NTNU
Norwegian University of
Science and Technology

Spectral impairment in HRTF based binaural reproduction systems

Magnus Hauge Skeide

Master of Science in Communication Technology
Submission date: July 2017
Supervisor: Odd Kr. Pettersen, IES
Co-supervisor: Audun Solvang, SINTEF

Norwegian University of Science and Technology
Department of Electronic Systems

Problem description

This master thesis will look at the spectral impairments that occurs when using HRTF based binaural reproduction systems. The main focus will be on reproduction by using Higher Order Ambisonics (HOA). The thesis will look at which orders are needed to avoid spectral impairments in the reproduction for different frequency intervals, and for different resolutions of HRTF databases. If possible, a recommendation for an ideal HRTF resolution shall be suggested, as well as methods or ideas for improving these impairments.

Preface

I have both dreaded and looked forward to this day for the last couple of months, as the writing of this final note represents the end of five years of studying in Trondheim, and the completion of my MSc. degree in Communication Technology at the Norwegian University of Science and Technology (NTNU). The work with this thesis took place from February to July, 2017.

I had a desire to work with spatial audio, and the initial topic of the thesis was suggested after discussions with co-supervisor Audun Solvang. Throughout the following months, the primary goal for investigation changed slightly even though the main topic remained the same, and I tried to include aspects of binaural hearing into the work, after my own interest.

I knew for a long time that I wanted to study in Trondheim and at NTNU. I found my place at Communication Technology, and with music being my favorite hobby, I was lucky enough to be able to specialize myself in acoustics. I have learned a lot from all of my activities in Trondheim, and gained a lot of new friends, and even though I certainly know more know than when I started, I still look forward to continue the learning process as I enter the working life.

For helping me out with the thesis, I would first of all like to thank my supervisors at SINTEF, Prof. Odd K. Pettersen and Audun Solvang, for guidance and assistance, and helping me come up with new ideas when I felt a bit stranded. Also, a special thanks to Tron Vedul Tronstad at SINTEF, who was of great help and provided invaluable discussions in the last and most intense period of the work.

I would also like to thank all my friends and co-students who have supported me through these years. Finally, a sincere thanks to my family who always supports me, and has helped reach the place I am today.

Magnus Hauge Skeide
Trondheim, July 2017.

Abstract

Higher Order Ambisonics is a method for capturing and reproducing sound fields. With Ambisonics‘ advantageous features and possibilities for binaural reproduction, the method has been established as one of the best ways of recreating 3D sound fields, and is widely used in both VR- and 360° applications. A drawback is that accurate reconstruction is only feasible inside a sphere, with radius r limited by $ekr < 2N - 1$, where e is the base of the natural logarithm, k is the wavenumber and N is the Ambisonics order.

This thesis presents the theory for spherical harmonics, binaural hearing and Ambisonics, and how they can be combined. HRTF datasets with different resolutions have been investigated to see if a recommended resolution can be suggested. Localization cues were also investigated. All evaluations were done with $\phi = 0°$ and for orders N=1-4.

For the HRTF reconstruction, the results behaved as expected. The lowest resolution dataset resembled its reference most for N=1-2. All of the datasets did sufficiently reconstruct the ITD_p correctly for N=1-2. The key features of the ILD were kept somewhat intact up to $f \approx$ 6kHz, with some ambiguity.

The results are based on a visual comparison model. The work in this thesis lacks a more clear, objective measure, and should also be complimented with a listening test showing the significance of the changes. Hence, this thesis does not present a strong enough foundation to provide a conclusion for a recommended HRTF resolution. However, ideas for future work are presented, which could be implemented in order to reach the desired conclusion.

Sammendrag

Høyere ordens Ambisonics er en teknikk for å ta opp og gjenskape lydfelt. På bakgrunn av fordelaktige egenskaper og muligheter for binaural reproduksjon, har Ambisonics etablert seg som en av de beste måtene for å gjenskape 3D lydfelt, og brukes derav mye i både VR- og 360°-applikasjoner. En kritisk ulempe er at god gjenskapelse kun er mulig inne i en sfære med radius r, avgrenset av $ekr < 2N - 1$, hvor e er basen til den naturlige logaritmen, k er bølgenummer og N er Ambisonics orden.

Denne oppgaven presenterer teorien for sfæriskharmoniske funksjoner, binaural hørsel og Ambisonics, og viser hvordan de kan kombineres. HRTF datasett med ulike oppløsninger har blitt undersøkt for å se om en anbefalt oppløsning kan foreslås. Lokaliseringshint har også blitt undersøkt. Alle evalueringer er foretatt for $\phi = 0°$ og orden N=1-4.

For HRTF gjenskapelsen oppførte resultatene seg som forventet. Datasettet med lavest oppløsning var enklest å gjenskape, og lignet referansen mest for N=1-2. Alle datasettene gjenskapte ITD_p i tilstrekkelig grad, mens ILD-gjenskapelsen inneholdt en del tvetydigheter, til tross for at de sterkeste responsene ble korrekte.

Resultatene er basert på en visuell sammenligningsmodell. Arbeidet i oppgaven mangler en tydelig, objektiv målestandard, og burde ideellt også kompletteres av en lyttetest, for å undersøke signifikansen til enkelte endringer. På bakgrunn av dette, presenterer ikke oppgaven et sterkt nok grunnlag til å foreslå en entydig konklusjon for en anbefalt HRTF oppløsning. Det er derimot foreslått ideer for fremtidig relevant arbeid, som kan implementeres slik at problemstillingen kan løses.

Table of Contents

List of Figures

List of Tables

Abbreviations

VR	=	Virtual reality
HRTF	=	Head related transfer function
WFS	=	Wave field synthesis
VBAP	=	Vector based amplitude panning
HOA	=	Higher order ambisonics
MMA	=	Minimum audible angle
JND	=	Just noticable difference
COC	=	Cone of confusion
FFT	=	Fast Fourier Transform
HRIR	=	Head related impulse response
ITD_p	=	Interaural phase delay difference
ITD	=	Interaural time difference
ILD	=	Interaural level difference
MML	=	MIT Media Lab

Chapter 1

Introduction

Applications for virtual reality (VR) has seen a boom the last few years, as the necessary equipment to experience VR has become both more accessible and affordable for the public audience. The range of what is possible to experience in VR is in a rapid growth, from movies and interactive games, to more practical applications such as city planning and tactical training for the military. Some of the latest additions as of 2017, is the introduction of omnidirectional treadmills, which in combination with separate hand controllers takes the immersion[1] to another level. Still, one could say that the graphical representation of the virtual world, and the possibility to interact with it, only makes up 2/3 of the experience, with a convincing 3D-sound being the last piece of the puzzle. Whether or not the 3D-sound makes up exactly one third of the experience is hard to quantify nor a point to do so, but it is beyond any doubt that sound needs as much focus as the visual aspects, and probably more due to how complex our hearing is.

Compared to seeing, human hearing is more directly wired to the human perception. The auditory system, which will be mentioned to some detail later, is incredibly quick to receive, recognize and process sound information, and can often utilize this information much quicker than visual information. This is why 3D audio is such an important part of the whole VR experience. Realistic 3D audio do more than to just immerse a user in some virtual space; it also gives a physical connection to the elements of that space. There are also several other applications which enjoy the benefits from 3D audio, such as movie theaters, teleconferencing and pure music listening, to name a few.

So what is *3D-sound*, and how is this different from traditional *stereo* and *surround* sound? While stereo[2] sound is most commonly associated with sound reproduction from two sources to the left and right of the listener, and surround sound with a horizontal loudspeaker array surrounding the listener, 3D-sound introduces elevation as a third dimension to the reproduced sound field.

During the last couple of decades, three main methods for 3D audio have been pro-

[1] Immersion into virtual reality is a perception of being physically present in a non-physical world.

[2] The term 'stereo' origins from *stereophonic sound*, which could also include more than two speakers, but is also used to describe 'quadraphonic' and 'surround' sound.

posed, and more or less continuously improved up until today. These are Vector Based Amplitude Panning (VBAP) [1], Wave Field Synthesis (WFS) [2] and Higher-Order Ambisonics (HOA) [3]. VBAP, in both 2D and 3D, is based upon panning methods[3] which are reformulated with vectors and vector bases. WFS is based on the quantification of Huygen‘s principle, stating that any point of a wave front can be considered as a secondary source.

HOA aims to decompose a spatial sound field by using a truncated spherical harmonics decomposition, and then reproduce that sound field over a given listening area called the *sweet spot*. An obvious drawback for these methods is the trade-off between spatial resolution and the amount of loudspeakers required for playback. Since an ideally perfect reproduction of the sound field over all frequencies would require functions of infinite order, and thus an infinite amount of loudspeakers, some kind of truncation is necessary. However, if one is able to adapt the reproduced sound field to the human hearing by binaural filtering, such that the amount of loudspeakers in practise becomes *virtual* loudspeakers, the aforementioned issue could be resolved. This will in turn introduce new challenges. Some of the them have already been investigated, while some of them will be investigated and discussed in this thesis.

1.1 Motivation

Of the previously mentioned methods for reconstructing 3D-sound, this thesis will solely focus on HOA. This is due to the solid theoretical foundation already available, an increasing amount of attention and papers released on the subject and having an co-supervisor, A. Solvang[4], with experience within the field. Though the standardization of a single 3D-audio format is still an arms race, HOA has truly been established as one of the strong contenders by being included as the main format for 360- and VR audio in both Youtube [4] and for Facebook [5].

The size of the sweet spot with radius r for which an accurate reconstruction of a sound field is possible has by a rule of thumb usually been limited to $r = \frac{N}{k}$, where N is the Ambisonics order and k is the frequency. The "loudspeaker cost" for such a setup is quite costly, since $M = (N+1)^2$ speakers are needed for any order N, which quickly leads to unfeasible setups[5] for HOA. One previously suggested solution to overcome this challenge, is to to convert the HOA signals into a binaural format, such that high quality headphones could be used for the sound reproduction, instead of complicated and expensive loudspeaker setups. Then, for an assumed head radius of $r \approx 0.10$m, one can instead investigate up to which frequency $f = \frac{Nc}{2\pi r}$ Ambisonics of a finite order N, is able to reconstruct the near perfect sound field within the sweet spot. Such a binaural filtered Ambisonics format could enable realistic *auralization* as suggested by Kleiner et al. in [6], which is the ability to render a spatial sound field along to go along with some kind of visualization. In order to do so, the HOA loudspeaker signal would have to be filtered with Head-Related Transfer Functions (HRTFs), which can be measured either on human subjects or on a specialized microphone setup simulating a human subject, such

[3] Panning is the distribution of a sound signal into a new stereo or multi-signal sound field.
[4] Research scientist at SINTEF.
[5] With an exception for research facilities with sufficient financial budget and space available.

as the KEMAR[6] mannequin. The concept of adapting HOA to fit a binaural setting is not completely new, and some main results and points from previous studies will be presented in section 1.2. However, there are fewer studies on objective evaluations of sound and how different HRTFs in terms of resolution affects the percieved sound quality, which this thesis will try to give some contribution to. Finally, to relate the order of Ambisonics that will be investigated to actual applications, orders N=1-4 will be looked into, since there are no recording equipment or software workstation that supports anything higher as of spring 2017.

1.2 Previous works and literature review

The idea of using spherical harmonics to decompose sound fields, which is the founding idea for Ambisonics, was suggested by Gerzon [7, 8] already in the early 1970s. He was dissatisfied with the quality of the quadraphonic sound reproduction. He suggested an improved 4-channel method, based on what he called the ABCD-formats, which he presented as Ambisonics in 1985 [9]. Since then, Ambisonics have been further developed, particularly by Daniel et al. [3], to include modes of higher orders, making it into what today is known as Higher Order Ambisonics (HOA). Particularly from the late 1990s and early 2000s, as a continuous increase in technology allowed larger and more advanced loudspeaker solutions, several studies on HOA and virtual source imaging were conducted, with Daniel et al. [3, 10] responsible for many of them, which resulted in a lot of developments.

Before presenting the most recent studies, it is useful to look back at some of the earlier studies on *binaural localization*, to get a sense of the development. Binaural localization is the ability to localize sound sources, and studies and listening experiments on this feature have been conducted for several decades. An often used measure for evaluation is the so called *minimum audible angle (MAA)*, more generally from the world of physics as the *just noticable difference (JND)*. This measure aims to quantify how small changes in source angles a subject can differentiate, when exposed to sound from equivalent sources, in order to say something about the resolution of human hearing. The ability to localize a sound is based upon several localization cues, and Wallach‘s studies [11] from as far back as in the late 1930s suggests that different localization cues are predominant depending on which frequencies that are being stimulated. It is also stated that localization to the sides are poor, due to an area to the side of the ears called the cone of confusion, which is a term still used to this day. In the late 1950s, Mills [12] did further studies on the suggested localization cues and the resolution of the human hearing, in terms of the MMA. His results concurred with previous studies, concluding with a very good resolution (i.e. small JND) at lower frequencies for sources positioned straight in front, and vice versa for sources to the sides. He also concluded that the cone of confusion which leaves the sides indeterminate, additionally affects the ability to differentiate sources in the front and back, so called front-back localization.

While the features of localization along the horizontal plane was investigated quite early, that was not the case for vertical localization, for which the horizontal localization

[6] Knowles Electronics Mannequin for Acoustics Research

cues collapses. In the late 1960s [13, Blauert] and early 1970s [14, Gardner, Gardner] [15, Hebrank, Wright], it was suggested and tested that high frequencies, filtering by the folds in the pinnae, contained the predominant cues for vertical localization, and that these were most applicable in the 4-16 kHz range. Following studies concurred with this, and in the late 1980s through the 1990s, Wightman & Kistler did thorough work [16, 17, 18] on free-field simulation through headphones, as well as sound localization and the use of non-individualized head-related transfer functions. Their studies were focused on *localization* in two dimensions, instead of only *lateralization* along one dimension (usually along the horizontal plane). They found that the robustness of horizontal plane localization were strong even for the non-individualized cases, and a surprisingly majority of the test group also did well for vertical localization, suggesting that individualized HRTFs might be unnecessary for a 3D sound field to be authentic. Noisternig et al. [19] also concluded with an successful use of non-individualized combined with a head-tracker device, where a combination of different Ambisonic orders, so called *mixed HOA*, have been used along with a room simulation model. The datasets used in that study are the same that are to be discussed later on in this thesis. Moreover, will it be interesting to compare these datasets to the study by Minnaar et al. [20], which concluded with a necessary directional resolution of 2°. Such a resolution would demand a total of 11975 HRTF pairs, but the study claimes that this number could be greatly reduced to just 1130 pairs, by interpolating neighboring minimum-phase components.

As is obvious from literature, evaluation of the hearing resolution from binaurally synthesized sources has become a more popular topic. Most of the studies aims to measure how well a reproduction system is able to reconstruct the spatial information. That could be either for a virtual source, or how well a reference HRTF can be reproduced via such a system. Apain et al. [21] did such studies on a mannequin in 2010, such that objective conclusions could be made. By using a system that allowed HOA of order 4, they were able to find that the ITD were near perfectly recreated within the frequency limits of the Ambisonics order, and also in a much higher frequency range. The ILD did on the other hand not recreate as accurate, and due to the frequency limit of the system, the perception of elevation was not ideal. However, since the results also showed some degree of flexibility regarding positioning of the sweet spot compared to the mannequin, it is fair to assume that some perception of elevation can be achieved simply by rotating the head. A study from 2017 by Xie et al. [22] supports this idea. They found that due to the limitations of HOA of lower orders, which prevents correct reproduction of high-frequency content, Ambisonic reproduction is not suited to reproduce either the previously mentioned spectral cues. However, it was found that the ITD changes in the Ambisonic reproduction that are caused by head turning, which are most predominant in the lower frequency content, matched those of the real source, implying that by vertical localization can still be solved by using a head tracker and using the dynamic cues.

1.3 Problem statement

With the methods available today for measuring head related transfer functions (HRTFs), it is not feasible in the near foreseeable future that every user of 3D-audio applications will have their own, individual HRTF. One should therefore rather look at how good a sound field can be reproduced with the means and methods that are already available, which is the focus of this thesis. More specifically:

How can the current state of Ambisonics be utilized in today's application, and how do different resolutions of HRTFs respond to this.

To say something about the primary task, some necessary sub tasks needs to be looked into first. These are:

- Suggesting a recommended resolution for HRTF sets based on the results from the experiments, and see how these differs from the relevant studies.
- Following the previous sub problem, suggesting an absoluyte *minimum* resolution for HRTF datasets, such that an even lower resolution than this suggestion, will render the expected experienced quality too poor for practical use.
- How will the HOA-induced errors look for Ambisonics of order N=1-4, and investigate how the binaural localization cues are restored for

1.4 Outline

The format of the rest of this thesis will be as follows:

Chapter 2	This chapter will contain the theoretical foundation necessary to follow what is attempted in this thesis, as well as the results and discussion. The theory chapter aims to be sufficiently detailed, and will refer to complimentary literature where only brief segments are included, or longer derivations are excluded.
Chapter 3	This chapter will describe how the results are acquired.
Chapter 4	The results are presented with some brief comments in this chapter.
Chapter 5	This chapter will provide a discussion about the results, some suggestions for immediate improvement, sources of error and finally suggest some topics for future work related to the described problem.
Chapter 6	A final conclusion for the work of this thesis.
Appendix	The only code provided in the appendix is the script *gen_sph_harm_mat.m*, provided by assistant supervisor Solvang. Here one can see how the spherical harmonics are estimated, in case a reader is only interested in this part, and wishes to build the rest for him- or herself.
Attachments	The digital hand in of this thesis will contain the scripts used to find the results. The author aims to comment these scripts in a sufficiently detailed manner, such that a reader with some experience within the field is able to use or adapt them for his or her own purpose.

Chapter 2

Theoretical foundation

The theoretical foundation of this thesis consists of three major parts. This chapter is structured such that the reader will first receive an introduction to spatial acoustics and 3D sound fields. Following is a chapter which reviews the relevant parts of binaural hearing, and adds more details to the terms introduced in chapter 1.2. Finally, Ambisonics will be reviewed in its entirety. First by explaining the concept in relation to the spatial acoustics, then proceeding to combine Ambisonics and binaural hearing in order to have a binaural representation of the final format.

2.1 Spherical acoustics

Spherical acoustics in the context of this thesis refers to being a collecting term for spatial coordinates and how acoustics can be described in three dimensions, which are to be presented in the following sections.

2.1.1 Spatial Coordinate Systems

The position of a point in a three dimensional space can be described by spherical, cylindrical or Cartesian coordinates, with spherical being the most appropriate choice for this thesis, as angular resolution is one of the points that will be investigated. A spherical coordinate system describes the position in terms of direction and distance to the origin. For obvious reasons when discussing spatial hearing, the origin is specified at the head center, leveled such that the entrances of the two ear canals lies along one of the axes, as shown in figure 2.1.

With the head set, the sound source position can be described exactly in space by the the directional vector pointing from the origin. To relate the sound source directions to the different planes of the body, three different planes are defined by the directional vectors: Two vectors pointing to the front and either direction defines the *horizontal plane*, two vectors pointing to the front and up defines the *median plane* and two vectors pointing

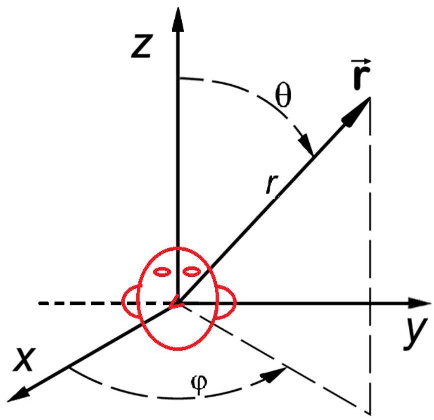


Figure 2.1: Spherical coordinate system showing how the head is centered. Figure from [23].

from the top to either direction defines the *frontal plane*. All of the planes are shown in figure 2.2.

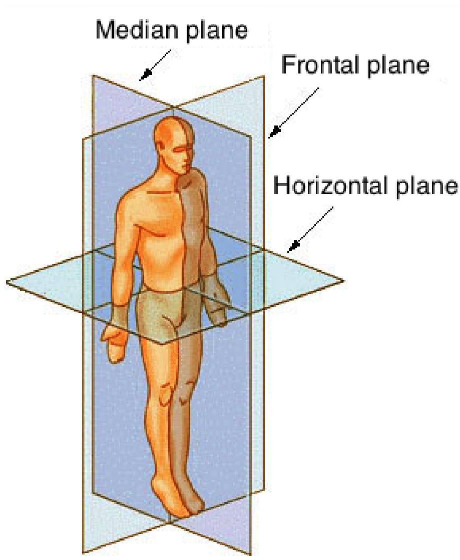


Figure 2.2: The three different body planes made up from the directional vectors in three dimensions. Other names for the different planes are *coronal*, *sagittal* and *transversal*, respectively for the frontal, median and horizontal planes in the figure. Figure from [24].

The coordinate system most commonly used in related research is referred to by Algazi et al. [25] as an *interaural-polar coordinate system*, which is shown in a counter-clockwise version in figure 2.3. The sound source position is described by (r, ϕ, θ), where the source distance with respect to the origin is denoted r with $0 \leq r < \infty$. ϕ defines the azimuth angle between the horizontal projection of the directional vector and the median plane, with $0 \leq \phi < 360°$, such that azimuth angles of $0°$, 90^o,180^o and 270^o represents the front, left, back and right directions of the horizontal plane. θ defines the angle between the horizontal plane and the directional vector, with $-90° \leq \theta \leq +90°$, where $-90°$ and $+90°$ represents the bottom and top positions. Some of the literature also refers to the clockwise version of this, where the azimuth is defined positively in the opposite direction, such that $\phi = +90°$ is to the right of the head.

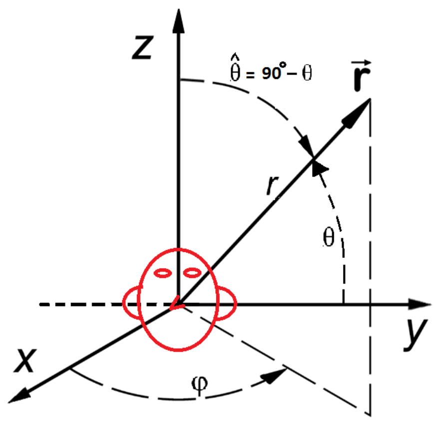


Figure 2.3: Interaural-polar coordinate system. Figure from [23].

Another instance of the spherical coordinate system is the system described referred to by Algazi et al. [25] as a *vertical-polar coordinate system*. This system has the same definition for the azimuth angle, and also appears with either clockwise or counter clockwise positive directions for the azimuth. The difference is the definition of the elevation angle, where θ for this system is defined in the range $0 \leq \theta \leq 180$, where $0°$ and $180°$ respectively represents the top and bottom position as in figure 2.1. The reason for defining two coordinate systems is that the first is the most used in literature when referring to HRTF datasets, which will be discussed further in 2.2.2, whereas the latter corresponds to the angular definitions used to describe spherical harmonics. Note that azimuth and elevation in figure 2.1 and 2.3 are denoted by respectively ϕ and θ, as opposed to the opposite which is found in most literature. This is commonly used in physics, and particularly with spherical harmonics. In accordance with the derivations and definitions from Williams [26, pp.680-685 and pp.698-700], these are the conventions that will be used henceforth.

2.1.2 Acoustics in three dimensions

The acoustic wave equation has a known solution for spherical coordinates, which enables that spherical sound fields can be decomposed in a relatively simple manner by using what is called *spherical harmonics*. This theory section will provide a sufficiently detailed introduction to the topic, such that the reader can follow through the rest of the report. A complete description of spherical acoustics and harmonics is provided by Williams [27, Chapter 6].

2.1.3 Spherical harmonics

Spherical harmonics are special functions that define the surface of a sphere. They are of significant importance in spherical acoustics, and according to Williams [27, p.190-192], can *any arbitrary function on a sphere* $g(\phi, \theta)$ be expanded in terms of spherical harmonics as

$$g(\theta, \phi) \sum_{n=0}^{\infty} \sum_{m=-n}^{n} A_{nm} Y_n^m(\theta, \phi), \tag{2.1}$$

where A_{nm} are complex constants and the spherical harmonics are defined as

$$Y_n^m(\theta, \phi) \equiv \sqrt{\frac{2n+1}{4\pi} \frac{(n-m)!}{(n+m)!}} P_n^m(cos\theta) e^{im\phi}. \tag{2.2}$$

P_n^m are the associated *Legendre polynomials*. These functions are all orthonormal, thus the arbitrary constants can be found from

$$A_{nm} \equiv \int d\Omega Y_n^m(\theta, \phi)^* g(\theta, \phi), \tag{2.3}$$

where the solid angle Ω is defined by

$$\int d\Omega \equiv \int_0^{2\pi} d\phi \int_0^{\pi} sin\theta d\theta. \tag{2.4}$$

For real-valued audio signals, the real-valued spherical harmonics are of particular interest, and one way to define these is suggested by Daniels [10][1] are

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

where δ_{nm} is the Kronecker delta function, which has the values

$$\delta_{nm} = \begin{cases} 1 \text{ for } n = m \\ 0 \text{ for } n \neq m \end{cases} \tag{2.6}$$

To get a visual understanding of what the spherical harmonics actually *looks like*, and how they can represent a spatial sound field, one can study the directivity of the spherical harmonics, as illustrated for the first orders in figure 2.4. One can especially take note of Spherical harmonics of first order, which currently is the most used order, due to limitations in most of today‘s software.

[1] The order of n and m in the related equation are interchanged to match the definitions in eq. 2.1 to 2.3

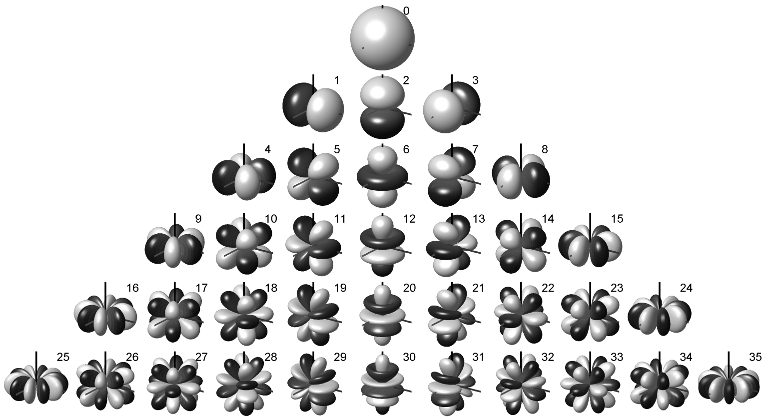


Figure 2.4: The directivity of spherical harmonics from the zeroth to forth order. The light and dark areas represents respectively positive and negative pressures. Figure from [28].

2.2 Binaural hearing

Most of the parameters that are to be evaluated later on are related to the acoustic phenomenons of binaural hearing, thus a review of the human hearing is needed. Binaural hearing refers to hearing with two ears instead of one, which would be *monoaural* hearing, and it offers significant advantages over monoaural hearing, including of sound localization and a perceptual feeling of the surroundings. The human auditory system is incredible complex, and has a lot of interesting features, but in order to keep it concise, only the key components of sound localization will be reviewed in this section.

2.2.1 Sound localization

There are a few prominent cues that enables us to determine the position of a sound source in terms of both distance and location relative to the listener. The important cues in the horizontal plane, also called lateralization cues, are *interaural time difference (ITD)* and *interaural level difference (ILD)*, while *dynamic* and *spectral* cues are the most important cues along the median plane. A lot of previous work is summarized in chapter 1.2, but a great and far more detailed summary of these cues could attained from either Blauert's classical work "Spatial Hearing" [29] or Xie's work on virtual auditory display [30, Ch.1]. The most relevant aspects are review in the following subsections.

2.2.1.1 Directional cues

Interaural Time Difference (ITD) ITD refers to the difference in time which the sound waves arrive at the left and right ear. When the source(s) deviate from the median plane and are distributed along the horizontal plane, the path lengths to the ear becomes different and the ITD becomes non-zero, as shown in figure 2.5. The ITD can be estimated for one quadrant from the Woodworth formula [30, p.9] as

$$ITD(\phi) = \frac{a}{c}(sin(\phi) + \phi), \tag{2.7}$$

where c is the speed of sound and a is the head radius. For the clockwise system in this example, a positive ITD here indicites a source closer to the right ear than the left. The ITD is shown in figure 2.6.

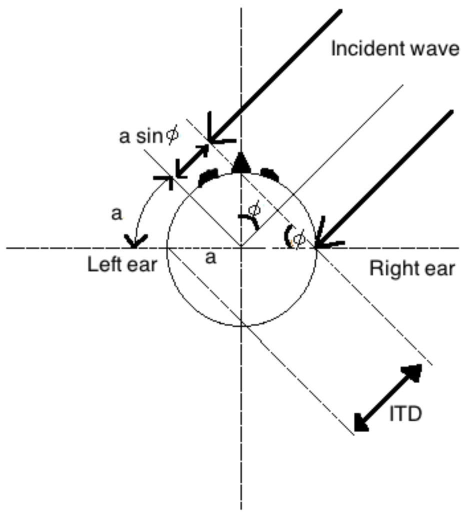


Figure 2.5: ITD calculation including the curved surface of the head

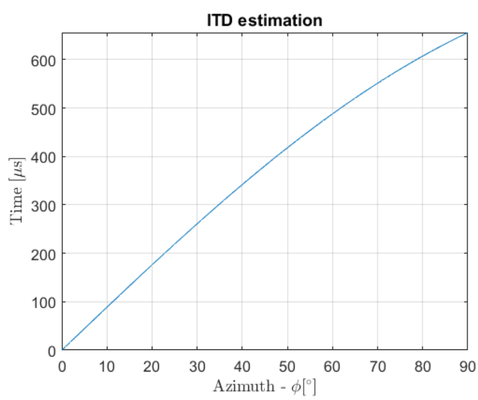


Figure 2.6: ITD for a source moving from 0 degrees azimuth to 90 degrees

Interaural phase delay difference Blauert also showed [29] that another important cue for localization is the *interaural phase delay difference*, which can thought of as the frequency domain version of the ITD, and it is prominent under 1.5 kHz where the wave length approximately equals the diameter of the head. The ITD_p derived from the interaural phase delay is defined by Xie as [30, p.10] as

$$ITD_p(\phi, f) = \frac{\Delta\Psi}{2\pi f} = -\frac{\Psi_L - \Psi_R}{2\pi f}, \tag{2.8}$$

where p is the phase delay, Ψ_L and Ψ_R are the sound pressure phases of the left and right ear, respectively, and the interaural phase difference between the ears are given by $\Delta\Psi = (\Psi_R - \Psi_L)$. Ψ_L and Ψ_R can be found as

$$\Psi_L = \arg[H_L(f)] = \arctan\left[\frac{\mathrm{Imag}\big(H_L(f)\big)}{\mathrm{Real}\big(H_L(f)\big)}\right] \tag{2.9}$$

and

$$\Psi_R = \arg[H_R(f)] = \arctan\left[\frac{\mathrm{Imag}\big(H_R(f)\big)}{\mathrm{Real}\big(H_R(f)\big)}\right] \tag{2.10}$$

Assuming that the time factor of a sinusoidal wave equals $exp(j2\pi ft)$, $\Delta\Psi > 0$ corresponds to leading in phase, and the opposite case where $\Delta\Psi < 0$ corresponds to lagging in phase. This cue is most accurate below 700 kHz. Around ~700Hz, the (average) path difference will equal half a wave length. This leads to a degree of ambiguity, since the sounds pressures at the two ears are out of phase. For $f \geq 1.5$ kHz the path difference between the ears becomes larger than the wave length. For these frequencies and above, the absolute value of the ITD_p may exceed 2π, making the cue completely ambiguous, thus giving no information.

Interaural Level Difference (ILD) ILD is the most dominant lateralization cue for higher frequencies. When a sound source deviates from the median plane, and the wave lengths are no longer comparable with the head size, the sound waves will be attenuated at the farther ear due to the shadowing effect of the head. The ILD is also strongly dependent on the azimuth angle, and can reach level differences up to 10-20 db, meaning that ILD is both dependent on frequency and direction.

$$ILD(r, \phi, \theta, f) = 20\log_{10}\left|\frac{P_R(r, \phi, \theta, f)}{P_L(r, \phi, \theta, f)}\right| (dB) \tag{2.11}$$

Spectral bands and dynamic cues For sources distributed along the median plane, both the ITD and ILD are near zero, making the spectral cue a kind of monaural cue. The spectral cues are formed in the folds of the pinnae, by giving relative delays between the incoming sound waves, depending on which elevation they hit the pinnae from. The spectral cues are also suggested by Blauert [29] to be a feature which could be built up empirically, by making a visual confirmation of the position of the sound source. Lastly, dynamic cues are a combination of all the aforementioned cues. When neither the spectral or lateral cues reveal the position, one could simply tilt or rotate the head in order to process the same sound from new angles, giving new information.

2.2.2 Head related transfer functions (HRTF)

As implied by the previous section, sound waves go through quite a lot of filtering before reaching the ears. The filtering is due to the sound waves interacting with different anatomical structures, such as the head, torso and pinnae. The interaction can take place in form of diffraction around the head for sound waves of similar wavelengths, or from reflections from the torso and pinnae. The resulting binaural sound pressure will then contain all of the localization cues reviewed in the former subsections, and finally the auditory system utilizes these in order to localize the source. One can look at the process of transmission from one point to each of the two ears, as a linear time-invariant (LTI) process, assuming that the head position is fixed. A *head related transfer function (HRTF)* will then completely describe the overall filtering imposed by the anatomical structures of the individual subject. The HRTFs are the acoustical transfer functions (from time to frequency domain) of the LTI process. For an arbitrary source position, a pair of HRTFs H_L and H_R for the left and right ear respectively can then be defined as:

$$\begin{aligned} H_L = H_L(r,\phi,\theta,f,a) &= \frac{P_L(r,\phi,\theta,f,a)}{P_0(r,f)}, \\ H_R = H_R(r,\phi,\theta,f,a) &= \frac{P_R(r,\phi,\theta,f,a)}{P_0(r,f)}, \end{aligned} \tag{2.12}$$

where r is the distance from the sound source to the individual ears, a is a variable denoting the individual uniqueness of the anatomical structures and P_L and P_R are the complex-valued sound pressures in the frequency domain at the left and right ear, respectively. P_0 is the free field complex-valued sound pressure, also in the frequency domain. The valid position of P_0 is in the center of the head, but with the head assumed physically absent. An illustration of two head-related impulse responses (HRIR) $h_L(t)$ and $h_R(t)$, which are the time-domain equivalent of the HRTFs connected by the Fourier transform[30, p.44], are shown in figure 2.7

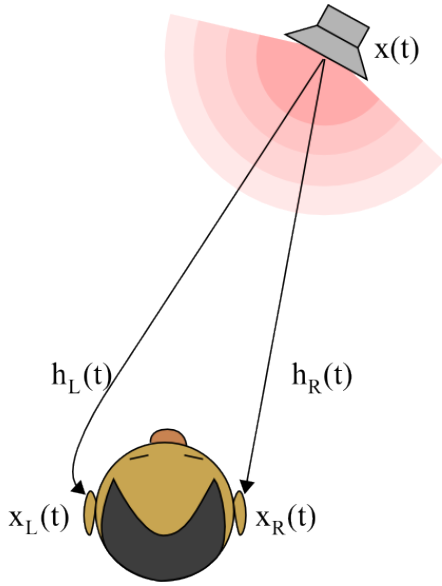


Figure 2.7: A time-varying signal x(t) passes a room and is filtered by $h_L(t)$ and $h_R(t)$ before reaching the inner ear as $x_L(t)$ and $x_R(t)$. Figure from [31].

Though eq. 2.26 is a multi-variable function, the distance r can be omitted, since the experiments will only include *far field*-HRTFs. The HRTF is proven by Xie et al. [30, pp.114-116] to be approximately distance-independent, and far field-HRTFs will be used exclusively, due to a larger availability of open databases containing exactly these.

2.3 Ambisonics

Michael Gerzon introduced the concepts leading up to Ambisoncs back in the 70s. His aim was to improve the current quadraphonic sound quality, and he suggested the *ABCD-formats* [8], which are later recognized as the formats used to design a first order Ambisonic encoder/decoder. In these terms, the A-format represents the raw form of which the four channel signal is derived (i.e. how it is recorded, taped, etc.), the B-format represents the channel signal as to be used within studio processing, C-format ('coded' format) represents how the signal is conveyed on disc, tape or radio via 2-4 channels, and lastly the D-format ('decoded' format) represents the set of signals fed to the listener's loudspeakers in order to provide the correct, subjective effect in the listening area. Moreover, original (1st order) Ambisonics, the B-format, consisted of 4 components: The omnidirective **W**-channel, which has constant signal, and the bi-directive **XYZ**-channels, which produce the polar patterns of three figure-of-eight microphones respectively in the front-back, left-right and up-down directions. 4 microphone capsules in a tetrahedral arrangement can in this way provide 1st order Ambisonics, and there are several commercially available microphones on the market today, for instance from Sennheiser [32].

In this section, the fundamentals of Ambisonics will be reviewed, and its relevant parts for the goals of this thesis. For a full review on both 3D-sound fields and Ambisonics, [3, Daniel et al.], [33, 34, Poletti], and [35, Nicol, Emerit] can be recommended.

2.3.1 The ambisonic approach

Daniel and Nicol et al. showed [3, 35] that Ambisonics is asymptotically holographic. According to holographic theory, any sound field can be expressed as a superposition of plane waves. The Kirchoff-Helmholtz integral relates the pressure inside a source free volume of space to the pressure and velocity on the boundary at the surface. Hence, reproduction of the original sound field is possible, by using an infinite amount of loudspeakers arranged on a closed contour, where the loudspeakers signals are assumed to be plane waves. However, since an infinite amount of loudspeakers generating exclusively plane waves contradicts reality, some error is introduced by this approach. Instead, a very good approximation of the original sound field can be synthesized by using a finite number of speakers arranged on said sphere, for a finite area known as *sweet spot*. The order of Ambisonics used defines the lower limit of transmit channels (i.e. the amount of speakers), as well as an upper limit to which the sound field is expected to be reconstructed correctly. The amount of transmit channels and Ambisonics order are related by $M = (N+1)^2$, where M is the amount of speakers/transmit channels and N is the Ambisonics order. The frequency limit for which accurate reproduction is possible up to, has been suggested by Bertet et al. [36] to be

$$f_{max} = \frac{Nc}{2\pi r}, \tag{2.13}$$

while Gumerov and Duraiswami [37] suggests

$$f_{max} = \frac{(2N-1)c}{e2\pi r}, \tag{2.14}$$

where r is the radius of the sweet spot, c the speed of sound[2] and e is base of the natural logarithm. The latter definition gives a reduction for the frequency limit, particularly for the orders 1-3. In figure 2.8, one can see that for N=1, f_{max} is reduced from $\sim$550 Hz to just 200 Hz, which is practically useless if this concurs with the results. The distance between the two limits do however converge to a constant distant quite quick. Hence, one can expect a more accurate representation under the lowest f_{max}, and look at the Δf_{max} area like a transition band.

There are two main methods generating HOA-signals: Actual recording of real sources, and placing virtual sources in space. The first method requires specialized equipment such as the aforementioned microphone. To capture sound fields and higher order Ambisonics requires a spherical microphone array[3] in the center of the coordinate system. The output of this array is then encoded the HOA-format by using spherical harmonics. As of spring 2017, microphones capable of recording Ambisonics up to 4th order are available to the public by M. H. Acoustics [38].

[2]Numerical value here: c= 343 m/s.

[3]Could also be a an array of pressure sensors scattered at a single, rigid sphere or grid spherical grid.

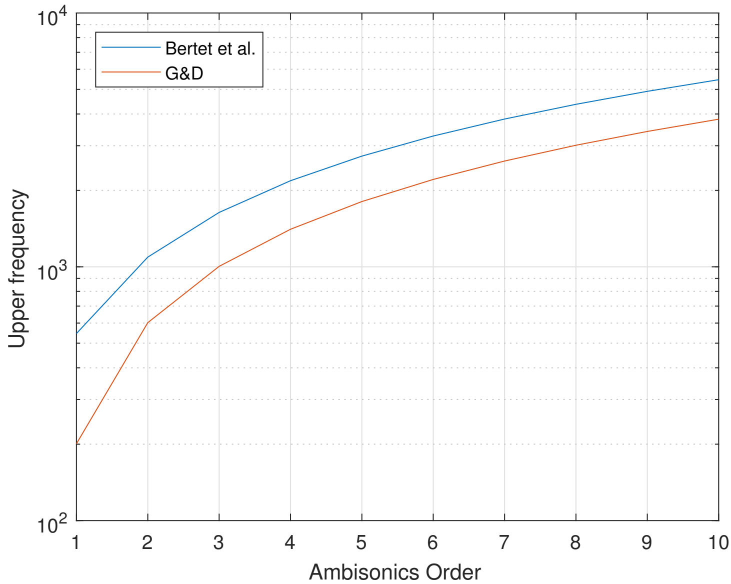


Figure 2.8: Difference between upper frequency definitions

The second way, which this thesis is focusing on, is to place virtual sources in space in relation to some coordinate system and origin. The source is placed at a position (r, ϕ, θ), and then directly encoded with spherical harmonics. This method is able to encode both plane waves (under the far field assumption) as well as spherical waves (near field).

2.3.2 Encoding and decoding virtual sources

Deriving Ambisonics from the homogeneous wave equation as done by Daniel [3] and Noisternig et al. [19],

$$\Delta p(t, \boldsymbol{r}) - \frac{1}{c^2} \frac{\partial}{\partial t^2} p(t, r) = 0, \tag{2.15}$$

where $p(t, r)$ is the sound pressure at position r and c is the speed of sound, yields the matching conditions

$$s \cdot Y^{\sigma}_{m,\eta}(\Phi, \Theta) = \sum_{n=1}^{N} p_n \cdot Y^{\sigma}_{m,\eta}(\phi_n, \theta_n). \tag{2.16}$$

The left side of eq. 2.16 is the Ambisonic encoding equation, which can be written as

$$\boldsymbol{B}_{\Phi,\Theta} = \boldsymbol{Y}_{\Phi,\Theta} \cdot s, \tag{2.17}$$

where $\boldsymbol{B}_{\Phi,\Theta}$ is the ambisonic channel vector representing the signal s from direction Φ, Θ with the corresponding spherical harmonics $Y_{\Phi,\Theta}$. The right side of 2.16 represents the reconstructed signal p_n at the nth loudspeaker in direction (ϕ, θ). For further simplifications in this thesis, the signal to be encoded will just be a vector of ones, $s = (1, 1, ..., 1)$, such that $B_{\Phi,\Theta} = Y_{\Phi,\Theta}$.

By defining

$$\boldsymbol{p} = [p_1, p_2, ..., p_N]^T \tag{2.18}$$

as the loudspeaker signal vector and

$$\boldsymbol{B} = [Y^1_{0,0}(\Phi, \Theta), ..., Y^{-1}_{M,M}(\Phi, \Theta)] \tag{2.19}$$

as the Ambisonics signal vector, eq. 2.16 can be rewritten as

$$\boldsymbol{B} = \boldsymbol{C} \cdot \boldsymbol{p}. \tag{2.20}$$

$\boldsymbol{C}$ is now a matrix that contains spherical harmonics which allows the reverse operation of eq. 2.17, namely encoding the loudspeaker signal into Ambisonics. Thus, the decoder can be calculated from the encoder as

$$\boldsymbol{D} = pinv(\boldsymbol{C}) = \boldsymbol{C}^T \cdot (\boldsymbol{C} \cdot \boldsymbol{C^T})^{-1}, \tag{2.21}$$

where $pinv(\boldsymbol{C})$ is the Moore-Penrose pseudo-inverse[4] of the matrix $\boldsymbol{C}$. This procedure highlights one of the main features of the Ambisonics format, a decoupling of the encoder and decoder. This means that the only focus in the encoding stage will be on the universal multi-channel format, and all the attention in the playback configuration can be dedicated towards the decoder. Moreover, this also implies that the number of loudspeakers is independent of the number of sources that are encoded. The minimum amount of loudspeakers, M, for such a 3D audio system is limited by the amount of transmit channels L, as previously shown as

$$M \geq L = (N + 1)^2. \tag{2.22}$$

Now, with the decoding matrix $\boldsymbol{D}$ being the pseudo-inverse of $\boldsymbol{C}$, one can multiply both sides of eq. 2.20 with eq. 2.21, which leads to

$$\boldsymbol{B} \cdot \boldsymbol{D} = \boldsymbol{C} \cdot pinv(\boldsymbol{C}) \cdot \boldsymbol{p} \Longrightarrow \boldsymbol{B} \cdot \boldsymbol{D} = \boldsymbol{p} \tag{2.23}$$

2.3.3 Ambisonics in a binaural sound reproduction

As explained in section 2.2.2, a pair of HRTFs for the left and right ear, respectively H_L and H_R, can be viewed as two ordinary filters in a LTI system. This enables to split a potentially very large and possibly unfeasible loudspeaker setup into two signals, respectively for the left and right ear. In frequency domain, this is simply done by multiplying either the decoding matrix $\boldsymbol{D}$ or the sound pressures $\boldsymbol{p}$ with the HRTFs. In this thesis, a left and right decoding matrix as been used, as shown in figure 2.9, in order to get a sound pressure matrix for each of the ears.

To study in more detail how errors are induced in the binaural signals by using HOA of finite orders, one possible way is to reconstruct the HRTFs from the binaural signals $\boldsymbol{p_L}$ and $\boldsymbol{p_R}$. Since the error in HRTF reproduction is introduced in the decoding step, and $\boldsymbol{p}$

[4] https://se.mathworks.com/help/matlab/ref/pinv.html

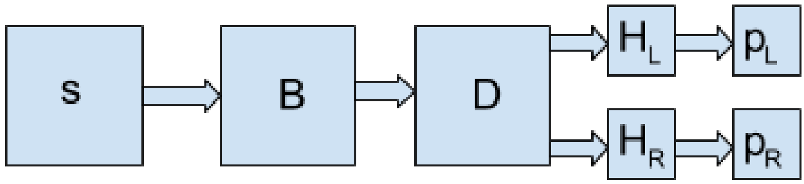


Figure 2.9: Block diagram showing how a signal *s*, whether it is real or virtual sources, can be encoded into the Ambisonics format *B*, then decoded with *D* into a given loudspeaker setup , and finally filtered into two binaural signals for the left and right ear with the respective HRTFs H_L and H_R.

can be found independent of the HRTFs, one can find the recreated HRTF signals $\hat{\boldsymbol{H}}_{\boldsymbol{L}}$ and $\hat{\boldsymbol{H}}_{\boldsymbol{R}}$ by in two steps. First, re-arranging the equation for $\boldsymbol{p_L}$ as

$$\begin{aligned} \boldsymbol{p_L} &= \boldsymbol{B} \cdot \boldsymbol{D_L} = \boldsymbol{B} \cdot \boldsymbol{D} \cdot \boldsymbol{H_L} = \boldsymbol{p} \cdot \boldsymbol{H_L}, \\ \boldsymbol{p_R} &= \boldsymbol{B} \cdot \boldsymbol{D_R} = \boldsymbol{B} \cdot \boldsymbol{D} \cdot \boldsymbol{H_R} = \boldsymbol{p} \cdot \boldsymbol{H_R}, \end{aligned} \tag{2.24}$$

and then using the pseudo-inverse in reverse which yields

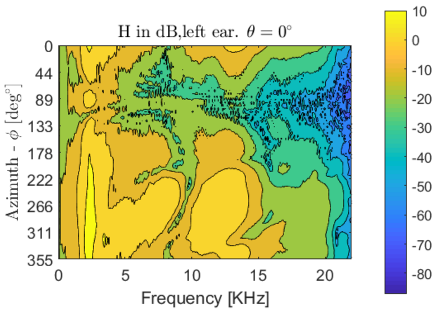
$$\begin{aligned} \hat{\boldsymbol{H}}_{\boldsymbol{L}} &= pinv(\boldsymbol{p}) \cdot \boldsymbol{p_L}, \\ \hat{\boldsymbol{H}}_{\boldsymbol{R}} &= pinv(\boldsymbol{p}) \cdot \boldsymbol{p_R}, \end{aligned} \tag{2.25}$$

The reconstructed HRTF can now be compared to its reference by plotting both datasets and visually inspecting the difference. To further simplify this comparison, difference plots showing the error can be added as well. These differences are simply found by

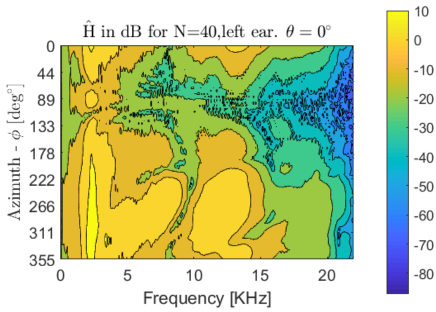
$$\begin{aligned} \boldsymbol{\Delta H_L} &= \hat{\boldsymbol{H}}_{\boldsymbol{L}} - \boldsymbol{H_L}, \\ \boldsymbol{\Delta H_R} &= \hat{\boldsymbol{H}}_{\boldsymbol{R}} - \boldsymbol{H_R}, \end{aligned} \tag{2.26}$$

To prove this relation between the different components, one would expect a perfect reconstruction from a very high Ambisonics order. According to eq. 2.14, an Ambisonics order of N=40 should suffice to cover the frequency range for the investigated HRTFs (0-22 kHz), for a sweet spot of $r = 0.1\text{m}$[5]. A comparison for the horizontal plane (i.e. $\phi = 0°$ is shown in figure 2.10, where one can see that the reference and reconstructed HRTFs are identical, and that the difference between them is apparently non-existent.

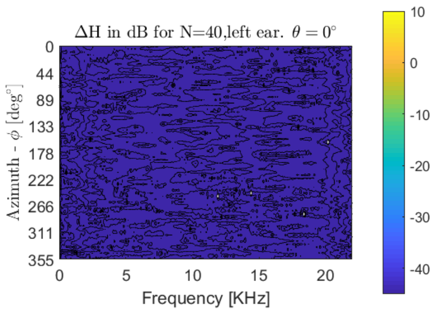
[5]I.e. the size of an average human head.



(a) Reference HRTF



(b) Reconstructed HRTF, N=40



(c) Error in dB between reference and reconstructed HRTF

Figure 2.10: Comparison between a contour plot of a reference HRTF in the top and a near-perfect reconstructed HRTF in the bottom.

Moreover, ITD_p and ILD from eqs. 2.8 and 2.11 can now be estimated and compared for both the reference and reconstructed HRTF, in order to make an error estimate. The reconstruction error estimate is the same as used by Epain et al. [21, p. 4]:

$$\Delta_{ILD} = \frac{1}{M} \sum_{i_1}^{M} |ILD_{HOA}(\phi_i, \theta_i) - ILD_{REF}(\phi_i, \theta_i)|, \tag{2.27}$$

and

$$\Delta_{ITD} = \frac{1}{M} \sum_{i_1}^{M} |ITD_{HOA}(\phi_i, \theta_i) - ITD_{REF}(\phi_i, \theta_i)|. \tag{2.28}$$

Chapter 3

Experiment

The practical part for this thesis is modelled and studied in MATLAB[1] As the goal for the study is to examine how different resolutions of HRTF-datasets behave for different orders of Ambisonics, the system is built as a console utility, where a range of parameteres can be altered in order to get the desired output.

3.1 System description

The script that is written in order to achieve the desired results, can be described as the different parts of theory explained in the previous chapter put into a systematic order, then iterated through. Some minor adjustments are done depending which of the tasks at hands, but these are all commented in the code in the digital hand-in. The results have been attained throughout implementation of the founding theory. More specific, the equations in section 2.3.2 have been implemented to construct the Ambisonic signals, which are then converted to binaural signals. Finally, the HRTF reconstructions are attempted based on the aforementioned results. The binaural cues are processed and reproduced as indicated by eqs. 2.8-2.10 and eq. 2.11.

3.1.1 Non-original code

There are two parts of the Matlab-code not written by the author. The first is a mathematical function, *gen_sph_harm_mat.m*, provided by the assistant supervisor A. Solvang, which generates the spherical harmonics. For an Ambisonics order N and a set of azimuth and elevation angles ϕ_i and θ_j, where $i = 1, 2, ..., M$ and $j = 1, 2, ..., K$, this function returns the spherical harmonics for the $(N+1)^2$ transmit channels that this order requires. There is no requirement for $K = M$, but the input parameters needs to cover every combination of azimuth and elevation angle, thus making it a natural requirement that for $K \neq M$,

[1] www.mathworks.com - Matlab is a multi-paradigm numerical computing environment and programming language.

the different combination of angles needs to be converted into matrices doing exactly this. The second part which is not written by the author, are some code snippets provided by the different instances from which the HRTF datasets are available. The HRTF datasets are stored in several different formats, and these snippets generally do is to assist in reading the data properly, such that it can be used in a meaningful way.

3.2 HRIR-datasets

Three HRIR databases have been chosen, due to their ease of access and variability in resolution. These are from MIT Media Lab [39], CIPIC [25] and the LISTEN-project [40]. Their resolution (i.e. amount of measured loudspeaker positions) ranges from 187 to 1250. A complete description of how the individual measurements have been performed are available on the respective web sites, and the key features are summarized in the following sections. All of the datasets are sampled at 44.1 kHz, which is sufficient to cover all of the dynamic range of human hearing.

3.2.1 MIT Media Lab

The MIT Media Lab-database, hereafter referred to as the MML-database, This is the oldest database (1995) of the three, but it is included due to its high amount of measured points, as well as it is subject for many references throughout the literature. Contrary to the other databases included in this thesis, the MML-database only contains HRIRs for a KEMAR. The KEMAR is fitted with two ears of different size, one large and one small, thus making it a fair assumption that the results might deviate some compared to the other datasets.

The MML-datasets are measured for a total of 710 loudspeaker positions. The distribution of measurement positions can be seen both in table 3.1 and visualized in figure 3.1. The distribution is most dense in the area $\pm 20°$, which is to be expected since all of the previous studies points to this being the most essential area for localization, particularly in the horizontal plane.

ϕ	-40	-30	-20	-10	0	10	20	30	40	50	60	70	80	90
N_ϕ	56	60	72	72	72	72	72	60	56	45	36	24	12	1

Table 3.1: Table showing the elevation angles for which the MML-dataset is measured from, and how many azimuth angles per elevation angle. Angles are given in the interaural polar-coordinate system.

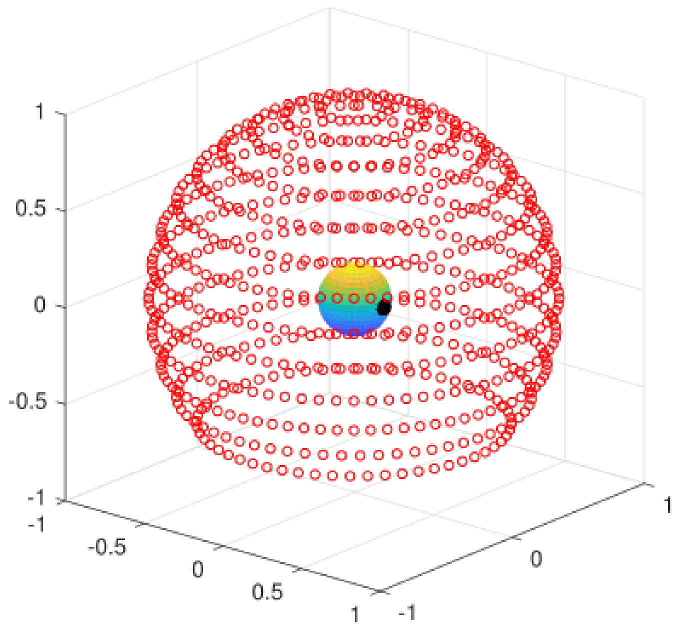


Figure 3.1: The red circles indicate the measured loudspeaker positions in the MML-database. The black dot indicates the nose of the subject, thus which direction is forward.

3.2.2 CIPIC

The CIPIC-datasets have the highest resolution of the datasets, and they are primarily made up from human subjects with the exception of two KEMAR measurements with large and small pinnae, as in the MML-datasets. It is also unique among most other available databases with HRIR-datasets, in the sense that it contains anthropometric[2] data. The anthropometric data corresponds to the variable a in eq. 2.26, and can be useful in showing correlation between the temporal and spectral features compared to individual sizes. 3.2

The CIPIC-datasets are made up by uniform elevation sampling in $360/64 = 5.625°$ steps from $-45°$ to $+230.625$, i.e. from front to the back of the head. For every elevation, azimuths were sampled at $\pm 80°$, $\pm 65°$, $\pm 55°$, and from $-45°$ to $45°$ in steps of $5°$. This makes a total of 1250 measurement points, and the distribution can be seen in figure 3.2.

[2]Anthropometric refers to measurements of a human invididual

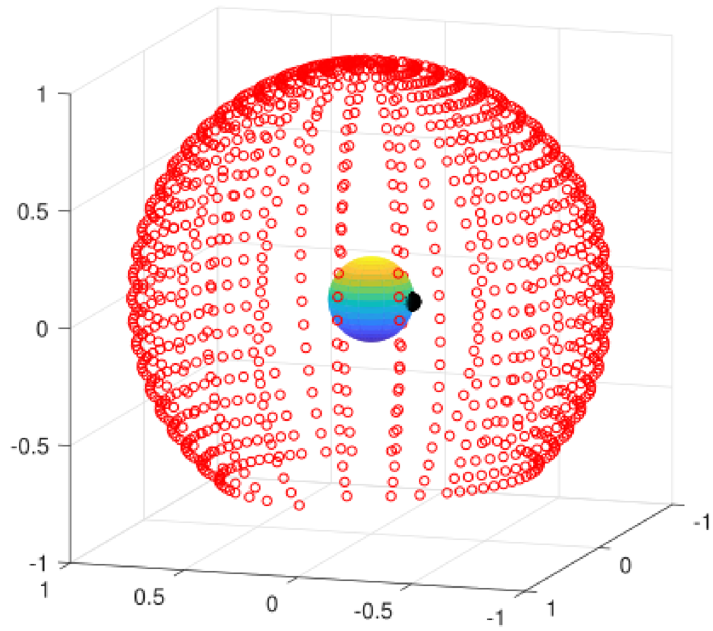


Figure 3.2: The red circles indicate the measured loudspeaker positions in the CIPIC-database. The black dot indicates the nose of the subject, thus which direction is forward.

3.2.3 LISTEN

A total of 187 loudspeaker positions makes up the LISTEN-datasets, giving it the lowest resolution. According to the literature study in section 1.2, one could argue that these datasets have such a low resolution that they are even not worth including. Despite this, since one of the goals for this thesis is to look into HOA-induced errors, they are included to give a range from low to high resolution datasets. The distribution of measurement positions can be seen both in table 3.2 and visualized in figure 3.3.

ϕ	-45	-30	-15	0	15	30	45	60	75	90
N_ϕ	24	24	24	24	24	24	24	12	6	1

Table 3.2: Table showing the elevation angles for which the LISTEN-dataset is measured from, and how many azimuth angles per elevation angle. Angles are given in the interaural polar-coordinate system.

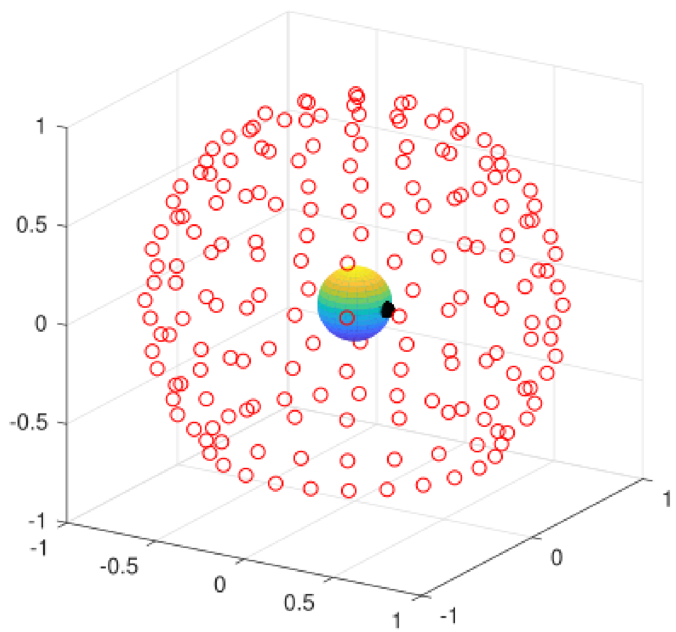


Figure 3.3: The red circles indicate the measured loudspeaker positions in the LISTEN-database. The black dot indicates the nose of the subject, thus which direction is forward.

Chapter 4

Results and analysis

The results are presented in two sections. The ability to reconstruct the reference for increasing Ambisonics order is reviewed first, and the quality is assessed by visual comparison. This method is partially both a subjective and objective evaluation. It is objective in the sense that the plots are based on actual data, and at the same time subjective to some extent, due to the nature of different readers having different opinions regarding what the thresholds for calling a change large or small is. This comparison method is inspired by the visualization of results done by Epain et al. in [21].

Secondly, the reconstruction of the ITD_p and ILD is reviewed. This is also done by the same visual comparison, but with the addition of a numerical error estimation as explained in section 2.3.3.

4.1 General HRTF reconstruction

As indicated by figure 2.8, one can at best hope for an accurate reconstruction up to 2 kHz for N=4, when decoding with only simple mode matching. There is also a greater degree of insensitivity for the higher frequencies. For these reasons, the HRTF reconstructions are scaled down to a frequency range $0 < f \leq 3.8$ kHz.

All figures related to the CIPIC datasets are created from the HRIRs from subject 12, henceforth, this will not be repeated when further discussing the CIPIC cases. The same applies for the LISTEN cases, which are reconstructed from the HRIRs for subject 08. Moreover, all of the HRTF reconstructions are evaluated for the left ear.

Plots for the same subject are made with the same dynamic range. For the difference-plots which are also in dB, the scale has been limited to [-45dB, 10dB], to clearly show which areas are accurately reconstructed (i.e. the dark blue areas), while one can expect audible glitches in areas with ± 10 dB.

4.1.1 CIPIC

Figure 4.1 shows reference HRTF as above. The magnitudes are highest for the positive azimuths, since the CIPIC datasets are measured counter-clockwise. In figures 4.2a-4.2d, close-ups of the reconstructed HRTFs are shown, and the difference between the reference and reconstruct HRTF are shown in figure 4.3a-4.3d.

N=1 Whereas a good reconstruction was expected up to ~500Hz according to eq. 2.13, it is safe to say that this reconstruction is barely similar at the lowest frequencies, which agrees with the lower f_{max} set by eq. 2.14. It can be noted that the sampling resolution for the CIPIC datasets is fairly rough, and equals a sampling frequency step of $\Delta f_s = 220$ Hz, which can explain some of the poor reconstruction for N=1. From figure 4.3a it also seems that the performance is best in the front of the listener at $-40° < \phi < 40°$.

N=2 The general pattern is starting to form for N=2, and it resembles the reference pattern for the lowest frequencies, as well as having no or very small errors for $f \approx 500$ Hz.

N=3 The pattern becomes both more accurate, and from figure 4.3c it seems like the reconstruction is fairly accurate for $f \leq 1$ kHz. There are however still some glitches for the right side ($-80° < \phi < 0°$) at $f > 1.5$ kHz, which is to be expected because of the shadowing effect of the head.

N=4 One can see a lot of more detailed nuances from the difference plot in figure 4.3d, but the are for accurate reconstruction is not improved by much from N=3.

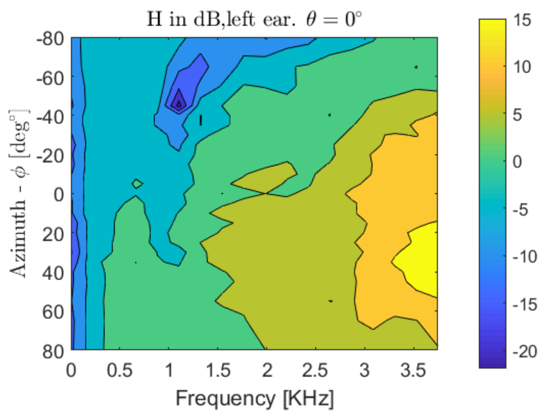
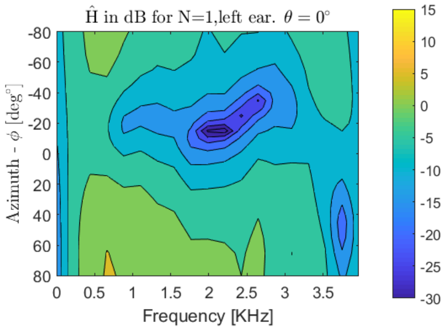
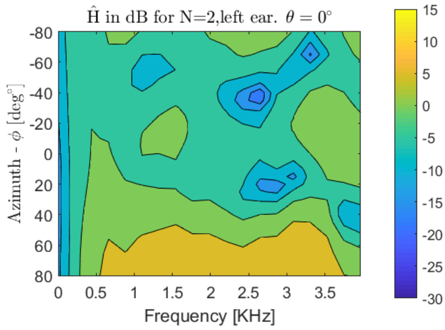


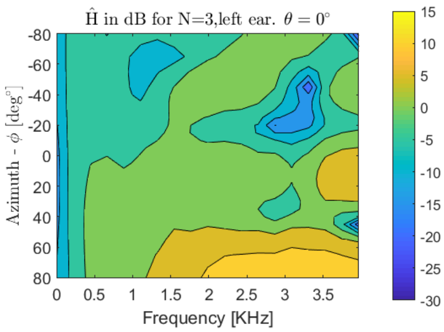
Figure 4.1: Reference HRTF for the left ear from the CIPIC datasets, with a limited frequency range.



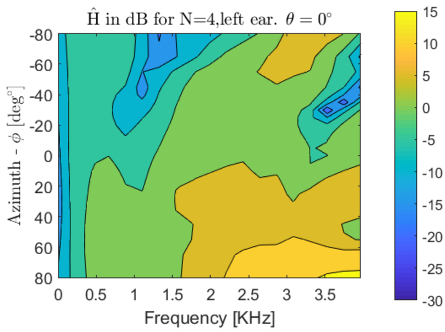
(a) Reconstructed HRTF N=1



(b) Reconstructed HRTF N=2

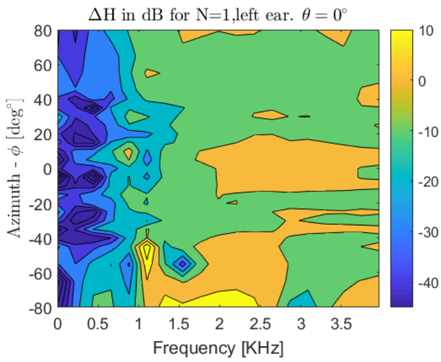


(c) Reconstructed HRTF N=3

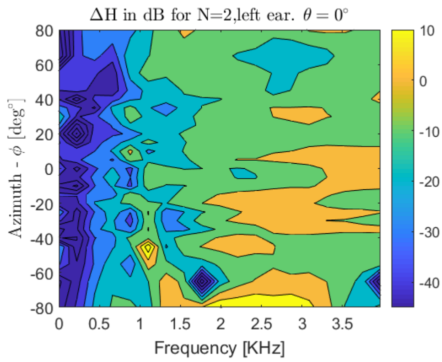


(d) Reconstructed HRTF N=4

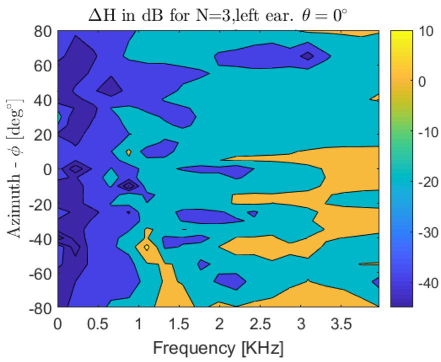
Figure 4.2: Close-up illustration of the reconstructed HRTFs for N=1-4 for the CIPIC datasets.



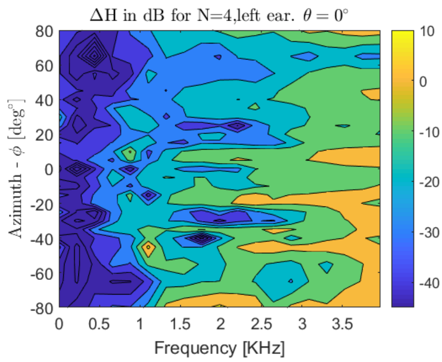
(a) Difference plot for N=1



(b) Difference plot for N=2



(c) Difference plot for N=3



(d) Difference plot for N=4

Figure 4.3: The difference plots between the reference and reconstructed HRTFs for N=1-4 for the CIPIC datasets.

4.1.2 LISTEN

The reference HRTF is shown in figure 4.4, the different reconstructions of order N=1-4 are shown in figures 4.5a-4.5d and the difference plots in figures 4.6a-4.6d. What is immediately clear for the LISTEN reconstructions, is that it performs a lot better compared to the reference than the CIPIC. Undoubtedly, the main reason for this is that the angular resolution is almost 7 times lower. The sampling frequency step is also smaller, $\Delta f_s = 86$ Hz, allowing more details along the x-axis, while the angular step is now increased to $\Delta\phi = 15°$. Moreover, the LISTEN-cases performs better than expected by both of the suggested frequency limits. However, as the frequency increase one can see the an area of ambiguity recurring on the right side of the subject, most likely caused by the shadowing effect of the head.

N=1 Contrary to the CIPIC-case, one can already see the general pattern from the reference figure for N=1. It resembles its reference accurately up to $f = 500$ Hz, and the difference plot in figure 4.6a also indicates good performance up to $f = 1$ kHz for the left side

N=2 The accurate reconstruction up to $f = 1$ kHz covers all of the azimuth angles now, and the performance is especially good at $\phi = 120°$. This angle is in the area associated with the cone of confusion, and there is no obvious reasons as to why the peak stands out.

N=3 In the transition from N=2 to N=3, the performance is improved primarily for the left side, while the differences for the right side remains almost unchanged. The peak appearing for N=2 is also widened.

N=4 Comparing the reconstructed HRTF to the reference, there does not seem to be any significant improvement from N=3. The difference plot does on the other hand indicate a stronger performance in the area $30° < \phi < 210°$.

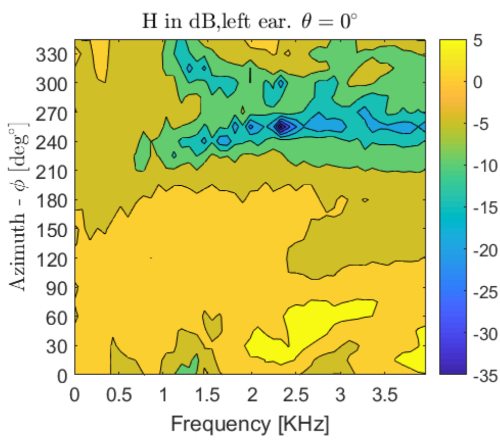
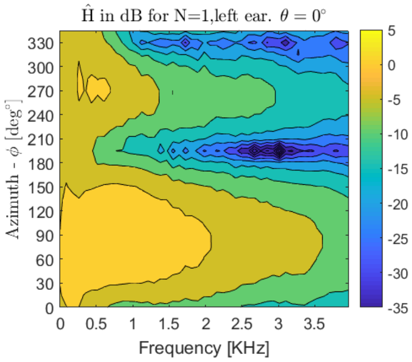
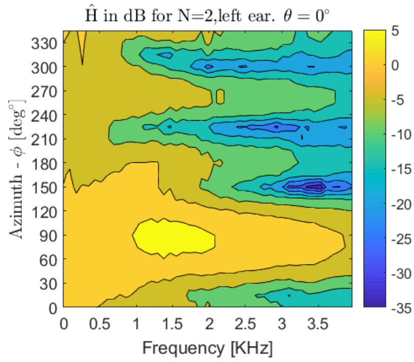


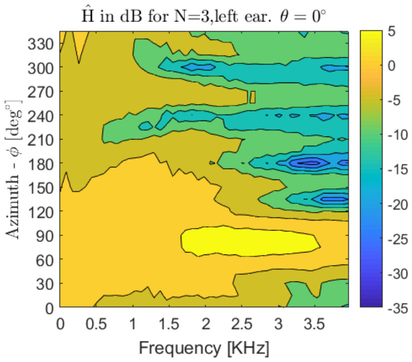
Figure 4.4: Reference HRTF for the left ear from the LISTEN datasets, with a limited frequency range.



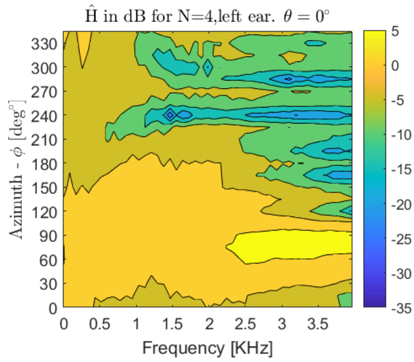
(a) Reconstructed HRTF, N=1



(b) Reconstructed HRTF, N=2

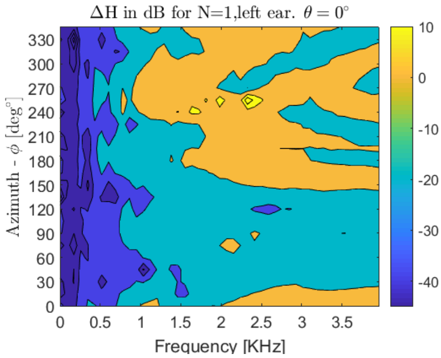


(c) Reconstructed HRTF, N=3

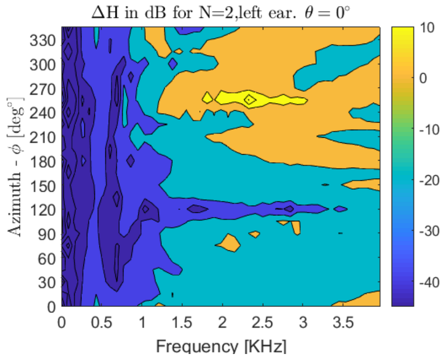


(d) Reconstructed HRTF, N=4

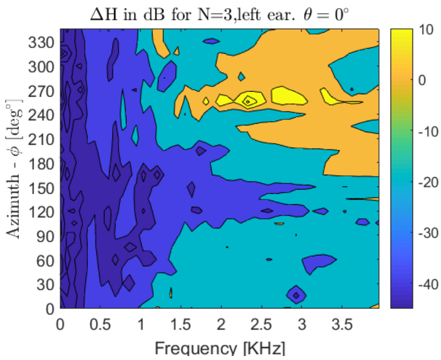
Figure 4.5: Reconstructed HRTFs for N=1-4 for the LISTEN datasets.



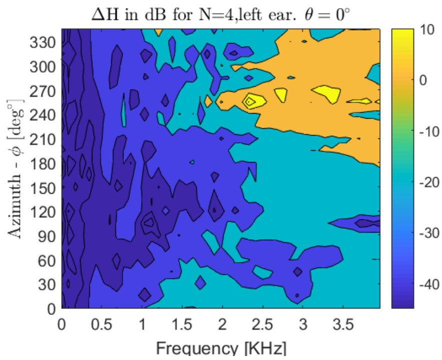
(a) Difference plot for N=1



(b) Difference plot for N=2



(c) Difference plot for N=3



(d) Difference plot for N=4

Figure 4.6: The difference plots between the reference and reconstructed HRTFs for N=1-4 for the LISTEN datasets.

4.1.3 MIT Media Lab (MML)

The KEMAR is equipped with pinnaes of two different sizes, whereas the left represents the ”normal” sized pinnae. As can be seen from the figures that follows, the MML has uses a clockwise interaural coordinate system for measuring, meaning that 0°, 90°, 180° and 270° are in front, to the right, back and to the left of the KEMAR, such that the strongest responses are located around $\phi \approx 270°$. The reference HRTF is given in figure 4.7, the reconstructions in figures 4.8a-4.8d and the difference plots in figures 4.9a-4.9d.

The MML-dataset stands out in the sense that it seems to have the same performance based on the difference plots, either good or bad, for all azimuth angles at the lower frequencies, indicating that the angular setup for the HRTF captures the information well. Another reason for this apparent stability, might be that the measurements are performed on a mannequin, which prohibits any sources of error from head movement which can be the case with the former datasets.

Comparing the reconstructions is difficult for the MML-dataset, since their patterns seems to be somewhat similar for all N. The difference plots does however reveal some more information, and the area for accurate reconstructions seems to rise with 500 Hz with every increment of N. As with the other datasets, expected artifacts can be seen to the area to the right of the KEMAR for the frequencies $f > 1.5$ kHz.

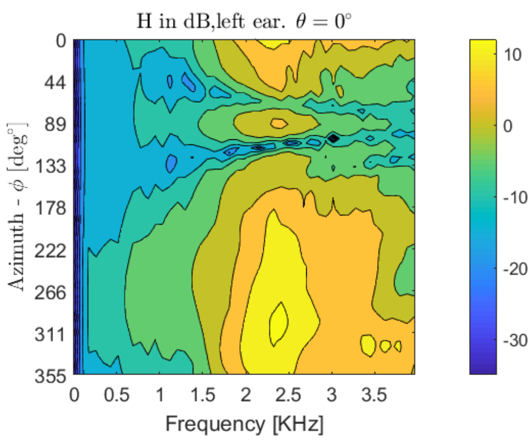
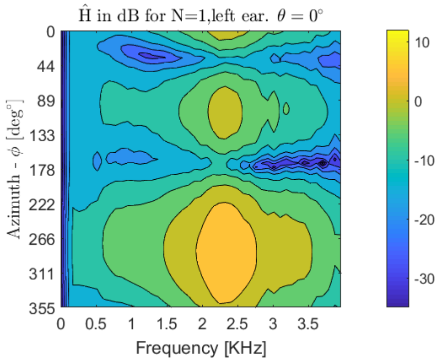
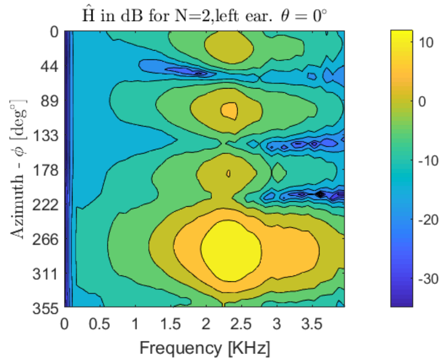


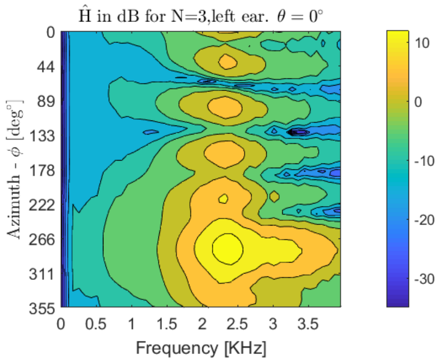
Figure 4.7: Reference HRTF for the left ear from the LISTEN datasets, with a limited frequency range.



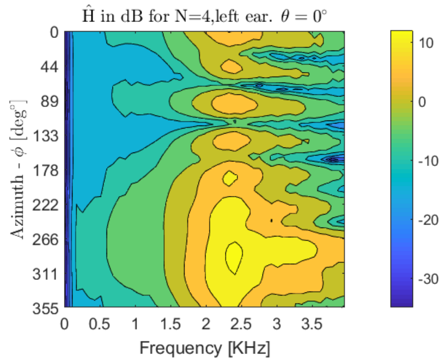
(a) Reconstructed HRTF, N=1



(b) Reconstructed HRTF, N=2

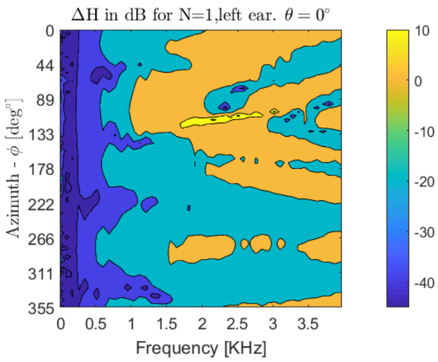


(c) Reconstructed HRTF, N=3

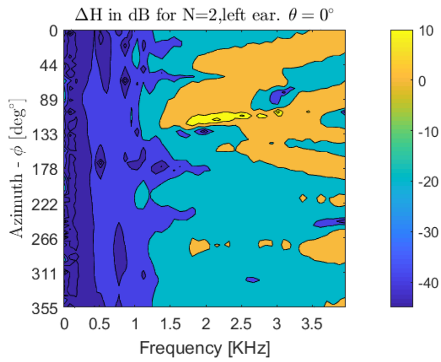


(d) Reconstructed HRTF, N=4

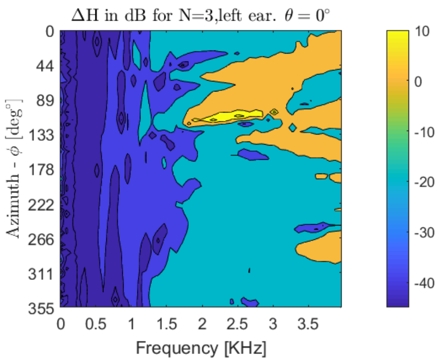
Figure 4.8: Close up illustration of the reconstructed HRTFs for N=1-4 for the MIT Media Lab datasets.



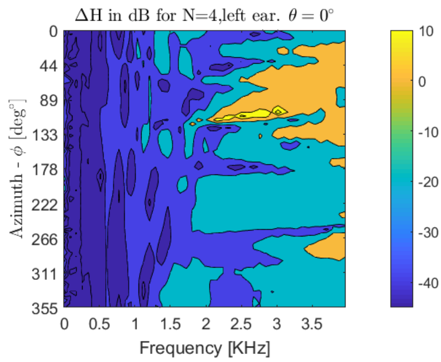
(a) Difference plot for N=1



(b) Difference plot for N=2



(c) Difference plot for N=3



(d) Difference plot for N=4

Figure 4.9: The difference plots between the reference and reconstructed HRTFs for N=1-4 for the MIT Media Lab datasets.

4.2 Reconstruction of localization cues

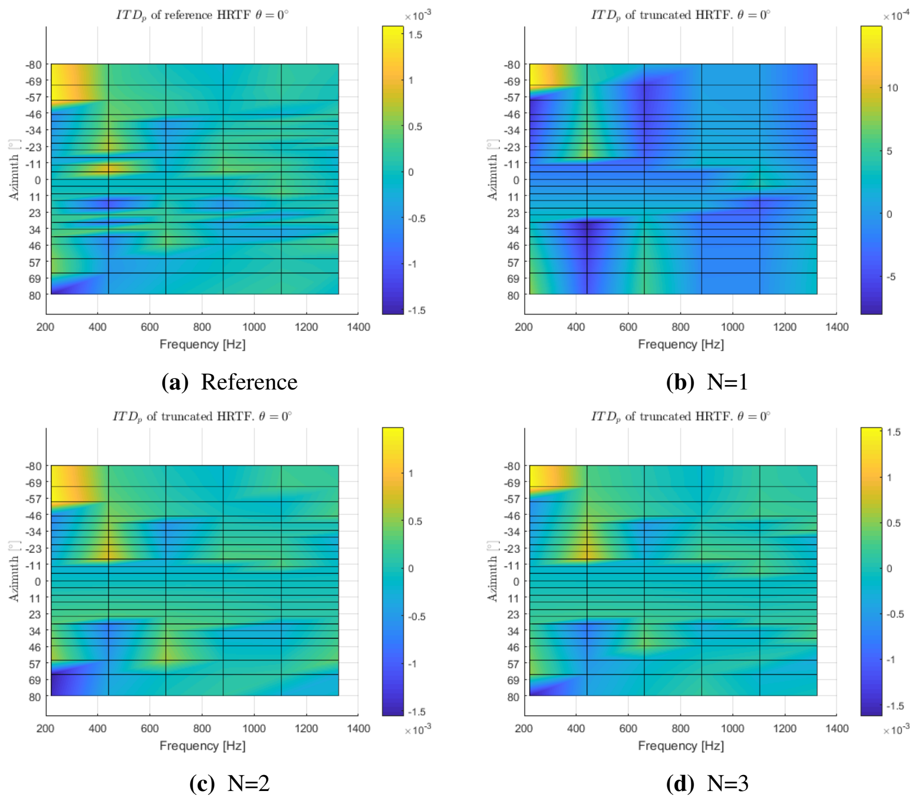
The ITD_p and ILD for the different datasets will be reviewed in this section. They are both investigated for one dataset, before moving on to the next, in the same order as the previous section. Only reconstructions up to order 3, which is expected to be accurate up to at least 1 kHz, are included. As explained in section 2.2, the ITD_p will become slightly ambiguous already at 700 Hz, hence the most interesting part is to see what happens between N=1 to N=3. Common for all of the reconstructions, is that the values for ITD_p appear to be near zero in the area near $\phi = 0°$, which is to be expected for sources in the median plane.

4.2.1 CIPIC

4.2.1.1 Interaural phase delay difference

The ITD_p for the reference is shown in figure 4.10a, whilst the reconstructions from N=1-3 are shown in figures 4.10b-4.10c[1]. For N=1, a lot of the information seems lost. The bright yellow spot for the right side ($\phi = -80°$) at f = 200 - 400 Hz is correctly reconstructed. Also the general pattern at f$\approx$ 400 Hz seem correct, though somewhat evened out. This might cause a more blurry reconstruction. For N=2, the reconstruction appears very similar to the reference, indicating that even though the general reconstruction reviewed in the previous section deviates a bit, the most important binaural cues are restored. Increasing the order further gives apparantly no new information, as the figures seems identical. Investigating the errors in reconstruction numerically, figure 4.11 shows that the phase delay difference is rather small for all orders. N=1 deviates, whereas N=2-4 follow each other quite closely. There also seems to be some harmonic relation in how it rises and sinks with frequency.

[1] Since the CIPIC has a slightly different loudspeaker setup than the two other datasets, in the sense that it has twice as many elevation angles as azimuth angles, a black grid is included in these figures to emphasize the somewhat lower horizontal resolution



(a) Reference **(b)** N=1

(c) N=2 **(d)** N=3

Figure 4.10: Interaural phase delay difference shown for the reference CIPIC HRTF in a, then for N=1-3 in b-d, respectively.

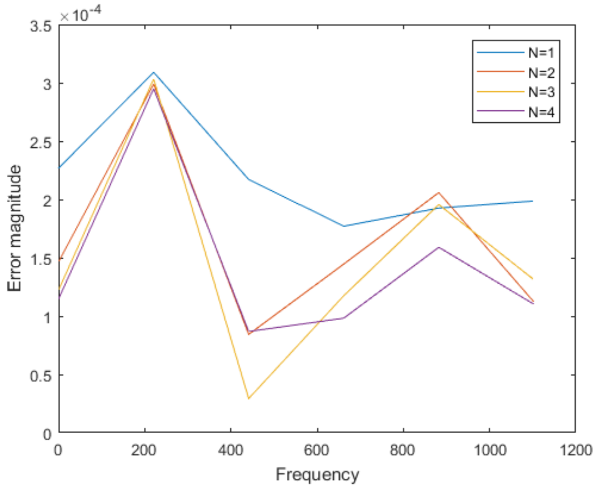
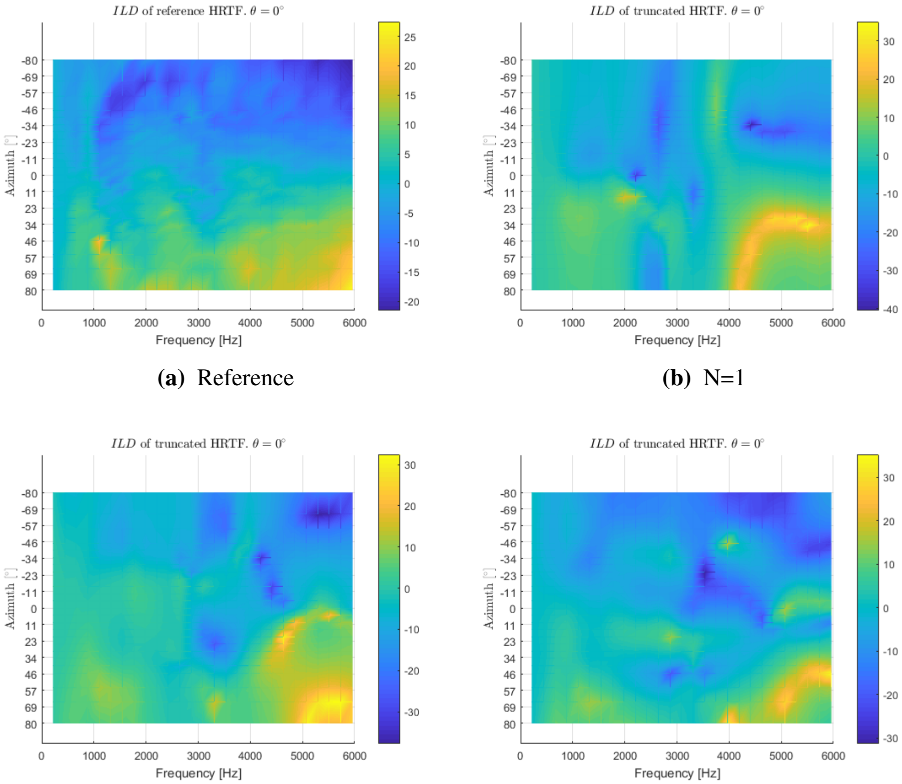


Figure 4.11: The ITD$_p$ error showed for increasing N.

4.2.1.2 Interaural level difference

The reference ILD is shown in figure 4.12a, the reconstructions in figures 4.12b-4.12d and the numerical error shown in figure 4.13. As expected for this binaural cue, a low order reconstruction with N=1 fails to reconstruct this cues, as well as providing a spike around $f = 2$ kHz. Interestingly to observe, is that the response that occurs in the reference to the left ($\phi = 80°$) of the subject at 6 kHz, is instead framed by a similar response. For $f < 2$ kHz, the reconstruction appears even more blurry than the reference, which already has weak responses at these frequencies. For N=2, the response to the left of the subject at 6 kHz is more correctly reconstructed, and the general ILD for $f > 4$ kHz appears good, which is supported by the error actually being the lowest in this area, even though it is a low order reconstruction. For N=3, the aforementioned strong response is again framed instead of reconstructed. The error also seems to have peaks at 4 kHz and 6 kHz . All of the $ILDs$ have weak responses to the right of the subject for $f > 4$kHz, which is to be expected.



(a) Reference

(b) N=1

(c) N=2

(d) N=3

Figure 4.12: Interaural level difference shown for the reference CIPIC HRTF in a, then for N=1-3 in b-d, respectively.

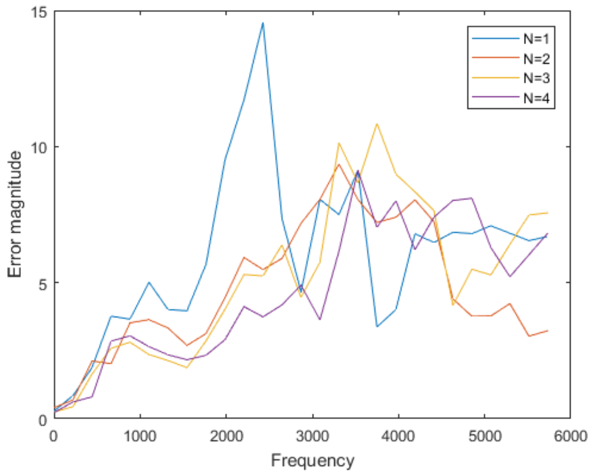
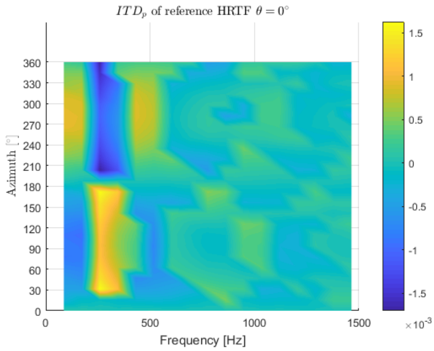


Figure 4.13: The ILD error showed for increasing N for.

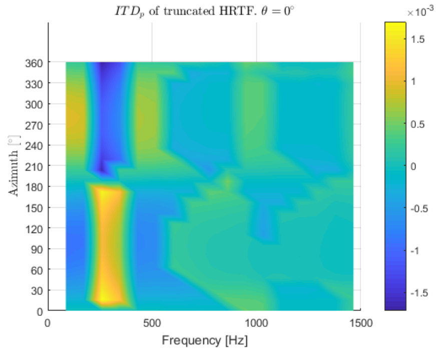
4.2.2 LISTEN

4.2.2.1 Interaural phase delay difference

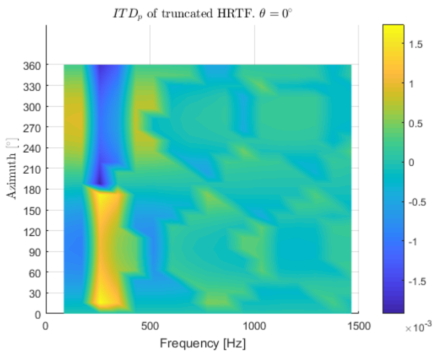
As in the previous section, the LISTEN-dataset seems to perform vastly better at the lowest orders due to its low resolution. Here, the reference is shown in figure 4.14a and the reconstructions in figure 4.14b-4.14d. Below 700 Hz, where the ITD_p is assumed to be of most significance, the reconstruction for N=1 is almost identical to the reference, indicating a solid reconstruction of this localization cue. For N=2 and N=3, the patterns resembles the reference more above 700 Hz, and one can see from figure 4.15 that the error is even lower, but because of earlier mentioned ambiguity in this area, this is of less importance.



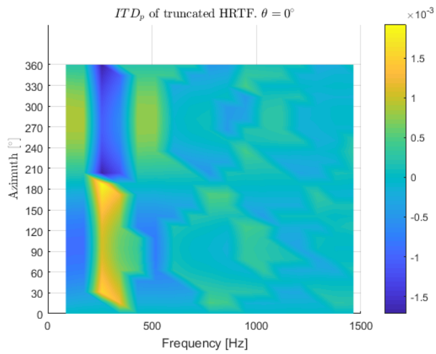
(a) Reference



(b) N=1



(c) N=2



(d) N=3

Figure 4.14: Interaural phase delay difference shown for the reference LISTEN HRTF in a, then for N=1-3 in b-d, respectively.

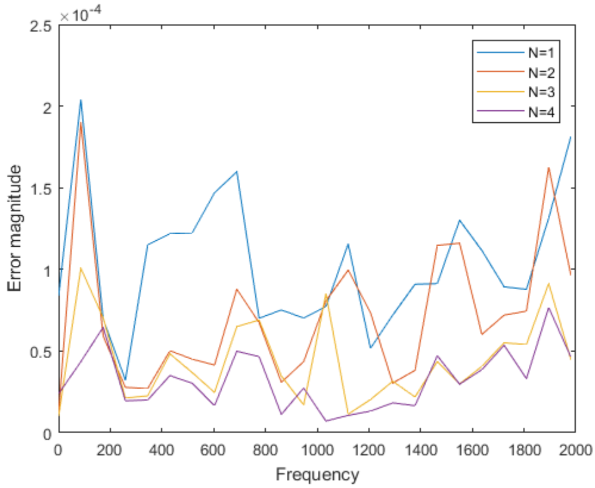
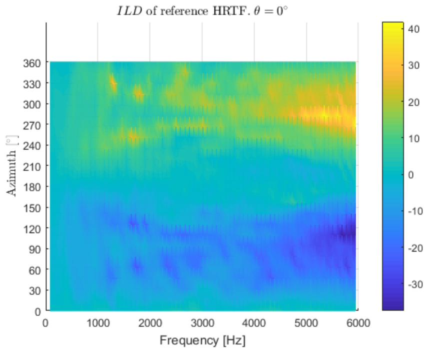


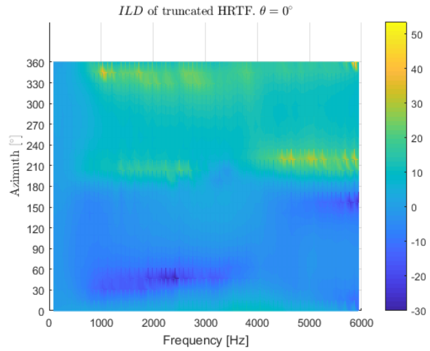
Figure 4.15: The ITD$_p$ error showed for increasing N.

4.2.2.2 Interaural level differences

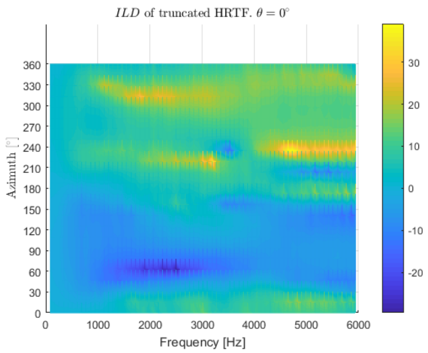
The reference ILD is shown in figure 4.16a, the reconstructions in figures 4.16b-4.16c and the error is figure 4.17. Most notable for these figures is that even though the responses shown for the left ear, the responses appear strongest for the loudspeaker positions at $270° < \phi < 330°$, which is to the right of the subject. Furthermore, the error is twice as high compared to the CIPIC-datasets. Even though some improvement are shown going to N=2 and N=3, the ILDs does not seem to be reconstructed in a good manner for the LISTEN-dataset.



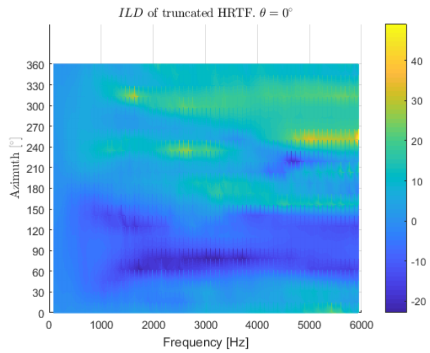
(a) Reference



(b) N=1



(c) N=2



(d) N=3

Figure 4.16: Interaural level difference shown for the reference LISTEN HRTF in a, then for N=1-3 in b-d, respectively.

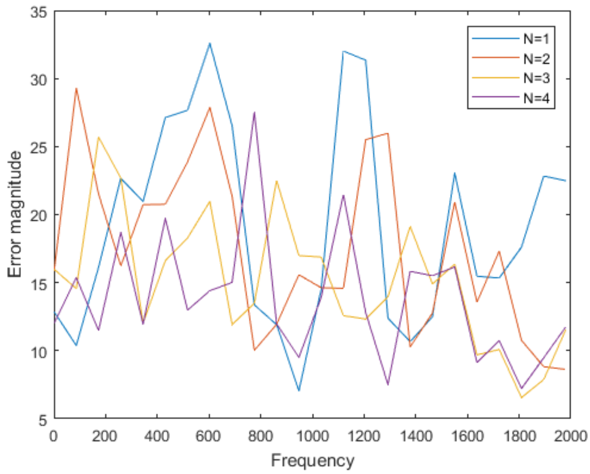
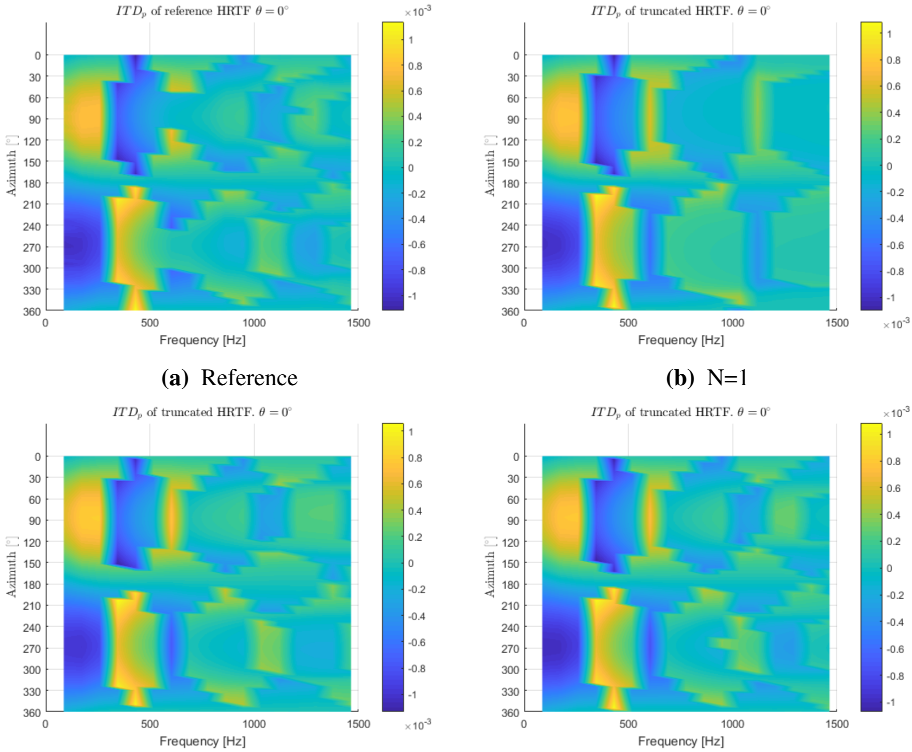


Figure 4.17: The ILD error showed for increasing N.

4.2.3 MIT Media Lab

4.2.3.1 Interaural phase delay difference

The refence ITD_p is shown in figure 4.18a, the reconstructions in figures 4.18b-4.18d and the error in figure 4.19. In agreement with previous datasets, this dataset also provides a very accurate representation for N=1 below 700 Hz, with the exception of the peak in the error plot at around 600 Hz. The improvements are rather small going to N=2 and N=3. There are some parts of the reference pattern that improve at around for f = 1 kHz for N=2-3, but it is fair to assume that these are more and more ambiguous as the frequency reaches 1.5 kHz. The only part that does not seem to reconstruct properly, is the gap at $f \approx 600Hz$ for $60° < \phi < 90°$. This is however exactly at the spot known as the cone of confusion, thus very likely to not have any significant impact on the localization.



(a) Reference **(b)** N=1

(c) N=2 **(d)** N=3

Figure 4.18: Interaural phase delay difference for the reference MIT Media Lab dataset and the reconstructed HRTFs for N=1-3.

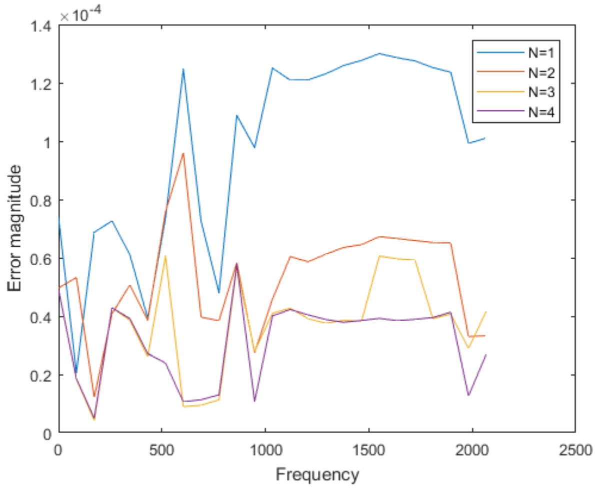
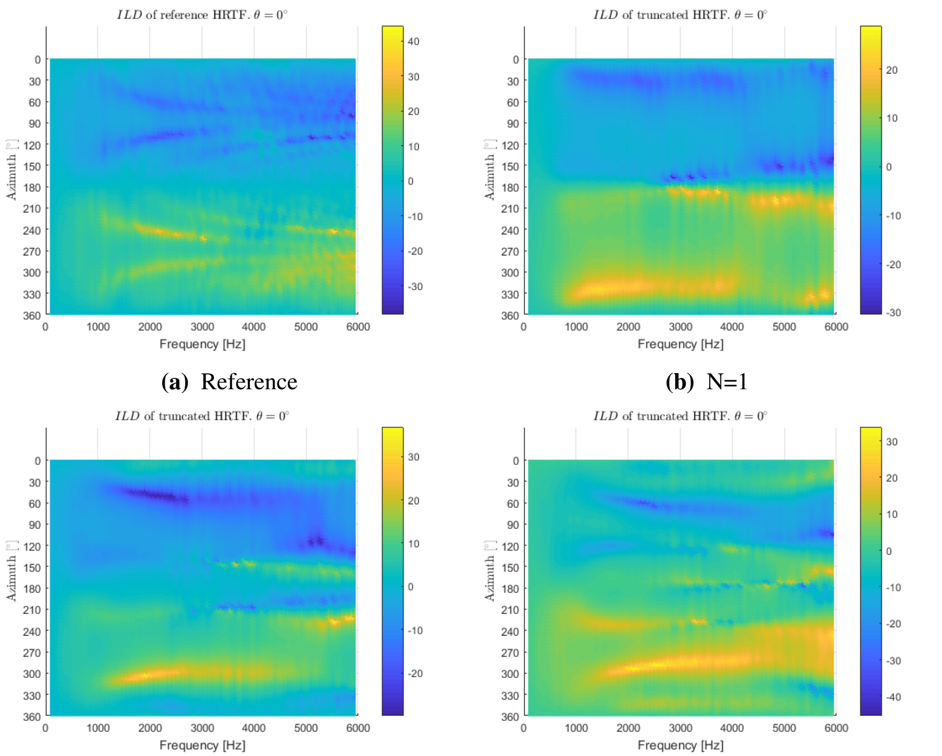


Figure 4.19: The ITD$_p$ error showed for increasing N.

4.2.3.2 Interaural level difference

The reference is shown in figure 4.20a, the reconstructions are shown in figures 4.20b-4.20d and the error plot in figure 4.21. For this dataset, every reconstruction seems to contain the strongest responses at the correct places. The pattern in the reference seems to be following some curved lines, both for the weak and strong responses, and these seems to be reconstructed partially correct at N=2 and N=3. The error is however not as small as the CIPIC-dataset, and the curves seems to follow each other quite closely, indiciating that there is not substantial improvement with increased order.



(a) Reference **(b)** N=1

(c) N=2 **(d)** N=3

Figure 4.20: Interaural level difference shown for the reference MIT Media Lab HRTF in a, then for N=1-3 in b-d, respectively.

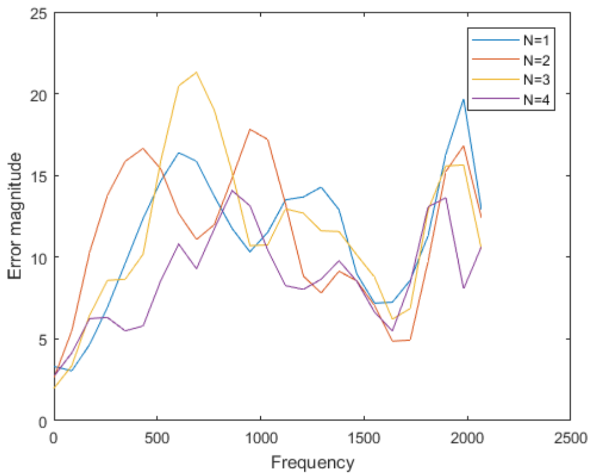


Figure 4.21: The ILD error showed for increasing N.

Chapter 5

Discussion

This chapter will summarize the results and provide a unifying discussion of the work presented in this thesis. First, it will briefly summarize the main findings of the previous chapter. Then a short discussion will follow about possible improvements that could improve the quality of the results with means already available in the framework. Possible sources of error is then provided before finally, topics for related future work are suggested.

5.1 Summarizing the presented results

Going into this work, some general expectations were made by the author:

1. As both suggestions for an f_{max} for accurate reconstruction are linear, a corresponding linear increase in quality was expected for increasing order, meaning that a higher N should equal higher reconstruction quality, until the f_{max} reaches the dynamic range of the HRTF.

2. Though it may not seem like it from the reference HRTFs, a lower amount of measured loudspeaker positions do necessarily mean a lower spatial resolution. As a consequence of this, it was expected that the lower resolution HRTFs should perform better at low order reconstructions.

3. As the most important range for ITD_p lies in $f < 700$ Hz, these localization cues were expected to be constructed fairly well even at N=1. Similarly, the ILD-cues were *not* expected to do so for lower orders, since the ILD is most dominant in the frequency above the one given by f_{max}.

Starting from the top, the results did not completely concur with the first expectation. Though the reconstructions do resemble the reference more with increased order, the biggest leaps were done from N=1 to N=2. For the last two datasets, surprisingly little improvement took place from N=3 to N=4. That said, it should be noted that the fields get more detailed, in the sense that a bigger dynamic range is used to give a more angular dependent response. Ideally, it should be backed up by a corresponding listening test, to get a measure one how these improvements are experienced.

The second expectation can more or less be answered simply by looking at the reconstruction for N=1. Though none of the first order reconstructions are *very* similar to the reference, it is easy to see that the *most* similar is the LISTEN-dataset, which is the one with lowest resolution. However, as the order increases, the apparent quality of reconstruction improves more for the higher resolution datasets than LISTEN, suggesting that HRTFs of resolution similar to that of LISTEN is not a viable option for the future, as it is expected that today‘s workstations will soon be capable of handling more than first order Ambisonics.

Finally, the results from the localization cues did partially concur with the third expectation. Except for the CIPIC-dataset, which had the highest resolution, the ITD_p seems to be reconstructed for N=1 in such a way that the cues are kept intact, and that localization at lower frequencies will remain sufficiently good. For the CIPIC, similar results occured at N=2. Regarding the ILD, the CIPIC had the lowest average error, indicating that a high spatial sampling resolution might help in reconstructing high frequency content, even at lower Ambisonics order.

5.2 Possible improvements with current data

The work in this thesis has revolved around simple visual comparison between a selected HRTF from three different dataset instances, as inspired by Epain et al. [21]. An obvious extension would be to include several HRTFs from the different datasets, in order to determine whether or not individual results are part of a trend in that database, or if they do all deviate somewhat from each other. Because of the low sensitivity for changes caused by the different anatomic structures that makes up individual HRTFs, it would be interesting to see how several subjects from one database performed, especially at the mid-frequencies. Evaluations for different elevations should also be performed.

Another useful aspect in such experiments which might also be a required step to make proper conclusions, is a sort of A/B-testing with listening tests. Even though some of the plot marks out strong differences in some directions, these might be directions with lower localization sensitivity. Thus it would be interesting to see how big differences a subject can detect in different directions. Ideally this test should also involve several subjects, such that one could establish an average measure of how small changes the human hearing can perceive, and at which frequencies they are most significant.

An effort should also be put in the development of more unambiguous evaluation criterion, which might make the comparison between the different datasets easier.

5.3 Sources of error

As the final implementation in this thesis ended up being rather simple, only following the existing theory, the programming itself can be considered to an unlikely source of error. If that were not the case, this would have been obvious from the different plots generated throughout the work. However, that data which the work relies upon is a far more vulnerable source of error. It will always be a change of errors occurring when using public available databases. As a specific example, the HRTF datasets available from Florida Institute University [41] was supposed to be included as the lowest resolution database to have an extreme in both ends. Initial investigations did however show that this database contained a lot of minor flaws, such as having the same values for $\pm 150°$, having a rather mixed order of measurements and so on. Contact with the responsible professor for the measurements revealed that they were performed by students, which in itself is a source of error with a high likelihood. Even though most of the other databases are made up by professors and post docs according to the documentation, human error will always be a chance.

In addition that the database itself might have numerical errors, HRTF measurements are by nature prone to error. An obvious source pf error is the occurrence of head movement when measuring the different HRTFs. This is a very difficult source of error to eliminate, since any mechanism added to prevent it would interfere with the sound waves and distorting the HRTF. Furthermore, the different microphones used in either the subject or mannequin may also alter the results. The different instances have documented the use of microphones with a frequency response as flat as possible, but some deviations do take place above $f > 1$ kHz according to the available documentation. The same source of error applies for the different speakers used in the measurements, where the frequency response can vary from 2-7 dB as in the LISTEN-dataset [40], as well as having different directivity gain.

5.4 Future work

While the suggestions in 5.2 are based on existing techniques and technology, the suggestions for future work are ideas that as far as the author know, has received little or no attention.

- Individualization of HRTF of different resolutions. It is agreed upon in the literature that optimized results will not occur unless an invididualized HRTF is used. Therefore, it would be interesting to see which angular resolution is required of an individualized HRTF to perform ”good enough”.

- While this thesis evaluations are mainly done by visual comparison, one could also further process the signals apart from just a bare Ambisonics encoding and decoding. An example could be the implementation of what has become traditional techniques such as the $max\ r_e$ and *in-phase*-techniques, as introduced by Daniel in [3]. These could improve the binaural signals before comparing them to the reference. As mentioned for the expected quality increase with increased order, it would be interesting to see how much the quality would improve for the different database resolutions.

- Similar comparison evaluated for different input signals (i.e. different source signals that are encoded to Ambisonics). It would especially be interesting to see the performance ratio between actual recordings and phantom sources.

- Interpolation has not been discussed in this work, as all of the encoded angles have corresponded to the angles of the individual HRTFs. It would be interesting to see how the low resolution datasets would perform at other encoded angles, since one could intuitively expect a dataset with higher spatial resolution to have smaller interpolation error. There are available articles on this topics, among others by Freeland et al. [42] and Duraiswami et al. [43].

Chapter 6

Conclusion

Higer Order Ambisonics has throughout this thesis been explained, both in its theoretical foundation and origin, and how one hope to use it for the future. The main problem that was sought out to answer was ”How can the current state of Ambisonics be utilized in today‘s application, and how do different resolutions of HRTFs respond to this”. The first part of this is rather simple to answer, since the workstations built for today‘s social media, such as Facebook and Youtube, only accepts first order Ambisonics. That effectively excludes detailed high frequency content for simple mode matching decoding, and requires other decoding methods such $max\ r_e$ and $in - phase$. There is also the suggestion of using Mixed Order Ambisonics in the more sensitive directions, but this is not feasible until the workstation supports at least 9 transmit channels for full 3D sound.

Considering how the different HRTFs performed for different reconstructions order, it was no surprise that the one with lowest resolution had the most resemblance to its reference. This suggests that one option for more realistic 3D sound rendering, is to focus on measuring smarter HRTFs, in order to get as much information from as low resolution as possible. However, one can conclude with Gumerov and Duraiswami‘s suggestion for the sweet spot radius, $r = \frac{2N-1}{e \cdot k}$, being more accurate than the former rule of thumb that $r = N/k$, especially for the lowest orders.

Most of the ITD_p information was constructed fairly well, while the ILD not quite so. Because of rather large jump in resolutions between the different datasets used, and the results not being completely unambiguous in determining the most important quality aspects due to too few evaluation aspects, a final recommendation for HRTF resolution cannot be made on basis of this thesis. More of the traditional methods should be implemented to make the final signal as good as possible, and it should be subjectively investigated by a thorough listening test, before a proper recommendation can be made.

Bibliography

[1] V. Pulkki, “Virtual sound source postioning using vector base amplitude panning,” *Journal of Audio Engineering Society (JAES)*, vol. 45, no. 6, pp. 456–466, 1997.

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

[3] J. Daniel, J. B. Rault, and J. D. Polack, “Ambisonics encoding of other audio formats for multiple listening conditions,” in *105th Convention (1998 September 26-29 San Francisco, California)*, (California, U.S.A.), 1998.

[4] “FAQ-site explaining the allowed audio formats for VR- and 360-videos on Youtube.” `https://support.google.com/youtube/answer/6395969?co=GENIE.Platform%3DDesktop&hl=en`. Accessed: 19-06-2017.

[5] “Facebook 360 Spatial Workstation.” `https://facebook360.fb.com/spatial-workstation/`. Accessed: 19-06-2017.

[6] M. Kleiner, B.-I. Dalenbäck, and P. Svensson, “Auralization-an overview,” *J. Audio Eng. Soc*, vol. 41, no. 11, pp. 861–875, 1993.

[7] M. Gerzon, “Periphony: With-height sound reproduction,” *Journal of Audio Engineering Society (JAES)*, vol. 21, no. 1, pp. 2–10, 1973.

[8] M. Gerzon, “What‘s wrong with quadraphonics,” 1974. , in Studio Sound (later edited and republished by The Gerzon Archives).

[9] M. Gerzon, “Ambisonics in multichannel broadcasting and video,” *JAES*, vol. 33, no. 11, pp. 859–871, 1985.

[10] J. Daniel, R. Nicol, and S. Moreau, “Further investigations of high order ambisonics and wavefield synthesis for holophonic sound imaging,” 2003.

[11] H. Wallach, “On sound localization,” *Journal of Audio Engineering Society (JAES)*, vol. 10, no. 4, pp. 237–240, 1939.

[12] A. W. Mills, “On the minimum audible angle,” *Journal of Audio Engineering Society (JAES)*, vol. 30, no. 4, pp. 237–246, 1958.

[13] J. Blauert, “Sound localization in the median plane,” *Acta Acustica united with Acustica*, vol. 22, no. 4, pp. 205–213, 1969.

[14] M. B. Gardner and R. S. Gardner, “Problem of localization in the median plane: effect of pinnae cavity occlusion,” *J. Audio Eng. Soc*, vol. 53, no. 2, pp. 400–408, 1973.

[15] J. Hebrank and D. Wright, “Spectral cues used in the localization of sound sources on the median plane,” *J. Audio Eng. Soc*, vol. 56, no. 6, pp. 1829–1834, 1974.

[16] F. L. Wightman and D. J. Kistler, “Headphone simulation of free-field listening. 1: Stimulus synthesis,” *J. Audio Eng. Soc*, vol. 85, no. 2, pp. 858–867, 1989.

[17] F. L. Wightman and D. J. Kistler, “The dominant role of low-frequency interaural time differences in sound localization,” *J. Audio Eng. Soc*, vol. 91, no. 3, pp. 1648–1661, 1992.

[18] F. L. Wightman and D. J. Kistler, “Localization using nonindividualized head-related transfer functions,” *J. Audio Eng. Soc*, vol. 94, no. 1, pp. 111–123, 1993.

[19] M. Noisternig, A. Sontacchi, T. Musil, and R. Holdrich, “A 3d ambisonic based binaural sound reproduction system,” in *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*, Jun 2003.

[20] P. Minnaar, J. Plogsties, and F. Christensen, “Directional resolution of head-related transfer functions required in binaural synthesis,” *J. Audio Eng. Soc*, vol. 53, no. 10, pp. 919–929, 2005.

[21] N. Epain, P. Guillon, A. Kan, R. Kosobrodov, D. Sun, C. Jin, and A. van Schaik, eds., *Objective evaluation of a three-dimensional sound field reproduction system*, The University of Sydney, International Congress on Acoustics, 8 2010.

[22] B. Xie, H. Mai, and X. Zhong, “The median-plane summing localization in ambisonics reproduction,” in *Audio Engineering Society Convention 142*, May 2017.

[23] “Wikipedia: Spherical polar coordinates.” `https://commons.wikimedia.org/wiki/File:BodyPlanes.jpg`. Accessed: 20-06-2017. Figure is reprinted and adapted with permission.

[24] “Wikipedia: Body planes.” `https://commons.wikimedia.org/wiki/File:BodyPlanes.jpg`. Accessed: 06-06-2017. Figure is reprinted and adapted with permission.

[25] V. Algazi, R. O. Duda, D. Thompson, and C. Avendano, “The cipic hrtf database,” in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics 2001*, 2001.

[26] G. Arfken, *Mathematical Methods for Physicists*. Orlando, FL: Academic Press, 3 ed., 1985. Chapter 12.6 and 12.9; ”Spherical Harmonics” and ”Integrals of the Products of three spherical harmonics.

[27] E. Williams, *Fourier Acoustics - Sound Radiation and Nearfield Acoustical Holography*. Academic Press, 1999.

[28] “Wikipedia: Spherical harmonics.” `https://commons.wikimedia.org/wiki/File:Spherical_Harmonics_deg5.png`. Accessed: 06-06-2017. Figure is reprinted with permission.

[29] J. Blauert, *Spatial Hearing - The Psychophysics Of Human Sound Localization*. The MIT Press, 1996.

[30] B. Xie, *Head-Related Transfer Function and Virtual Auditory Display*. J. Ross Publishing, 2013.

[31] “Wikipedia: Head-related transfer functions.” `https://upload.wikimedia.org/wikipedia/commons/5/53/Hrtf_diagram.png`. Accessed: 20-06-2017. Figure is reprinted with permission.

[32] “Sennheser 1st order Ambisonics microphone.” `http://no-no.sennheiser.com/microphone-3d-audio-ambeo-vr-mic`. Accessed: 20-06-2017.

[33] M. A. Poletti, “Three-dimensional surround sound systems based on spherical harmonics,” *J. Audio Eng. Soc*, vol. 53, no. 11, pp. 1004–1025, 2005.

[34] M. A. Poletti, “A unified theory of horizontal holographic sound systems,” *J. Audio Eng. Soc*, vol. 48, no. 12, pp. 1155–1182, 2000.

[35] R. Nicol and M. Emerit, “3d-sound reproduction over an extensive listening area: A hybrid method derived from holophony and ambisonic,” in *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*, Mar 1999.

[36] S. Bertet, J. Daniel, and S. Moreau, “3d sound field recording with higher order ambisonics - objective measurements and validation of spherical microphone,” in *Audio Engineering Society Convention 120*, May 2006.

[37] N. A. Gumerov and R. Duraiswami, *Fast multipole methods for the Helmholtz equation in three dimensions*. Elsevier, 2004.

[38] “Eigenmike EM32 4th order Ambisonics microphone.” `https://mhacoustics.com/products#eigenmike1`. Accessed: 20-06-2017.

[39] W. G. Gardner and K. D. Martin, “HRTF measurements of a KEMAR,” *The Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3907–3908, 1995.

[40] "LISTEN HRTF Database." http://recherche.ircam.fr/equipes/salles/listen/. Accessed February 2017.

[41] "Florida Institute University - HRTF Database." http://dsp.eng.fiu.edu/HRTFDB/main.htm. Accessed February 2017.

[42] F. P. Freeland, L. W. P. Biscainho, and P. S. R. Diniz, "Efficient hrtf interpolation in 3d moving sound," in *Audio Engineering Society Conference: 22nd International Conference: Virtual, Synthetic, and Entertainment Audio*, Jun 2002.

[43] R. Duraiswaini, D. N. Zotkin, and N. A. Gumerov, "Interpolation and range extrapolation of hrtfs [head related transfer functions]," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. iv–45–iv–48 vol.4, May 2004.

Appendix

Generating spherical harmonic

```
function Y=gen_sph_harm_mat(N,theta,phi,varargin)
% Y=gen_sph_harm_mat(N,theta,phi,imag)
% Generates orthonormal spherical harmonics for at azimuth 0<=phi<=2pi and
% elevation 0<=theta<=pi (measured from the polar z axis).
% Y=gen_sph_ha rm_mat(N,theta,phi) and Y=gen_sph_harm_mat(N,theta,phi,0)
% returns real valued spherical harmonics (HOA formulation) while
% Y=gen_sph_harm_mat(N,theta,phi,1) returns complex valued spherical
% harmonics.
%   See also LEGENDRE
% Audun Solvang SINTEF 2014


if (nargin==3)
    Y=real_harm_mat(N,theta,phi);
else
    if varargin{1}==1
        Y=imag_harm_mat(N,theta,phi);
    else
        Y=real_harm_mat(N,theta,phi);
    end
end

function Y=imag_harm_mat(N,theta,phi)
n=0;
m=(-n:n);
P = legendre(n,cos(theta))';
P=[P(:,end:-1:2),P];
w=(-1).^(m).*sqrt((2*n+1)/(4*pi)*factorial(n-abs(m))./factorial(n+abs(m)));
phi_comp=exp(1i*phi*m);
```

```
Y=w(ones(length(theta),1),:).*P.*phi_comp(:,ones(2*n+1,1));
for n=1:N
    m=(-n:n);
    P = legendre(n,cos(theta))';
    P=[P(:,end:-1:2),P];
    w=(-1).^(m).*sqrt((2*n+1)/(4*pi)*factorial(n-abs(m))./
        factorial(n+abs(m)));
    phi_comp=exp(1i*phi*m);
    Y=[Y,w(ones(length(theta),1),:).*P.*phi_comp];
end

function Y=real_harm_mat(N,theta,phi)
n=0;
m=(-n:n);
delta=0*m;
delta(n+1)=1;
P = legendre(n,cos(theta))';
P=[P(:,end:-1:2),P];
w=sqrt((2-delta).*(2*n+1)*factorial(n-abs(m))./factorial(n+
    abs(m)));
phi_comp=1;
Y=w(ones(length(theta),1),:).*P.*phi_comp(:,ones(2*n+1,1));
for n=1:N
    clear phi_comp
    m=(-n:n);
    delta=0*m;
    delta(n+1)=1;
    P = legendre(n,cos(theta))';
    P=[P(:,end:-1:2),P];
    w=sqrt((2*n+1)*(2-delta).*factorial(n-abs(m))./
        factorial(n+abs(m)));
    phi_comp(:,(m<0))=sin(phi*m(m<0));
    phi_comp(:,(m>=0))=cos(phi*m(m>=0));
    Y=[Y,w(ones(length(theta),1),:).*P.*phi_comp];
end
```