoustic wavefield decomposition*, vol. 348. Springer, 2007.

[34] C. Landone and M. Sandler, "Applications of binaural processing to surround sound reproduction in large spaces," in *Proc. of the IEEE International Symposium on Circuits and Systems, (ISCAS), Geneva*, vol. 3, pp. 217–220, IEEE, 2000.

[35] M. Noisternig, A. Sontacchi, T. Musil, and R. Hóldrich, "A 3D ambisonic based binaural sound reproduction system," in *AES 24th International Conference on Multichannel Audio, The New Reality*, Audio Engineering Society, 2003.

[36] R. Duraiswami, D. N. Zotkin, Z. Li, E. Grassi, N. A. Gumerov, and L. S. Davis, "High order spatial audio capture and binaural head-tracked playback over headphones with HRTF cues," in *AES 119th Convention*, 2005.

[37] Z. Li and R. Duraiswami, "Headphone-based reproduction of 3D auditory scenes captured by spherical/hemispherical microphone arrays," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2006 (ICASSP'06)*, vol. 5, pp. V–V, IEEE, 2006.

[38] D. Menzies and M. Al-Akaidi, "Nearfield binaural synthesis and Ambisonics," *The Journal of the Acoustical Society of America*, vol. 121, no. 3, pp. 1559–1563, 2007.

[39] D. N. Zotkin, R. Duraiswami, and N. A. Gumerov, "Regularized HRTF fitting using spherical harmonics," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 257–260, IEEE, 2009.

[40] M. Pollow, K.-V. Nguyen, O. Warusfel, T. Carpentier, M. Müller-Trapet, M. Vorländer, and M. Noisternig, "Calculation of head-related transfer functions for arbitrary field points using spherical harmonics decomposition," *Acta Acustica united with Acustica*, vol. 98, no. 1, pp. 72–82, 2012.

[41] S. Bertet, J. Daniel, L. Gros, E. Parizet, and O. Warusfel, "Investigation of the perceived spatial resolution of higher order Ambisonics sound fields: A subjective evaluation involving virtual and real 3D microphones," in *AES 30th International Conference on Intelligent Audio Environments*, Audio Engineering Society, 2007.

[42] S. Bertet, J. Daniel, E. Parizet, and O. Warusfel, "Investigation on localisation accuracy for first and higher order Ambisonics reproduced sound sources," *Acta Acustica united with Acustica*, vol. 99, no. 4, pp. 642–657, 2013.

[43] A. Solvang, "Spectral impairment of two-dimensional higher order Ambisonics," *Journal of the Audio engineering Society*, vol. 56, no. 4, pp. 267–279, 2008.

[44] J.-M. Batke, S. Abeling, S. Balke, and G. Enzner, "Investigation of HRTF sets using content with limited spatial resolution," in *AES 135th Convention*, Audio Engineering Society, 2013.

[45] G. Enzner, M. Weinert, S. Abeling, J.-M. Batke, and P. Jax, "Advanced system options for binaural rendering of Ambisonic format," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2013 (ICASSP'13)*, pp. 251–255, IEEE, 2013.

[46] N. R. Shabtai and B. Rafaely, "Generalized spherical array beamforming for binaural speech reproduction," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 1, pp. 238–247, 2014.

[47] N. Epain, P. Guillon, A. Kan, R. Kosobrodov, D. Sun, C. Jin, and A. van Schaik, "Objective evaluation of a three-dimensional sound field reproduction system," in *Proceedings of the 20th International Congress on Acoustics, Sydney, Australia*, 2010.

[48] S. Clapp, A. Guthrie, J. Braasch, and N. Xiang, "Evaluating the accuracy of the Ambisonic reproduction of measured soundfields," in *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*, 2014.

[49] J. Sheaffer, S. Villeval, and B. Rafaely, "Rendering binaural room impulse responses from spherical microphone array recordings using timbre correction," in *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*, 2014.

[50] E. G. Williams, *Fourier acoustics: Sound radiation and nearfield acoustical holography.* Academic press, 1999.

[51] E. Hellerud and U. P. Svensson, "Lossless compression of spherical microphone array recordings," in *AES 126th Convention*, Audio Engineering Society, 2009.

[52] E. Hellerud, A. Solvang, and U. P. Svensson, "Spatial redundancy in higher order Ambisonics and its use for lowdelay lossless compression," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2009 (ICASSP'09)*, pp. 269–272, IEEE, 2009.

[53] C. H. Choi, J. Ivanic, M. S. Gordon, and K. Ruedenberg, "Rapid and stable determination of rotation matrices between spherical harmonics by direct recursion," *The Journal of Chemical Physics*, vol. 111, no. 19, p. 8825, 1999.

[54] M. A. Blanco, M. Flórez, and M. Bermejo, "Evaluation of the rotation matrices in the basis of real spherical harmonics," *Journal of Molecular Structure: THEOCHEM*, vol. 419, no. 1, pp. 19–27, 1997.

[55] D. Murillo, F. Fazi, and M. Shin, "Evaluation of ambisonics decoding methods with experimental measurements," in *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics*, 2014.

[56] J. Meyer and G. W. Elko, "Handling spatial aliasing in spherical array applications," in *IEEE Hands-Free Speech Communication and Microphone Arrays (HSCMA 08)*, pp. 1–4, IEEE, 2008.

[57] A. W. Mills, "On the minimum audible angle," *The Journal of the Acoustical Society of America*, vol. 30, no. 4, pp. 237–246, 1958.

[58] S. A. Gelfand, *Hearing: An introduction to psychological and physiological acoustics.* Marcel Dekker New York, 1998.

[59] J. Blauert, *Spatial hearing: The psychophysics of human sound localization.* MIT press, 1997.

[60] A. W. Mills, "Lateralization of high-frequency tones," *The Journal of the Acoustical Society of America*, vol. 32, no. 1, pp. 132–134, 1960.

[61] W. Feddersen, T. Sandel, D. Teas, and L. Jeffress, "Localization of high-frequency tones," *The Journal of the Acoustical Society of America*, vol. 29, no. 9, pp. 988–991, 1957.

[62] R. Klumpp and H. Eady, “Some measurements of interaural time difference thresholds,” *The Journal of the Acoustical Society of America*, vol. 28, no. 5, pp. 859–860, 1956.

[63] J. Hebrank and D. Wright, “Spectral cues used in the localization of sound sources on the median plane,” *The Journal of the Acoustical Society of America*, vol. 56, no. 6, pp. 1829–1834, 1974.

[64] D. R. Begault, E. M. Wenzel, and M. R. Anderson, “Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source,” *Journal of the Audio Engineering Society*, vol. 49, no. 10, pp. 904–916, 2001.

[65] E. M. Wenzel, “What perception implies about implementation of interactive virtual acoustic environments,” in *AES 101st Convention*, Audio Engineering Society, 1996.

[66] D. S. Brungart and W. M. Rabinowitz, “Auditory localization of nearby sources. head-related transfer functions,” *The Journal of the Acoustical Society of America*, vol. 106, no. 3, pp. 1465–1479, 1999.

[67] M. Otani and S. Ise, “Fast calculation system specialized for head-related transfer function based on boundary element method,” *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 2589–2598, 2006.

[68] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman, “Localization using nonindividualized head-related transfer functions,” *The Journal of the Acoustical Society of America*, vol. 94, no. 1, pp. 111–123, 1993.

[69] B. Bernschütz, “A spherical far field HRIR/HRTF compilation of the Neumann KU 100,” in *AIA-DAGA 2013 : Proceedings of the International Conference on Acoustics*, DEGA, 2013.

[70] J. Nam, J. S. Abel, and J. O. Smith III, “A method for estimating interaural time difference for binaural synthesis,” in *AES 125th Convention*, Audio Engineering Society, 2008.

[71] J. C. Middlebrooks and D. M. Green, “Directional dependence of interaural envelope delays,” *The Journal of the Acoustical Society of America*, vol. 87, no. 5, pp. 2149–2162, 1990.

[72] G. F. Kuhn, “Model for the interaural time differences in the azimuthal plane,” *The Journal of the Acoustical Society of America*, vol. 62, no. 1, pp. 157–167, 1977.

[73] E. Benjamin, A. Heller, and R. Lee, “Design of ambisonic decoders for irregular arrays of loudspeakers by non-linear optimization,” in *AES 129th Convention*, Audio Engineering Society, 2010.

[74] A. V. Oppenheim, R. W. Schafer, J. R. Buck, *et al.*, *Discrete-time signal processing.* Upper Saddle River, NJ: Prentice-hall, 2 ed., 1999.

[75] J. D. Miller, M. R. Anderson, E. M. Wenzel, and B. U. McClain, "Latency measurement of a real-time virtual acoustic environment rendering system," in *Proceedings of the International Conference on Auditory Display (ICAD03)*, pp. 111–114, 2003.

[76] K. Inanaga, Y. Yamada, and H. Koizumi, "Headphone system with out-of-head localization applying dynamic HRTF (head-related transfer function)," in *AES 98th Convention*, Audio Engineering Society, 1995.

[77] P. Minnaar, S. K. Olesen, F. Christensen, and H. Møller, "The importance of head movements for binaural room synthesis," in *Proceedings of the International Conference on Auditory Display (ICAD01)*, 2001.

[78] Hillcrest Labs, http://hillcrestlabs.com/, *HCOMM Reference Manual*, 0.3 ed., March 2013.

[79] E. W. Weisstein, "Euler angles," *From Mathworld - A Wolfram Web Resource - http://mathworld.wolfram.com/EulerAngles.html, Retrieved 28.04.2014.*

[80] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano, "The CIPIC HRTF database," in *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp. 99–102, IEEE, 2001.

[81] B. Rafaely and A. Avni, "Interaural cross correlation in a sound field represented by spherical harmonics," *The Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. 823–828, 2010.

[82] D. Brungart, A. J. Kordik, and B. D. Simpson, "Effects of headtracker latency in virtual audio displays," *Journal of the Audio Engineering Society*, vol. 54, no. 1/2, pp. 32–44, 2006.

APPENDIX A

Truncation error on a rigid sphere

The normalised truncation error is defined as

$$\epsilon_N(kr) = \frac{\displaystyle\int_S |p_\infty(r,\theta,\phi,k) - p_N(r,\theta,\phi,k)|^2 \,\mathrm{d}S}{\displaystyle\int_S |p_\infty(r,\theta,\phi,k)|^2 \,\mathrm{d}S} \tag{A.1}$$

where $p_\infty(r,\theta,\phi,k)$ is the original sound field, and $p_N(r,\theta,\phi,k)$ is the sound field obtained by truncating the spherical harmonics representation to order N. S denotes the spherical surface.

Now, consider a plane wave impinging on a rigid sphere with radius R. The total pressure is a sum of the incident plane wave and the scattered wave. On the rigid sphere, the radial derivative of the total pressure must be zero:

$$\frac{\partial}{\partial r}\Big(p_i(r,\theta,\phi) + p_s(r,\theta,\phi)\Big)\bigg|_{r=R} = 0 \tag{A.2}$$

The scattered field is represented with outgoing waves [50]

$$p_s(r,\theta,\phi) = \sum_{n=0}^{\infty}\sum_{m=-n}^{n} C_n^m(\omega) h_n(kr) Y_n^m(\theta,\phi), \tag{A.3}$$

the incident field is represented with the plane wave, Equation (2.15) truncated to order N,

$$p_i(\theta,\phi) = 4\pi \sum_{n=0}^{N} \mathrm{i}^n j_n(kr) \sum_{m=-n}^{n} Y_n^m(\theta,\phi) Y_n^m(\theta_i,\phi_i)^* \tag{A.4}$$

and the solution to Equation (A.2) is

$$\begin{aligned} &4\pi \sum_{n=0}^{N} \mathrm{i}^n k j_n'(kR) \sum_{m=-n}^{n} Y_n^m(\theta,\phi) Y_n^m(\theta_i,\phi_i)^* \\ &+ \sum_{n=0}^{\infty}\sum_{m=-n}^{n} C_n^m(\omega) k h_n'(kR) Y_n^m(\theta,\phi) = 0 \end{aligned} \tag{A.5}$$

Due to the orthonormality of the spherical harmonics, the summation terms for $n > N$ must be zero, i.e. $C_n^m(\omega) = 0$, $n > N$. The expression for C_n^m becomes

$$C_n^m(\omega) = -4\pi \mathrm{i}^n \frac{j_n'(kR)}{h_n'(kR)} Y_n^m(\theta_i,\phi_i)^* \tag{A.6}$$

and the scattered pressure field $p_i + p_s$ can simply be expressed as a truncated version of Equation (2.17),

$$p_{tot}(r,\theta,\phi,\omega) = 4\pi \sum_{n=0}^{N} \mathrm{i}^n \left[j_n(kr) - \frac{j'_n(kR)h_n(kr)}{h'_n(kR)} \right] \sum_{m=-n}^{n} Y_n^m(\theta,\phi), Y_n^m(\theta_i,\phi_i)^* \quad \text{(A.7)}$$

and at the sphere surface:

$$p_{tot}(r,\theta,\phi,\omega) = 4\pi \sum_{n=0}^{N} \frac{\mathrm{i}^{n+1}}{(kR)^2 h'_n(kR)} \sum_{m=-n}^{n} Y_n^m(\theta,\phi), Y_n^m(\theta_i,\phi_i)^* \quad \text{(A.8)}$$

Now, the squared error over the rigid sphere, simplifying the pressure notation, is

$$\begin{aligned}\int_S |p_\infty - p_N|^2 \,\mathrm{d}S &= \int_S \left| 4\pi \sum_{n=N+1}^{\infty} \frac{\mathrm{i}^{n+1}}{(kR)^2 h'_n(kR)} \sum_{m=-n}^{n} Y_n^m(\theta,\phi) Y_n^m(\theta_i,\phi_i)^* \right|^2 \mathrm{d}S \\ &= (4\pi)^2 \sum_{n=N+1}^{\infty} \sum_{n'=N+1}^{\infty} \sum_{m=-n}^{n} \sum_{m'=-n'}^{n'} \frac{\mathrm{i}^{n+1}(\mathrm{i}^{n'+1})^*}{(kR)^4 h'_n(kR) h'_{n'}(kR)^*} \\ &\times Y_n^m(\theta_i,\phi_i)^* Y_{n'}^{m'}(\theta_i,\phi_i) \int_0^{2\pi} \int_0^{\pi} Y_n^m(\theta,\phi) Y_{n'}^{m'}(\theta,\phi)^* R^2 \sin\theta \,\mathrm{d}\theta \,\mathrm{d}\phi \quad \text{(A.9)}\end{aligned}$$

which, due to the spherical harmonics orthonormality property (Equation (2.7)), reduces to:

$$\int_S |p_\infty - p_N|^2 \,\mathrm{d}S = (4\pi)^2 \sum_{n=N+1}^{\infty} \sum_{m=-n}^{n} \frac{1}{(kR)^4 |h'_n(kR)|^2} |Y_n^m(\theta_i,\phi_i)|^2 R^2 \quad \text{(A.10)}$$

For a real plane wave ($N \to \infty$), the integrated squared sound pressure over the rigid sphere can similarly be expressed as:

$$\int_S |p_\infty|^2 \,\mathrm{d}S = (4\pi)^2 \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \frac{1}{(kR)^4 |h'_n(kR)|^2} |Y_n^m(\theta_i,\phi_i)|^2 R^2 \quad \text{(A.11)}$$

Thus, Equation (A.1) becomes

$$\epsilon_{N,s}(kR) = \frac{\displaystyle\sum_{n=N+1}^{\infty} \sum_{m=-n}^{n} \frac{1}{|h'_n(kR)|^2} |Y_n^m(\theta_i,\phi_i)|^2}{\displaystyle\sum_{n=0}^{\infty} \sum_{m=-n}^{n} \frac{1}{|h'_n(kR)|^2} |Y_n^m(\theta_i,\phi_i)|^2} \quad \text{(A.12)}$$

where $(4\pi)^2/k^4R^2$ cancels on both sides of the fraction. Rearranging and splitting the sums yields the normalised truncation error with the scatterer present

$$\epsilon_{N,s}(kR) = 1 - \frac{\displaystyle\sum_{n=0}^{N} \sum_{m=-n}^{n} |h'_n(kR)|^{-2} |Y_n^m(\theta_i,\phi_i)|^2}{\displaystyle\sum_{n=0}^{\infty} \sum_{m=-n}^{n} |h'_n(kR|^{-2} |Y_n^m(\theta_i,\phi_i)|^2} \quad \text{(A.13)}$$

that can be further simplified with the addition theorem (Equation (2.8)):

$$\sum_{m=-n}^{n} |Y_n^m(\theta_i, \phi_i)|^2 = \frac{2n+1}{4\pi} P_n(\cos 0) \tag{A.14}$$

and, with $P_n(\cos 0) = 1$, the final result is:

$$\epsilon_{N,s}(kR) = 1 - \frac{\sum_{n=0}^{N} |h_n'(kR)|^{-2}(2n+1)}{\sum_{n=0}^{\infty} |h_n'(kR)|^{-2}(2n+1)} \tag{A.15}$$

In practice, the denominator term must be calculated for a finite order high enough to ensure convergence of the error function. This is not a problem, since the $|h_n'(kR)|^{-2}(2n+1)$ term decreases rapidly with n above $kR = n$, as shown in Figure A.1.

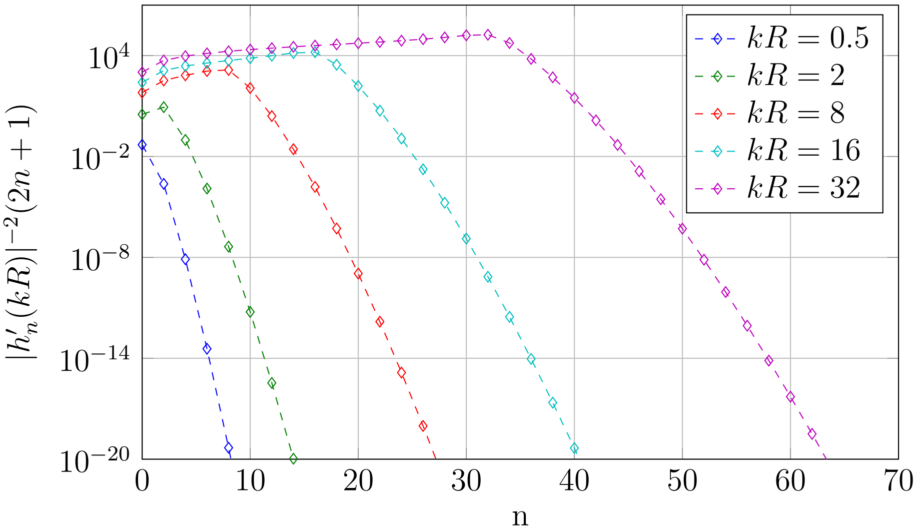


Figure A.1: Behaviour of the denominator terms in Equation (A.15).

APPENDIX B

Overview of the MATLAB code

This appendix describes the MATLAB files in the attached `zip`-file in Table B.1 and B.2. Further on, the most important scripts and functions are also included.

Table B.1: MATLAB scripts

File Name	**Description**
`SHrealTime.m`	Real-time processing of Eigenmike® microphone signals, converted to Higher Order Ambisonics, and reproduced binaurally with head-tracking. Requires Psychophysics Toolbox and DSP System Toolbox
`SHrealTimePortAudio.m`	Same as above, but does not require DSP Toolbox
`micFilterDesign.m`	Designs and stores microphone EQ FIR filters
`resampleHRTFs.m`	This script resamples the 2354-point KU 100 HRTF library to a 32-point library using a spherical harmonics representation. The sample rate (f_s) can also be changed
`chXfigY.m`	Plots the figures in this thesis, where X is the chapter, and Y is the figure number in that chapter. Placed in the attached `figures` folder

Table B.2: MATLAB functions

File name	Description
`calcBinaural-ReproductionErrors.m`	Calculates and plots various reproduction errors when reproducing HOA binaurally. Used by the `chXfigY` scripts. Look at those to better understand how this function works.
`FSM9Comm.m`	Communication with the Freespace FSM-9 with PsychHID
`FSM9motionConversion.m`	Converts Freespace FSM-9 data to quaternions
`FSM9quaterions2eulers.m`	Computes euler angles from a reference quaternion (q_0) and a motion input (q_1). First, a reverse q_0 rotation is applied, then q_1 is applied (i.e. $q = \text{conj}(q_0) \times q_1$).
`SHdecodeHOA2Bin.m`	Finds the resulting HRTFs when a virtual source is encoded with Nth order HOA, and decoded to a binaural signal with the given HRTF database.
`SHdistanceToPointOn-Sphere.m`	Finds the closest distance from a point in space, to a point on a sphere.
`Shhankelderivative.m`	Calculates the derivative of the spherical hankel function of the first kind, $h'_n(x)$.
`SHphaseCorrectHRTF.m`	Applies one of the developed phase correction methods to the HRTFs.
`SHrotationMatrix.m`	Creates the rotation matrix $\mathbf{R}$ with the 3-2-3 euler angles α, β, γ, so that the soundfield can be rotated by $\hat{\mathbf{B}} = \mathbf{R}\mathbf{B}$. Only for HOA order $N \leq 4$.
`SHsynthesizePlaneWave-OnSphere.m`	Synthesises plane wave pressure on a rigid sphere
`SHtimbreCorrection-Filter.m`	Calculates Villeval's Timbre correction filter [49]
`SHtransform.m`	Calculates the Spherical Harmonics coefficients Y_n^m.

SHrealTimePortAudio.m

```
% SHrealTimePortAudio.m
%
% Real-Time processing of EigenMike microphone signals, converted to
% Higher Order Ambisonics, and reproduced binaurally with head-
% tracking. Requires Psychophysics Toolbox
%
% This version uses PsychPortAudio to perform audio I/O
%
% Jakob Vennerød, NTNU, 2014.
% jakob.vennerod@gmail.com


%% Define input parameters
T = 25;                     % Auralization duration

Fs = 44100;                 % Sampling rate
N = 4;                      % HOA capture and reproduction order
frame_size = 1024;          % Processing frame size
nfft = frame_size*2;        % FFT size
buffer_size = 1024;         % Sound card buffer size (Input & Output)
amplification = 30;         % Signal amplification, in dB
useAudioIO = 1;             % Use audio I/O or just simulate
useMotionInput = 1;         % Use motion input or just simulate

% If no Eigenmike, either select useVirtualSource or useRecordedSig:
useVirtualSource = 1;       % Whether to use virtual source or EM32
useRecordedSig = 0;         % Whether to use a prerecorded EM32 track
FSM_sampleperiod = 4000;    % FSM-9 sample period in microseconds
PsychHID_sampleperiod = 4;  % PsychHID sample period in milliseconds

% For the phase correction
c = 343;        % Speed of sound
r_lsp = 3.25;   % HRIR Loudspeaker radius
r = 0.09;       % Assumed head radius
T0 = 4.6e-4;    % Delay from loudspeaker to origin

% Phase correction method
phCorrMethod = 'none'; %'none', 'reduceRadius' or 'linearPhase'

%% Read audio data if virtual source
if(useRecordedSig)
    % Filename (em32 recorded audio)
    fname = 'audio/em32_evidence.wav';
    mic_sig = single(audioread(fname,[1 ceil(Fs*T*1.01)])).';
elseif(useVirtualSource)
    % Filename (any audio file)
    fname = 'audio/Evidence_stereo_30s.wav';
    sig = (audioread(fname,[1 ceil(Fs*T*1.01)]));

    % Source straight ahead:
    B = SHtransform(N,pi/2,0,0);
    HOA_sig = (sig(:,1)*B).';
end

%% Compute encoding matrix
% Load elevations and azimuths for the EM32 microphone. Convert to
% radians
load('EM_mic_locations.mat')
mic_elev = EM_elevations/180*pi;
% Correct for mic rotation
mic_azi = EM_azimuths/180*pi + pi/2;

% Generate real spherical harmonics matrix for microphones
Y = SHtransform(N,mic_elev,mic_azi,0);
% Find Encoding matrix
E = single(pinv(Y));


%% Compute decoding matrix
% Load HRIR data: Impulse responses, elevations and azimuths
load('KU100_32_44100.mat');

% Phase correct HRTFs if necessary:
[hrir_L,hrir_R] = SHphaseCorrectHRTF(phCorrMethod,hrir_L,hrir_R,...
    elevations,azimuths,Fs,N,c,r_lsp,r,T0);

% Generate spherical harmonics matrix for HRIRs
Y = SHtransform(N,elevations,azimuths,0);
% Find Decoding matrix
D = pinv(Y);

% Multiply with HRIR data and find new binaural filtering matrix,
% corresponding to one HRIR filter per SH component per ear.
% Also, convert to frequency domain
DbinL = single(fft(D*hrir_L,nfft,2));
DbinR = single(fft(D*hrir_R,nfft,2));

%% Load mic equalizer FIR coefficients, convert to freq. domain
load('EQ_256samples.mat')
% Convert to frequency domain and put in a matrix

EQ(1,:) = fft(eq(1,:),nfft,2);
for i=2:4
    EQ(i,:) = fft(eq(2,:),nfft,2);
end
for i=5:9
    EQ(i,:) = fft(eq(3,:),nfft,2);
end
for i=10:16
    EQ(i,:) = fft(eq(4,:),nfft,2);
end
for i=17:25
    EQ(i,:) = fft(eq(5,:),nfft,2);
end
EQ = single(EQ);

```

```
%% Initialize PsychHID and the FSM-9
if(useMotionInput)
    % Find FSM-9 device number
    devNo = FSM9Comm('Discover');
    if(devNo == -1)
        return;
    end

    % Set FSM-9 SampleRate in microseconds
    outputdata = FSM9Comm('SetSampleRate',devNo,FSM_sampleperiod);

    % Set PsychHID sample rate in milliseconds
    FSM9Comm('SetPsychHIDSampleRate',devNo,PsychHID_sampleperiod);

    % Set configuration Full Motion On
    outputdata = FSM9Comm('FullMotionOn',devNo);

    input('Place the sensor in position for calibration and press ENTER...')


    % Flush some reports
    FSM9Comm('Flush',devNo);

    % Find the calibration state
    motionInput = FSM9Comm('ReceiveMotionData',devNo);
    if(isempty(motionInput))
        return;
    end
    if(motionInput(1) == 38)
        q0 = FSM9motionConversion(motionInput(11:18));
        q_old = q0;
        q = q0;
    end
    % Pause for keycheck
    input('OK, when ready press enter...')
    pause(0.2)
end
% Initialize rotation matrix
R = single(SHrotationMatrix(0,0,0));

%% Pre-initialize stuff
n = 1;                                  % Sample counter
newMotionInput = 0;                     % Motion input flag

% Some audio arrays
HOA_frame_in = single(zeros((N+1)^2,frame_size));
HOA_frame = HOA_frame_in;
HOA_frame_L = HOA_frame;
HOA_frame_R = HOA_frame;
tail_eq = HOA_frame_in;
tail_L = HOA_frame_in;
tail_R = HOA_frame_in;
tmp = single(zeros((N+1)^2,nfft));
Left = single(zeros(1,frame_size));
```

```matlab
Right = single(zeros(1,frame_size));

% Load KbCheck and WaitSecs from PsychToolbox
KbCheck;
WaitSecs;

%% Initialize PortAudio
if(useAudioIO)
    % Number of input channels (set to 32 for EigenMike)
    if(useVirtualSource || useRecordedSig)
        inChannels = 2;
    else
        inChannels = 32;
    end
    % Perform low-level initialization of the sound driver:
    InitializePsychSound(1);
    % Level of debug output:
    PsychPortAudio('Verbosity', 1);
    % Open PortAudio input
    painput = PsychPortAudio('Open', [], 2, 2, Fs, inChannels, 0,[]);
    % Preallocate an internal audio recording  buffer with a capacity
    % of 10 seconds:
    PsychPortAudio('GetAudioData', painput, 10);
    % Open default audio device [] for playback
    paoutput = PsychPortAudio('Open', [], 1, 2, Fs, 2, buffer_size, []);
    % Start audio capture immediately and wait for the capture to start.

    % Perform output warmup start
    PsychPortAudio('FillBuffer', paoutput, zeros(2,1024));
    PsychPortAudio('Start', paoutput, 1, 0, 1);
    PsychPortAudio('Stop', paoutput, 1);


    painputstart = PsychPortAudio('Start', painput, 0, 0, 1);

    % Quickly readout available sound and initialize sound output
    % buffer with it:
    [audiodata, ~, ~, capturestart]= PsychPortAudio(...
       'GetAudioData',painput, [], frame_size/Fs, frame_size/Fs, 1);
    % The frame_size variable actually determines the (desired)
    % latency here.

    % Feed everything into the initial sound output buffer:
    PsychPortAudio('FillBuffer', paoutput, audiodata(1:2,:));

    % Start the playback engine immediately and wait for start.
    playbackstart = PsychPortAudio('Start', paoutput, 0, 0, 1);
    % Expected latency (I->O)
    expecteddelay = (playbackstart - capturestart) * 1000;
end

% Start timer
tic
[keyIsDown, secs, keyCode] = KbCheck;
while(n < Fs*T)
```

```

    if(useVirtualSource)
        HOA_frame = HOA_sig(:,n:n+frame_size-1);

        % Dummy read from audio input
        [audiodata_in] = PsychPortAudio('GetAudioData', painput,...
            [], frame_size/Fs,frame_size/Fs,1);
    else
        if(useAudioIO)
            % Record audio and amplify
            [audiodata_in] = PsychPortAudio('GetAudioData',...
                painput, [], frame_size/Fs, frame_size/Fs, 1);
            mic_signals = single(audiodata_in*10^(amplification/20));
        end

        if(useRecordedSig)
            mic_signals = mic_sig(:,n:n+frame_size-1);
        end


        % Encode to HOA format
        HOA_frame_in = E*mic_signals;

        % Microphone EQ filtering. Store the filter conditions in
        % tail matrices
        tmp = ifft(fft(HOA_frame_in,nfft,2).*EQ,nfft,2);
        HOA_frame = tmp(:,1:frame_size) + tail_eq;
        tail_eq = tmp(:,frame_size+1:end);

    end

    if(useMotionInput)
        % Get motion input from FSM-9
        % Flush last reports. Needed?
        reports = FSM9Comm('Flush',devNo);

        % Get motion input
        motionInput = FSM9Comm('ReceiveMotionData',devNo);
        if(motionInput(1) == 38)
            % New motion report
            newMotionInput = 1;
        elseif(motionInput(1) == -1 && (~isempty(reports)))
            % Check last report
            motionInput = reports(end).report;
            if(motionInput(1) == 38)
                newMotionInput = 1;
            else
                newMotionInput = 0;
            end
        else
            % No new motion input
            newMotionInput = 0;
        end

        % Find new orientation if we have new motion input
        if(newMotionInput)
            % Extract quaternions
            q = FSM9motionConversion(motionInput(11:18));
            [phi, theta, psi] = FSM9quaternions2eulers(q0,q);
        end

        % Check if the new motion input is really new:
        angle_movement = sqrt(sum((q(2:4)-q_old(2:4)).^2));

        if( angle_movement < 0.001)
            newMotionInput = 0;
        end

        % Find new rotation matrix if necessary
        if(newMotionInput)
            % Store old motion data
            q_old = q;
            %Reverse (for head-tracking):
            R = single(SHrotationMatrix(-psi,theta,-phi));
            %Normal (for point-to-source):
            %R = single(SHrotationMatrix(phi,-theta,psi));
        end
    end

    % Rotate HOA signals
    HOA_frame_rotated = R*HOA_frame;

    %HRIR filtering. Can be done before rotation.
    HOA_f_r_F = fft(HOA_frame_rotated,nfft,2);

    tmp = ifft(HOA_f_r_F.*DbinL,nfft,2);
    HOA_frame_L = tmp(:,1:frame_size) + tail_L;
    tail_L = tmp(:,frame_size+1:end);

    tmp = ifft(HOA_f_r_F.*DbinR,nfft,2);
    HOA_frame_R = tmp(:,1:frame_size) + tail_R;
    tail_R = tmp(:,frame_size+1:end);

    % Sum the HOA signals to L+R
    Left = sum(HOA_frame_L,1);
    Right = sum(HOA_frame_R,1);

    % Create output vector
    outputdata = [Left;Right];

    if(useAudioIO)
        PsychPortAudio('FillBuffer', paoutput, outputdata, 1);
    end

    % Increase sample counter
    n = n + frame_size;
end
toc
%%

```

```
% Clean up PortAudio
if(useAudioIO)
    % Done. Stop the capture engine:
    PsychPortAudio('Stop', painput, 1);
    % Drain its capture buffer...
    [audiodata_drain,~] = PsychPortAudio('GetAudioData', painput);
    % Stop the playback engine:
    PsychPortAudio('Stop', paoutput, 1);
    % Ok, done. Close engines and exit.
    PsychPortAudio('Close');
end

% Stop the motion sensor
if(useMotionInput)
    FSM9Comm('Sleep',devNo);
    FSM9Comm('Flush',devNo);
    FSM9Comm('StopReports',devNo);
end
```

SHdecodeHOA2Bin.m

```matlab
function [HRTF_L,HRTF_R,HRIR_L,HRIR_R] = SHdecodeHOA2Bin(elev_in,azi_in,hrir_elev,hrir_azi,N,hrir_L,hrir_R,nfft)
%SHdecodeHOA2Bin Finds the resulting HRTFs when a virtual source
%is encoded with Nth order HOA, and decoded to a binaural signal
%with the given HRTF database.
%
% Input arguments:
%   elev_in     Elevation of the virtual source
%   azi_in      Azimuth of the virtual source
%   hrir_elev   Elevations for the HRIRs
%   hrir_azi    Azimuths for the HRIRs
%   N           Desired HOA truncation order
%   hrir_L      Left ear HRIRs
%   hrir_R      Right ear HRIRs
%   nfft        FFT size
%
% Output arguments:
%   HRTF_L  Resulting left ear HRTF
%   HRTF_R  Resulting right ear HRTF
%   HRIR_L  Resulting left ear HRIR
%   HRIR_R  Resulting right ear HRIR
%
%   Jakob Vennerød, NTNU, 2014.
%   jakob.vennerod@gmail.com

    % SH representation of the virtual source
    B = SHtransform(N,elev_in,azi_in,0);

    % Find the sperhical harmonic coefficients corresponding to the
    % HRIR angles
    Y = SHtransform(N,hrir_elev,hrir_azi,0);

    % Decoding matrix
    D = pinv(Y);

    % Multiply with HRIR components and find new decoding matrix
    DbinL = D*hrir_L;
    DbinR = D*hrir_R;

    % Find HRIRs
    HRIR_L = (B*DbinL).';
    HRIR_R = (B*DbinR).';

    % Find HRTFs
    HRTF_L = fft(HRIR_L,nfft);
    HRTF_R = fft(HRIR_R,nfft);
end
```

SHphaseCorrectHRTF.m

```
function [hrir_L_new, hrir_R_new] = SHphaseCorrectHRTF(method,hrir_L,
    hrir_R,elevations,azimuths,Fs,N,c,R,r,T0,nfft)
%SHphaseCorrectHRTF Apply one of the developed phase correction
%methods to the HRTFs.
%
% Input arguments:
%   method      String, 'linearPhase', 'reduceRadius' or 'none'
%   hrir_L      Left ear HRIRs
%   hrir_R      Rgiht ear HRIRs
%   elevations  HRIR elevations
%   azimuths    HRIR azimuths
%   Fs          Sampling rate
%   N           HOA truncation order
%   c           Speed of sound
%   R           HRIR loudspeaker radius
%   r           Assumed head radius
%   T0          Delay from loudspeaker to origo
%   nfft        FFT size (does not really matter as long as it is
%               larger than the HRIR length)
%
% Output arguments:
%   hrir_L_new  Left ear corrected HRIRs
%   hrir_R_new  Right ear corrected HRIRs
%
%
%   Jakob Vennerød, NTNU, 2014.
%   jakob.vennerod@gmail.com

% Length of HRIRs
M = size(hrir_L,2);

switch(method)

    case{'linearPhase'}
        % Set the phase to linear for each HRTF, for all frequencies
        % above kr = N. The linear phase is set so that the group
        % delay corresponds to T0.

        nAngles = size(hrir_L,1);

        % Convert to HRTF
        HRTF_L = fft(hrir_L,nfft,2);
        HRTF_R = fft(hrir_R,nfft,2);

        % Freq vector
        f = linspace(0,Fs-Fs/nfft,nfft);

        % Calculate frequency limit and index
        flim = N/r*c/(2*pi);
        idx = find(f>=flim);
        idx = idx(1);

        % Width of radial frequency bins
        dw = 2*pi*(f(2)-f(1));

        % Phase correct HRTFs
        for i = 1:nAngles
            % Original phase
            phi_L = unwrap(angle(HRTF_L(i,:)));
            % Modify phase
            phi_L(idx:nfft/2) = phi_L(idx)-T0*dw*(0:nfft/2-idx);
            phi_R = unwrap(angle(HRTF_R(i,:)));
            phi_R(idx:nfft/2) = phi_R(idx)-T0*dw*(0:nfft/2-idx);
            % Modify HRTFs
            HRTF_L(i,idx:nfft/2) = abs(HRTF_L(i,idx:nfft/2)).*...
                exp(1i*phi_L(idx:nfft/2));
            HRTF_R(i,idx:nfft/2) = abs(HRTF_R(i,idx:nfft/2)).*...
                exp(1i*phi_R(idx:nfft/2));
            HRTF_L(i,nfft/2+2:end) = conj(HRTF_L(i,nfft/2:-1:2));
            HRTF_R(i,nfft/2+2:end) = conj(HRTF_R(i,nfft/2:-1:2));
        end

        % Convert to HRIRs
        hrir_L_new = ifft(HRTF_L,nfft,2);
        hrir_R_new = ifft(HRTF_R,nfft,2);

    case{'reduceRadius'}
        % Reduce the head radius such that kr is constant (=N)
        % above flim.

        nAngles = size(hrir_L,1);

        % Convert to HRTF
        HRTF_L = fft(hrir_L,nfft,2);
        HRTF_R = fft(hrir_R,nfft,2);

        % Freq vector
        f = linspace(0,Fs-Fs/nfft,nfft);

        % Calculate frequency limit and index
        flim = N/r*c/(2*pi);
        idx = find(f>=flim);
        idx = idx(1);

        % New radius (kr = N);
        new_r = N./(2*pi*f/c);

        % Phase correct HRTFs
        for i = 1:nAngles
            % Find distances from the loudspeaker to the ears
            dist_L = SHdistanceToPointOnSphere(R,elevations(i),...
                azimuths(i),r,[0 r 0],1000);
            dist_R = SHdistanceToPointOnSphere(R,elevations(i),...
                azimuths(i),r,[0 -r 0],1000);

            % Depending on in which half-sphere the loudspeaker is..
            if(azimuths(i)>=0 && azimuths(i)<pi)
                % Left ear is on "sunny side"
                costheta = (r^2+R^2-dist_L^2)/(2*r*R);
                new_dist_L = sqrt(new_r.^2+R^2-2*new_r*R*costheta);
                % Right ear is on "shadow side". Find Great
                % Circle Distance
                gcd = dist_R - sqrt(r^2+R^2); %
                new_gcd = gcd/r*new_r;
                new_dir = sqrt(R^2+new_r.^2);
                new_dist_R = new_gcd + new_dir;
            else
                % Right ear is on "sunny side"
                costheta = (r^2+R^2-dist_R^2)/(2*r*R);
                new_dist_R = sqrt(new_r.^2+R^2-2*new_r*R*costheta);
                % Left ear is on "shadow side". Find Great
                % Circle Distance
                gcd = dist_L - sqrt(r^2+R^2);
                new_gcd = gcd/r*new_r;
                new_dir = sqrt(R^2+new_r.^2);
                new_dist_L = new_gcd + new_dir;
            end

            % Calculate distance differences
            dist_diff_L = new_dist_L - dist_L;
            dist_diff_R = new_dist_R - dist_R;

            % Phase correction terms
            phasecorr_L = exp(-1i*2*pi*f.*dist_diff_L/c);
            phasecorr_R = exp(-1i*2*pi*f.*dist_diff_R/c);

            % Modify HRTFs
            HRTF_L(i,idx:nfft/2) = HRTF_L(i,idx:nfft/2).*...
                phasecorr_L(idx:nfft/2);
            HRTF_R(i,idx:nfft/2) = HRTF_R(i,idx:nfft/2).*...
                phasecorr_R(idx:nfft/2);
            HRTF_L(i,nfft/2+2:end) = conj(HRTF_L(i,nfft/2:-1:2));
            HRTF_R(i,nfft/2+2:end) = conj(HRTF_R(i,nfft/2:-1:2));
        end

        % Convert to HRIRs
        hrir_L_new = ifft(HRTF_L,nfft,2);
        hrir_R_new = ifft(HRTF_R,nfft,2);

    case{'none'}
        % No phase correction, just pass the HRIRs on..
        hrir_L_new = hrir_L;
        hrir_R_new = hrir_R;
end

% Remove zeros.
hrir_L_new = hrir_L_new(:,1:M);
hrir_R_new = hrir_R_new(:,1:M);

end
```

SHtransform.m

```
function [ Y ] = SHtransform(N, theta, phi, def)
%SHtransform Calculates the Spherical Harmonics coefficients Ynm
% Usage: Y = SHtransform(N, theta, phi, def)
%
%   Takes in HOA order N, angle vectors theta and phi
%   Returns the Spherical Harmonics amplitudes
%
%   Yq(n,m) =
%
%   [Y1(0,0) Y1(1,-1) Y1(1,0) Y(1,1) Y1(2,-1) ... Y1(N,N);
%    Y2(0,0) ...
%    ...
%    YQ(0,0) ...                           ...    YQ(N,N)]
%
%   def is the definition parameter:
%       1 = Williams' definition (complex Y)
%       0 = Daniel's definition  (real Y)
%
%
%   Jakob Vennerød, NTNU, 2014.
%   jakob.vennerod@gmail.com

    % Number of calc. angles
    Q = length(theta);

    % Initialize
    Y = zeros(Q,(N+1)^2);

    % Compute factorials
    factorials = [1 1 cumprod(2:(2*N))];

    % Compute phase values
    phase = zeros(Q,N);

    for m = 1:N
        phase(:,m) = exp(1i*m*phi);
    end
    if(def == 1) % Williams' definition

        Y(:,1) = 1/sqrt(4*pi);
        for n = 1:N

            % Find Legendre polynomials of degree n
            Pn = legendre(n,cos(theta)).';

            % Calculate spherilca harmonics
            idx = n^2 + 1 + n ;
            for m = 1:n
                tmp = sqrt((2*n+1)/(4*pi)*factorials(n-m+1)/...
                    factorials(n+m+1))*Pn(:,m+1).*phase(:,m);
                % Positive m
                Y(:,idx+m) = tmp;
                % Negative m
                Y(:,idx-m) = (-1)^m*conj(tmp);
            end
            % m = 0
            Y(:,idx) = sqrt((2*n+1)/(4*pi))*Pn(:,1);

        end

    else % Daniel's definition

        Y(:,1) = 1;
        for n = 1:N

            % Find Legendre polynomials of degree n
            Pn = legendre(n,cos(theta)).';

            % Calculate spherical harmonics
            idx = n^2 + 1 + n ;
            for m = 1:n
                tmp = sqrt((2*n+1)*2*factorials(n-m+1)/...
                    factorials(n+m+1))*Pn(:,m+1);
                % Positive m
                Y(:,idx+m) = tmp.*real(phase(:,m));
                % Negative m
                Y(:,idx-m) = tmp.*imag(phase(:,m));
            end
            % m = 0
            Y(:,idx) = sqrt((2*n+1)*1)*Pn(:,1);

        end
    end
end
```

View publication stats