Anthropometric Personalisation of Head-Related Impulse Responses

- An application to the Scattering Delay Network and Higher Order Ambisonics -

Master's Thesis

Jason-Yves Tissières

Aalborg University Copenhagen
Sound and Music Computing

Here you can write something about which tools and software you have used for typesetting the document, running simulations and creating figures. If you do not know what to write, either leave this page blank or have a look at the colophon in some of your books.

Sound and Music Computing
Aalborg University Copenhagen
http://www.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
Anthropometric Personalisation of Head-Related Impulse Responses: An application to the Scattering Delay Network and Higher Order Ambisonics

Theme:
Spatial Audio

Project Period:
Spring Semester 2018

Participant(s):
Jason-Yves Tissières

Supervisor(s):
Stefania Serafin
Michele Geronazzo
Wookeun Song

Page Numbers: 79

Date of Completion:
November 9, 2018

Abstract:

This thesis suggests a small set of essential parameters for a personalised and effective dynamic binaural synthesis with headphones. An image-guided procedure with two 2D images of the head guides the personalisation of head-related transfer function (HRTF), combining a spherical head model with ear displacement with the HRTF magnitude selected from a database according to anthropometry. Room acoustics phenomena are simplified following the scattering delay network (SDN) approach which allows an accurate spatialisation of first order reflections. The proposed model is compared to the common higher-order ambisonics (HOA) rendering. Improvements in localisation and externalisation performances within a virtual reality listening experiment attest the benefits of HRTF personalisation compared to the use of generic HRTFs, and indicates the success of the proposed customised SDN model once compared to HOA.

Contents

1 Introduction **1**
- 1.1 Binaural Audio . . . 1
- 1.2 The Question of Personalisation . . . 2
- 1.3 Aim of the Thesis . . . 3

2 Background **5**
- 2.1 Human Sound Localisation . . . 5
- 2.2 Anthropometric Personalisation of HRTFs . . . 6
  - 2.2.1 Computational Model Approach . . . 7
  - 2.2.2 Spherical Head Model . . . 8
  - 2.2.3 Ear Offsets and Optimal Head Radius . . . 10
  - 2.2.4 Image-Guided HRTF selection . . . 11
  - 2.2.5 HRTF Manipulation . . . 14
- 2.3 Binaural Synthesis Support . . . 15
  - 2.3.1 Headphones equalisation . . . 15
  - 2.3.2 HRTF interpolation . . . 15
- 2.4 Scattering Delay Network . . . 16
  - 2.4.1 Overview . . . 16
  - 2.4.2 Network Structure . . . 17
  - 2.4.3 Scattering and Permutation Matrix . . . 19
- 2.5 Ambisonics . . . 19
  - 2.5.1 Spherical Harmonics . . . 20
  - 2.5.2 Encoding a Virtual Source . . . 20
  - 2.5.3 Encoding a Soundfield Recording . . . 20
  - 2.5.4 Sound Field Rotation . . . 22
  - 2.5.5 Binaural Decoding . . . 23
  - 2.5.6 Spatial Room Impulse Response Rendering . . . 24

3 Room Impulse Response Measurements **25**
- 3.1 Hardware . . . 25
  - 3.1.1 The Eigenmike . . . 26

3.2 Procedure 26

4 Implementation **29**
4.1 HRTF Personalisation 29
4.1.1 Design 29
4.1.2 HRTF Database 29
4.1.3 HRTF Selection Software 30
4.1.4 ITD and head-shadow personalisation 31
4.1.5 Anthropometric Data Acquisition 32
4.2 The Binaural SDN 32
4.2.1 ScatAR 33
4.2.2 Binaural Rendering of ScatAR 34
4.2.3 Material Absorption Filter 37
4.2.4 Room Parameterisation 37
4.3 Ambisonics Processing 39
4.3.1 SPARTA plugins 39
4.3.2 Ambisonics System Design 40

5 Evaluation **42**
5.1 Objective Analysis 42
5.1.1 HRTF personalisation: a case study on the KEMAR 42
5.1.2 Measured vs Modelled BRIR 43
5.2 Localisation Test 55
5.2.1 Experimental Design 55
5.2.2 Subjects 55
5.2.3 Apparatus 56
5.2.4 Stimuli 57
5.2.5 Experimental Procedure 58
5.2.6 Results 59

6 Discussion **65**
6.1 Personalisation Wins 65
6.2 Minimal, but External 66
6.3 Future Works 66

7 Conclusion **68**

Bibliography **69**

A Bilinear Transform of the Head Shadow Effect Filter **79**

Chapter 1

Introduction

1.1 Binaural Audio

Spatial Audio and Artificial Reverberation are two topics that have known important progress during the past 50 years. In several important areas such as human/computer interfaces, computer gaming, hearing aids for visually-impaired, virtual and augmented reality systems, pilots and air-traffic controllers and teleconferencing, the synthesis of spatial sound is of great value [1]. In particular, the rise of virtual and augmented reality (AR/VR) systems invokes the incorporation of high quality 3-D audio technologies to provide to the users an immersive experience [2]. Creating an immersive spatial acoustic virtual reality requires the understanding of the mechanisms underlying the spatial auditory imagery of a listener in a space and involves knowledge in, inter alia, acoustics, psychoacoustics, spatial hearing and digital signal processing.

Spatial sound is often required to be delivered via headphones, referred to as binaural synthesis, or Auralization. Auralization is defined by Kleiner et al. as "*the process of rendering audible, by physical or mathematical modeling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modeled space.*" [3]. One simple approach consists in a convolution of a dry signal to a measured Binaural Room Impulse Response (BRIR) which is an impulse response recorded with microphones placed in-ear of a human or a "dummy head" and therefore contains the information of sound propagation in a space from the point source to the listener's ear canal entrance. The BRIR depends on the room and the position and orientation of the source and listener. Although a high quality binaural recording rarely fails to surprise, to the point where the person takes out the headphones to confirm that the sound was really coming from them, convolution based "reverb" has its drawbacks, notably in terms of resources such as computational power, memory and recording equipment. Moreover, an

acoustic scene is far more complex than a simple static source and listener. In complex cases with moving source or listener, a high quantity of BRIRs is required to fulfil the properties of a natural sound environment. Therefore, the use of BRIR in an application come quickly to its limitations.

Alternatively, the computer simulation and synthesis of the environment, the source and the listener are created separately and fused on later stages, typically before the sound is output. This type of audio renderers are called "binaural reverberators". The most common way is to simulate a room with an artificial reverberation algorithm, either one from the "delay network" family, in which the input signal is delayed, filtered and fed back into the system according to reverberation characteristics, or one from the "computational acoustic" category, based on room geometry and acoustic energy [4]. In both approaches, attention is put on precisely modelling the direct path and reflections and applying appropriate manipulations according to spatial aspects of auditory perception. Binaural rendering is usually added at the end of the processing chain by the process of Head-Related Impulse Response (HRIR) filtering. A HRIR is the equivalent of a BRIR without the room information. It is the impulse response describing the filtering effect of the torso, head and pinna when an emitting point source in a specific position in space is received at the ear canal under free-field conditions [5]. Its frequency domain representation is the Head-Related Transfer Function (HRTF).

1.2 The Question of Personalisation

The most common approach in a binaural application is the use of a generic set of HRTFs measured with a "dummy head" which represent an average of a wide human population. However, the use of such datasets, and in general the use of non-individual HRTFs, introduces an increase in errors such as localisation errors, front-back confusion and a lack of externalisation [6], occurring with a high variability between individuals.

A question that arises when considering headphone based rendering systems is: do we need individualisation? In fact, every human being hears sound differently. Our auditory system and its processing in the brain evolved since our young age according to our morphology. Shape of the head, ears, torso and shoulders are well know elements that alters the sound propagation before entering the ear canal and finally hitting the ear drum. Therefore, if the creation of an immersive spatial acoustic virtual reality is aimed, the individualisation of HRTFs plays an important role.

The term "personalisation" came from the fact that acquiring a set of HRTFs

for a specific individual requires special equipment and conditions [7], thus individualisation is impractical for most applications. Personalisation is the process of modeling, manipulating, or selecting the best HRTF filters for an individual, based on his/her anthropometric dimensions. Although the localisation accuracy with personalised HRTFs is slightly worse than with individual HRTFs, it is nevertheless still better than with generic HRTFs, considering the laborious process of measurements. Over the years, several databases of human HRTFs have been publicly accessible for commercial, artistic and research purposes. Examples are the ones of CIPIC [8], LISTEN [9] and ARI [10]. These databases contain HRTF sets of individuals with a methodical and precise spatial sampling with measurement conditions varying from a database to another [11, p. 28]. Some contain photographs and anthropometric dimensions of the subjects for a deeper investigation on the influence that head and body parts have on HRTFs. Numerous studies have been made to understand, model and personalise HRTFs according to a particular individual.

1.3 Aim of the Thesis

This thesis concerns the design, implementation and evaluation of an anthropometry-based HRTF personalisation resulting from two 2-D images of the head. The aim of the thesis is to objectively and subjectively investigate the performance of this personalisation with respect to a generic set of HRTFs. The personalisation process will be implemented in two spatial Audio dynamic rendering systems, namely:

1. Scattering Delay Network (SDN) reverberation algorithm: considering a state-of-the-art artificial reverberator, and based on simplified physical modeling of room geometry and sound propagation.

2. Ambisonics approach: a multichannel recording- and reproduction-based system that mathematically encodes the sound field, a technique widely used in virtual reality systems.

BRIRs and multichannel room impulse responses were measured in two rooms from our University and will serve as case studies and acoustic ground truth for this research.

The thesis is organised as follow. **Chapter 2** presents the background of this research. The chapter is divided into three main parts. First, recent techniques of HRTF personalisation are described, followed by a presentation of the SDN and finally, a description of the fundamentals of the Ambisonics approach. **Chapter 3** presents the impulse response measurements of the two classrooms with a B&K dummy head and an Eigenmike spherical array microphone. **Chapter 4** describes

the implementation of the HRTF personalisation process and its incorporation in the SDN and the Ambisonics systems, to form two dynamic auralization systems. The evaluation is presented in **Chapter 5**. First an objective analysis is performed on the proposed HRTF personalisation, followed by the parameterisation of the SDN according the measured BRIRs of our two classrooms. An objective comparison is performed between measured and modeled BRIRs. The HRTF personalisation in combination with the rendering systems are then perceptually evaluated in terms of localisation performances and externalisation judgements. **Chapter 6** discusses these results and proposes future developments. Finally, **Chapter 7** summarises the work and results of this research and conclude the thesis.

Chapter 2

Background

2.1 Human Sound Localisation

Assuming an emitting omnidirectional point source in the far-field, its sound wave will propagate towards the head and its ears. In addition to the direct sound path, the wave is also refracted and diffracted by the body parts, mainly the head, torso and pinnae, before entering the ear canals. These phenomenon, which are also frequency dependent, cause signal delays and attenuations that are used by the auditory system and its processing in the brain to locate the position of the source.

The determination of the sound source location is firstly a binaural process, where the auditory system takes into account the differences in sound level and time, namely the interaural level- and time-difference (ILD and ITD). The differences are mainly dependent on the size and shape of the head, the position of the ears and the frequency, and reaches its maximum when the sound source is positioned around 90 degrees to the right or left of the listener. The Duplex theory, proposed by Lord Rayleigh in 1907, describes the dominance of the ITD cue for frequencies below 1.5 kHz and the ILD cue is used for greater frequencies [12]. Also, it is known that ITD is frequency-independent for frequencies below 500 Hz and above 3 kHz, with an additional 50% low-frequency delay compared to high frequencies, and being variable in the mid-range [13]. On the other hand, ILD is highly frequency-dependent due to the diffraction and shadowing effect introduced by the head, with a loss in high frequencies when the source is on the far side of the head [14].

Furthermore, since there are locations where the ITD and ILD are the same, due to the symmetry of the ears relative to the median plane, it is almost impossible by the simple means of these two cues to locate the sound source. These points are located on the so called Cones of Confusion, centred at the interaural axis [15]. Therefore, sound localisation for sources on the Cones of Confusion is determined by monaural cues which are produced by the fine structure of the pinna at high fre-

quencies. In fact, sound waves of a wavelength comparable to the size of our ears can be reflected against the morphological structure of the pinna, thus spectrally and temporally modifying the sound before entering the ear canal. It is also known that the determination of elevation in the median plane is essentially a monaural process [16]. Additionally, natural small head movements help solve these confusions by slightly changing interaural differences. These are called dynamic cues.

The effect of reverberation on localisation performance is not fully understood. Literature on the precedence effect [17] and the selection cue model [18] states that room reflections have little to no influence on sound source localisation. Instead, reverberation can provide information for sound source distance [19] and is known to strongly influence the impressions of source width and envelopment [20]. However, other studies show that reverberation affects the performance in localisation, compared to an anechoic signals [21], depending on room acoustics and nature of the sound source [22] and the position and orientation of the source and listener is the room [23].

Perceptual Misjudgements in Binaural Synthesis

In binaural synthesis, the use of non-individual HRTFs typically introduces a degradation in localisation accuracy. Figure 2.1 illustrates errors that are characteristically analysed in the literature. A *localisation error* is a misjudgement of the direction of a sound stimulus relative to a target response, typically reported as azimuth and elevation errors (number 1 in the figure). A *reversal error*, also known as front-back confusion, is referred to a confusion in direction relative to the interaural axis (number 2 in the figure). These errors occur due sound stimuli on the cones of confusion. Finally, when a sound stimulus is perceived as coming from inside the head, it is referred to an *externalisation error* (number 3 in the figure).

It has been shown that reversal and externalisation errors are significantly reduced by head-tracking and reverberation, respectively [21].

2.2 Anthropometric Personalisation of HRTFs

HRTF personalisation has been increasingly investigated in recent literature. A focus has been put on matching anthropometric features to temporal and spectral characteristics found in HRTFs. The goal of HRTF personalisation is to get as close as possible to the perceptual attributes of an individual without going through the laborious exercise of HRTF measurements. In the literature, there appears to be two main trends regarding HRTF personalisation. The first is the generation of synthetic responses derived from computational models. The second, and more recent, is the selection of a best set of HRTF from an existing database. The following

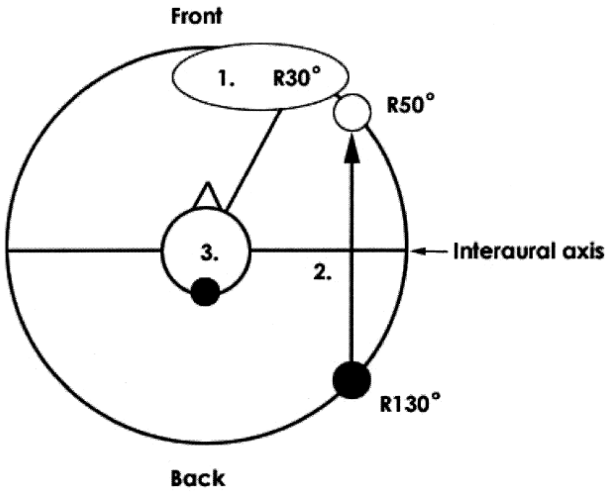

Figure 2.1: Top view illustrations of typical perceptual misjudgements in binaural synthesis. 1. Localisation error: an emitting sound source from direction 30° can be perceived with a deviation from the target. 2. Reversal error: a source source at 50° is confused with a source symmetrically opposed relative to the interaural-axis (130°). 3. "In-head" localisation: a misjudged distance of the sound source, perceived as coming from inside the head. Image from [21].

describes relevant methods.

2.2.1 Computational Model Approach

Computational modelling techniques vary from simple amplitude change and signal delay producing interaural differences, to complete mathematical models. The most common approach consists in exploiting the linear property of HRTFs and to separate the synthesis of the most influential elements forming a HRTF, i.e. torso, head and pinna, into isolated sub-components. This approach is referred to as "structural modelling" [14]. The sub-components are studied separately and modelled in the form of low-order Digital Signal Processing (DSP) filters which parameters are fit to certain anthropometric dimensions of the listener, e.g. head radius, pinna shape. The advantage of this approach is that it is well suited to real-time applications and is flexible due to the ability to improve or add components to the system. The different blocks forming the system are derived from numerical solutions of wave propagation on simple geometric shapes such as sphere or ellipse, from ray tracing computations, or from in-depth analysis of measured HRTFs using advanced signal processing algorithms.

The spherical head model is the most recurrent model in the literature [24, 5, 13, 14, 25, 26]. It is an excellent first approximation of the human head and facilitates the design of signal processing algorithms. It is in general used to model or predict interaural differences and is subject to numerous improvements such as

ear displacements [27], optimised ITD [28] and optimised head radius [25]. The ellipsoidal head model was also proposed as an upgrade to the sphere regarding ITD prediction [29]. However, no HRTF was derived from the ellipsoidal model due to its complex analytical solution.

The torso and the shoulders have a relatively weak effect on HRTFs compared to the influence of the head and the pinnae. [14]. The torso introduces a shadowing effect for sound sources (or reflections) coming from below the listener. In addition, if the sound source is emitting from above, its sound waves will reflect on shoulders and potentially support elevation perception resolve localisation on the cones of confusion [30]. Algazi et al. modelled a spherical head in combination with a spherical torso in the form of a structural model derived notably from ray tracing analysis [31]. The system is composed of two sub-systems that are switching according to weather the sound source is above (shoulder reflections) or below (torso shadow) a certain threshold called the shadow limit. The model was extended to an ellipsoidal torso [32].

The pinna has a major role in elevation localisation and is very delicate to model. It introduces peaks and notches in the high end of the spectrum (> 3 kHz) that are intricately related to the shape and size of the ears. Geronazzo et al. identified two resonances and three reflections that were strictly pinna-related [33, 34]. Their analysis is based on the extraction of the "pinna-related transfer function" (PRTF) from measured HRTFs and mapping PRTFs' notches and peaks to anatomical parts of the ear (helix, concha, etc.). Interestingly, their work resulted in parameterising the filters of the model by automatic feature extraction from a picture of the pinna [35] or acoustic selfies [36] .

More complex physical models of HRTFs exist, for example a boundary element method approach [37], or from a 3D scan of the head [38, 39].

2.2.2 Spherical Head Model

Based on Brown and Duda's structural model [14] and Algazi's adaptation [31], the head model is defined as a shadowing filter H in cascade with a time delay ΔT, as illustrated in Figure 2.3. Both filters are dependent on the radius a of the sphere and the observation angle γ formed with an incident sound wave.

Defining γ as the dot product between the source vector $\vec{S_v}$ and the ear vector $\vec{e_v}$ with origin at the centre of the sphere [27]

$$\gamma = cos^{-1}\left[\frac{\vec{S_v}\cdot\vec{e_v}}{\|\vec{S_v}\|\|\vec{e_v}\|}\right], \tag{2.1}$$

the head-shadow filter is defined as [31]:

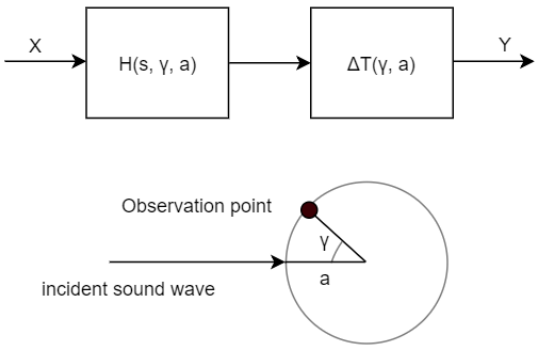

Figure 2.2: Illustration of the spherical head structural model from Brown. **Top**: The angle γ formed by an incident sound wave and an observation point on a sphere of radius a. **Bottom**: The shadow filter in cascade with a time delay, both depending on sphere radius and observation angle.

$$H(s,\gamma,a) = \frac{\alpha(\gamma)s+\beta}{s+\beta}, \tag{2.2}$$

where

$$\beta = \frac{2c}{a} \tag{2.3}$$

where c is the speed of sound and $\alpha(\gamma)$ is an asymptotic high-frequency gain

$$\alpha\left(\gamma\right) = \left(1+\frac{0.1}{2}\right) + \left(1-\frac{0.1}{2}\right) cos\left(\frac{\gamma}{5\pi/6}\pi\right). \tag{2.4}$$

The delay block ΔT is the time difference between the time the sound wave arrives at the centre of the sphere (if the sphere was absent) and the time it hits the observation point. Given the head radius a and the speed of sound c, it is approximated by Woodworth and Schlosberg's frequency-independent formula [24]:

$$\Delta T(\gamma,a) = \begin{cases} -\frac{a}{c}cos\gamma & \text{if } 0 \leq \mid \gamma \mid < \frac{\pi}{2} \\ \frac{a}{c}\left[\mid \gamma \mid -\frac{\pi}{2}\right] & \text{if } \frac{\pi}{2} \leq \mid \gamma \mid < \pi \end{cases} \tag{2.5}$$

This formula is valid for frequencies above 600 Hz [11]. From Equation 2.5 and for ears that are diametrically opposed, the ITD is given by [27]:

$$ITD(\gamma,a) = \begin{cases} \frac{a}{c}\left(sin\gamma+\gamma\right) & \text{if } 0 \leq \mid \gamma \mid < \frac{\pi}{2} \\ \frac{a}{c}\left(\pi-\gamma+sin\gamma\right) & \text{if } \frac{\pi}{2} \leq \mid \gamma \mid < \pi \end{cases} \tag{2.6}$$

Note that the head-shadow filter introduces a group delay in the low end of the spectrum, accounting for the 50% additional low-frequency delay found by Kuhn [14]. This head model is particularly attractive for its simplicity and its ability to estimate ITD for any given head radius and ear positions.

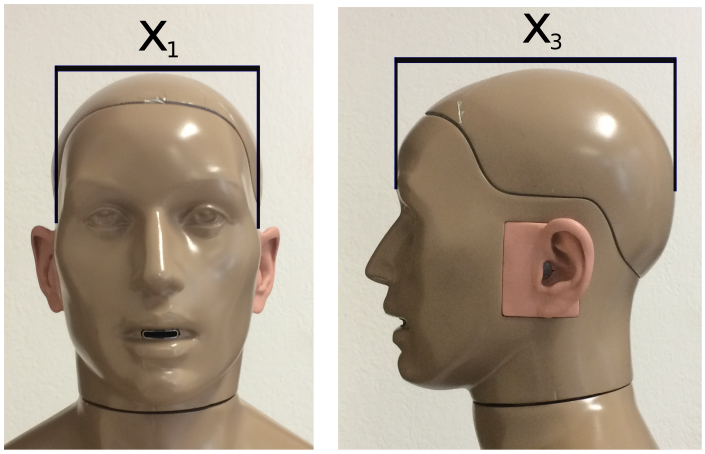

Figure 2.3: Frontal and lateral picture of KEMAR with anthropometric dimensions (a) head width (X_1), (b) head depth (X_3).

2.2.3 Ear Offsets and Optimal Head Radius

Diametrically opposed ears on the spherical head model gives a rough estimation of the ITD. However, this modelling doesn't allow patterns appearing with elevation-dependent features of measured HRTFs and in particular on the cones of confusion [27]. It is known that the sensitivity to ITD in the frontal sector is of tens of microseconds [5]. Therefore, an improved ITD model is necessary.

Recently, Bahu et al. investigated the effect of ear offsets on ITD accuracy with the spherical head as well as on the ellipsoidal head model [27]. In their work, they compared different ear offset scenarios with the CIPIC database and furthermore calculated optimised head radius and ear offsets by least-mean square and linear regression in regards to the anthropometric data from the database. For the spherical head, results of frontward (e_b cm) and downward (e_d cm) ear shifts showed significant improvement over the antipodal model. Differences between ellipsoidal and spherical surprisingly appeared as minim. The regression formula for head radius is [27]:

$$a_{opt} = 0.41\frac{X_1}{2} + 0.22\frac{X_3}{2} + 3.7 \tag{2.7}$$

where X_1 and X_3 are the head width and depth, respectively (see Figure 2.3). Note that the head height (X_2) was found to be non-significant in the determination of the optimal head radius.

The optimal ear offsets has been found to be optimal with a frontward shift(for 36 of 37 CIPIC subjects) by $e_b = -0.94$ cm and downward (for 100% of the CIPIC subjects) by $e_d = 2.10$ cm.

Introducing an ear displacement relative to the +/- 90 degrees position is equivalent to translating the position of the sphere relative to the interaural axis (see Figure 2.4). Therefore, the centre of the new sphere is translated to $x_0 = [0, e_b, e_d]$ and γ of Equation 2.1 becomes [27]:

$$\gamma = cos^{-1}\left[\frac{(\vec{S_v} - \vec{x_0}) \cdot (\vec{e_v} - \vec{x_0})}{\|\vec{S_v} - \vec{x_0}\|\|\vec{e_v} - \vec{x_0}\|}\right] \tag{2.8}$$

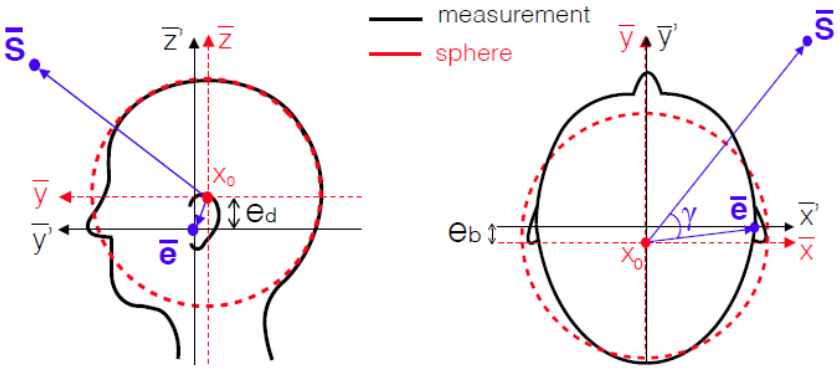

Figure 2.4: Illustration from [27] showing an upward and backward shift of the sphere of the original coordinate system, the equivalent of shifting the ears downwards by e_d and frontwards by e_b.

2.2.4 Image-Guided HRTF selection

Due to the increasing amount of HRTF databases available, HRTF selection has been increasingly investigated. The concept relies in selecting the best dataset for a given individual. In the literature, most methods have been tested with the CIPIC database, released by Algazi et al. in 2001 [8]. This database contains HRTFs of 45 subjects measured for both ears from 1250 positions. The database also includes pictures and and complete anthropometric data for 35 subjects, and partially for the remaining subjects.

Zotkin et al.'s hypothesis was that if two persons had similar anthropometric features, they would have compatible HRTFs [40]. They proposed a 2-D image based system which selects the best match using seven dimensions from the pinna. Results from a localisation test with head-tracker showed a general decrease in localisation error, but not observable on all participants.

In the same lines, Geronazzo et al. proposed an image-guided selection tool based on the relation between ear contours and frequency notches appearing in HRTFs [41]. Their work was the continuation of their proposed structural pinna model composed of two resonances and three reflections, briefly described in Section 2.2.1.

The HRTF selection problem is based on anthropometric features of the pinna, mapped into the HRTF domain by a ray tracing method. In fact, incoming sound waves of sufficiently high frequency are small enough to reflect on pinna contours, before entering the ear canal. Therefore, because of the high variability in pinna shapes between individuals, important differences between HTRF sets occur, thus modifying our perceived sound location, in particular influencing the vertical localisation. Based on the distance $d_i(\phi)$ between the entrance of the ear canal, a pinna contour C and the speed of sound c, the time delay $t_d(\phi)$ between the arrival of the direct sound (sound that enters the ear canal directly) and the reflected wave, depending on elevation ϕ is defined as [41]:

$$t_d(\phi) = \frac{2d_i(\phi)}{c}, i = 1, 2, 3 \tag{2.9}$$

Figure 2.5 shows a picture of an ear with traced pinna contours and intersection points relative to a specific elevation angle and ear canal entrance.

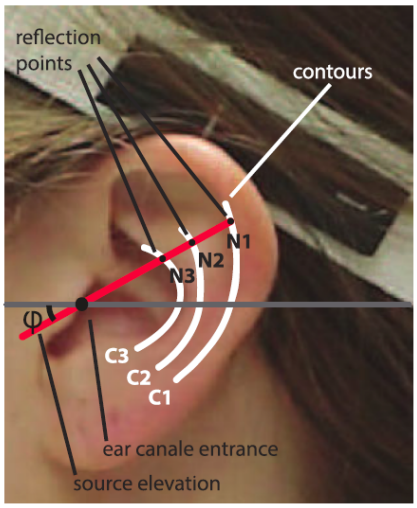

Figure 2.5: Picture of an ear, with pinna traced pinna contours C1, C2 and C3 and the intersection points relative to elevation angle and ear canal entrance [42].

Due to the reflections, spectral notches appear in the corresponding HRTF. The frequencies at which a notch occur is found with:

$$f_n(\phi) = \frac{n+1}{t_d(\phi)} = \frac{c(n+1)}{2d_i\phi}, \quad n = 0, 1, ..., \quad i = 1, 2, 3 \tag{2.10}$$

assuming the reflection coefficient to be negative (which is the case for 80% of the CIPIC subjects [34]). Therefore, the first notch is found at ($n = 0$):

$$f_0(\phi) = \frac{c}{2d_i(\phi)}, \quad i = 1, 2, 3 \tag{2.11}$$

Now, given N estimate of a contour (let's say contour C1) with K different ear canal entrance points, a combination of all these possibilities with all possible elevation angles ([$-45°$,$45°$]), gives a matrix of notch frequencies. By comparing these notch frequencies to notch frequencies of a measured HRTF, a new metric is introduced, called the "mismatch". The mismatch function is defined as follow and will give a mismatch matrix corresponding to the N and K estimates [41]:

$$m_{(m,k)} = \frac{1}{N_\phi} \sum_{\phi} \frac{|f_0^{(k,n)}(\phi) - F_0(\phi)|}{F_0(\phi)} \tag{2.12}$$

where F_0 corresponds to the notch frequency extracted from the measured HRTF.

Iterating this comparison process with each HRTF set of a database yields ranking lists based on 3 metrics derived from mismatch matrices [41]:

- *Mismatch: each HRTF is assigned a similarity score that corresponds exactly to increasing values of the mismatch function calculated with Eq. 2.12 (for a single (k, n) pair).*
- *Ranked position: each HRTF is assigned a similarity score that is an integer corresponding to its ranked position taken from the previous mismatch values (for a single (k, n) pair).*
- *Top-M appearance: for a given integer M, for each HRTF, a similarity score is assigned according to the number of times (for all the (k, n) pairs) in which that HRTF ranks in the first M positions.*

Mismatch matrices can as well be calculated for the other contours C2 and C3. However, results showed that C1 (the outer contour) is the most significant [43].

Results

This method has been evaluated several times in the literature essentially in terms of localisation performance. In [43], a static localisation test was performed on 8 subjects for 80 points around the listener. Results showed an improvement in vertical localisation of 28%compared to the generic KEMAR dataset. Furthermore, the model has been evaluated on a virtual auditory system and also confirmed by significant improvements [41]. It also has been evaluated dynamically in a VR environments [44, 45].

Note that all these evaluation were performed in anechoic conditions and that in terms of azimuth errors, no improvements is observed.

2.2.5 HRTF Manipulation

HRTF manipulations are for instance useful for ITD modification of an original dataset or for combining synthetic and measured HRTFs.

Overwriting the ITD of an existing HRTF dataset can be achieved in the frequency domain by minimum-phase plus delay reconstruction [46]. The method consists in estimating the delay to be applied, convert this delay to phase and finally add it to the minimum phase of the data. The estimated ITD can be used as a time-constant delay τ_{itd}. However, because the delay of the minimum phase τ_{mp} is not zero, τ_{itd} must be compensated [11]:

$$\tau_{itd} = ITD - \tau_{mp} \tag{2.13}$$

In this way, the reinserted ITD won't be overestimated or underestimated.

Additionally, it is known that measured HRTFs are unreliable at low-frequencies because of the use of relatively small loudspeakers for the measurements. Therefore, as suggested by Zotkin [40] and Algazi [31] compensation can be achieved by combining the frequency magnitude response of the spherical head model (or snowman model) HRTF and the measured HRTF and keeping the phase of the head model. This is expressed in the following way:

Let H_h be the head model HRTF and H_c the corresponding HRTF from the database, therefore the combined HRTF is H_s:

$$A_s(\omega) = \begin{cases} A_h(\omega) & \text{if } \omega < \omega_l \\ A_h(\omega) + \frac{A_c(\omega) - A_h(\omega)}{\omega_h - \omega_l}(\omega - \omega_l) & \text{if } \omega_l < \omega < \omega_h \\ A_c(\omega) & \text{if } \omega > \omega_h \end{cases} \tag{2.14}$$

with

$$A_s(\omega) = \log \mid H_s(\omega) \mid, \quad A_h(\omega) = \log \mid H_h(\omega) \mid, \quad A_c(\omega) = \log \mid H_c(\omega) \mid$$

.

$$\omega_l = 250Hz \qquad \omega_h = 1000Hz$$

.

Thus, below ω_l the magnitude of the head model is used. Above ω_h, is the magnitude of the measured HRTF, and in between, a cross-fading between both measured and modelled. Finally, the ITD is reinserted with the minimum-phase reconstruction explained above.

Zotkin et al. proposed this model in combination with their pinna image-guided HRTF selection process [40]. After the selection of the best match from

the CIPIC database, they combined the selected dataset with the snowman model in order to improve low-frequency localisation cues. Results showed significant improvement in localisation performance relative to the generic HRTF dataset.

This "hybrid" between measured and modelled HRTF can be considered as a mixed structural modelling [47] which concept is to create a HRTF from a combination of "partial" HRTFs that can either be synthetic or measured, thus making a system modular.

2.3 Binaural Synthesis Support

This section describes two issues related to headphone based spatial audio rendering that will be taken into account in the current work. Methods are proposed to remedy to the matters.

2.3.1 Headphones equalisation

Due the spectral sensitivity in spatial hearing, it is important for an accurate binaural synthesis to compensate for the spectral coloration introduced by the headphone transducers and the high-frequency resonances of the listener's pinnae [48, 49]. A simple method to compensate these effects consist in filtering the output with the inverse of the headphone's transfer function (HpTF) measured on a dummy head or human subjects. Ideally, the HpTF should be the one measured on the individual using the system. However, databases containing HpTFs of popular headphones exist [50].

2.3.2 HRTF interpolation

HRTF databases contain measurements for a finite set of points, and most of the time they don't cover a full sphere around the listener. For a real-time dynamic rendering involving head-tracking and moving sources, it is crucial to have a high-quality time-varying interpolation between HRTFs. Otherwise, audible "jumps" will occur during the rendering.

HRTF interpolation commonly consists in finding the weights of the closest N neighbouring points. Gamper's method [51] consists in triangulating the original set of points formed by a pair of azimuth and elevation[1]. As a result, these points become vertices of triangles. Consider a triangle formed by vertices **A**, **B** and **C**, therefore any point **X** inside the triangle can be represented by a linear combination of the vertices [51]:

$$\mathbf{X} = g_1\mathbf{A} + g_2\mathbf{B} + g_3\mathbf{C} \tag{2.15}$$

[1]The method from Gamper originally includes distance as a third variable, therefore forming tetrahedrons. Here we assume the dataset to be measured at the same distance.

where g_i are scalar weights representing the barycentric coordinates of X, with the constraint:

$$\sum_i g_i = 1 \tag{2.16}$$

These weights are used as interpolation weights to compute the estimated HRTF $\mathbf{H_X}$ at point $\mathbf{X}$ from the three neighbouring HRTFs $\mathbf{H_i}$ at points $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$:

$$\mathbf{H_X} = \sum_i g_i \mathbf{H_i} \tag{2.17}$$

The first two weights g_1 and g_2 are found by solving the equation:

$$\begin{bmatrix} g_1 & g_2 \end{bmatrix} = (\mathbf{X} - \mathbf{C})\mathbf{T}^{-1} \tag{2.18}$$

where

$$\mathbf{T} = \begin{bmatrix} \mathbf{A} - \mathbf{C} \\ \mathbf{B} - \mathbf{C} \end{bmatrix} \tag{2.19}$$

And finally, the last weight g_3 is found with Equation (2.16):

$$g_3 = 1 - g_1 - g_2 \tag{2.20}$$

The interpolated HRTF at point X is therefore found by searching for the triangle that encloses X and by calculating the corresponding weights for the three neighbouring HRTFs. The algorithms complexity is O(N) for a brute-force search, with N being the number of HRTF points in the dataset.

2.4 Scattering Delay Network

This section describes the general structure of the Scattering Delay Network (SDN) reverberation algorithm, introduced by Hacıhabiboğlu et al. in 2011 [52, 53].

2.4.1 Overview

The SDN is a physically-based digital reverberator part of the digital waveguides family. It has its roots in the Digital Waveguide Mesh (DWM) [54], the Image-source (IM) [55], and the Feedback-Delay Network (FDN) [56] for the following reasons:

- The DWM, because of its general structure based on room geometry and the properties of sound waves simulated by digital waveguides arriving and scattered at junctions.
- The IM, because of its method for finding first order reflection points.

- The FDN, because of its internal structure of feedback loop (matrix) to reinsert energy into the system.

The SDN system renders accurately in time and space the paths of the direct sound (source-listener) and first-order reflections by creating a fully connected mesh topology connected at scattering junctions via uni- and bi- directional delay lines. The scattering junctions are situated where first order reflection points occur. Higher-order reflections utilise this same network, thus they are rendered with a decreasing accuracy as the order gets higher. Nevertheless, the SDN renders a significant subset of reflections that would occur in an enclosed space while keeping the efficiency of a FDN structure and the flexibility of a physical room model. Figure 2.6 illustrates the conceptual design of the SDN.

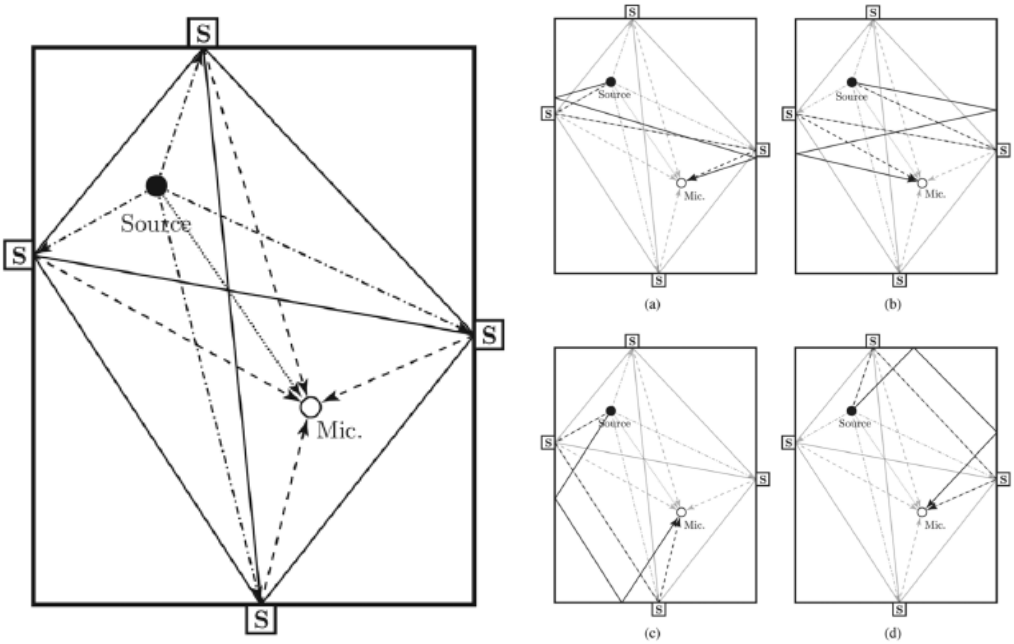

Figure 2.6: [53]. **Left**: Conceptual illustration of the SDN. This figure represent the top view of a rectangular room where a source and microphone are placed. The network is created by delay lines interconnected via scattering junctions **S**. The source transmits its signal to each junction (point-dashed lines) and to the microphone directly (dotted line). The junctions are connected to the the microphone to transmit the reflections (dashed lines). The junctions are also connected to each other to form paths for higher order reflections (solid lines). **Right**: Illustrated comparison of second order paths. Second order paths use the SDN network (dashed lines) instead of their true paths (solid lines).

2.4.2 Network Structure

A block diagram representing the operation of the SDN is depicted in Figure 2.7. Here, only the essential elements are explained. A detailed description of each

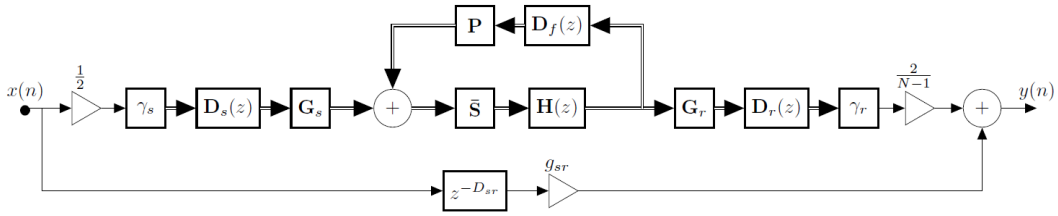

Figure 2.7: Block diagram overview of the Scattering Delay Network. Image from [57].

stage is found in [52, 53].

The input $x(n)$ is separated in two paths: the direct path and the network path. The direct path connects $x(n)$ directly to the output y(n) by applying a delay D_{sr} and a gain g_{sr}. D_{sr} is defined according to the distance d_{sr} between the source and the receiver, the speed of sound c and the sampling rate F_s, so that $D_{sr} = d_{sr} \times F_s/c$. g_{sr} is the attenuation gain according to the $1/r$ law: $g_{sr} = 1/d_{sr}$.

For the network path, $x(n)$ is injected though the source node and and is transmitted to the wall-junctions after the application of delay and attenuation matrix operators, $\mathbf{D}_s(z)$ and $\mathbf{G}_s$, containing the set of delays and attenuation gains for each source-junction connection. The signal is then scattered between wall-junctions connections and the junction-receiver connection with the scattering matrix $\overline{\mathbf{S}}$. This operation is accompanied by frequency-dependent absorption $\mathbf{H}(z)$ filters, determining the frequency response of the corresponding wall. This usually depends on the wall surface material. In addition to the frequency modification introduced by the wall, a certain amount of energy is absorbed when the incident wave hit the wall, called the absorption coefficient $\alpha \in [0;1]$. α is then introduced in the scattering matrix $\overline{\mathbf{S}} = \beta\overline{\mathbf{S}}$, where $\beta = \sqrt{1-\alpha}$. Furthermore, for the higher order reflections, a feedback loop operation is applied. It consists of inter-junction delays $\mathbf{D}_f$ followed by a permutation symmetric matrix $\mathbf{P}$ whose elements depend on the network topology. For a simple rectangular cuboid (shoebox), $\mathbf{P} = \delta_{i,f(j)}$, where $\delta_{i,j}$ is the Kronecker delta, where [53]

$$f(j) = (6j - (j-1)\%N - 1)\%(N(N-1)) + 1 \tag{2.21}$$

Finally, the gateway to $y(n)$ is the application of attenuation gain matrix $\mathbf{G_r}$ and delay $\mathbf{D}_s$, whose elements are associated to each junction-receiver respective connection and sum these signals to the direct path signal.

Note that the input scaling factor of $\frac{1}{2}$ is an adjustment due to waveguide properties to provide the intended pressure at the junctions [53]. This operation is then compensated at the end of the system by $\frac{2}{N-1}$, where N is the total number of junctions (N = 6 for a shoebox model). Note also that γ_s and γ_r are the source and receiver directivity patterns which can be modelled as gains depending on the direction or as frequency-dependent filters.

2.4.3 Scattering and Permutation Matrix

Given N wall junctions in the system, each junction has to manage one incoming source unidirectional delay line, one outgoing microphone unidirectional delay line and $(N-1)$ bidirectional inter-junction delay lines. The scattering main matrix $\overline{\mathbf{S}}$, of size $N(N-1) \times N(N-1)$, is composed of smaller identical scattering matrices $\mathbf{S}$ that determine the spread of acoustic energy that runs in the system, in between the bidirectional connections [52]:

$$\mathbf{S} = \frac{2}{N-1}\mathbf{1}_{(N-1)(N-1)} - \mathbf{I} \tag{2.22}$$

where $\mathbf{1}$ is the all-ones matrix and $\mathbf{I}$ is the Identity matrix. S is utilised to calculate the outgoing sound pressure signal p_{ij}^{-} from the incoming sound pressure signals p_{ij}^{+}:

$$p_{ij}^{-} = \mathbf{S}p_{ij}^{+} \tag{2.23}$$

In the recursive loop, because the order of the signal matrix was altered by the scattering matrix, signals must be reorder before being fed back into the system, hence the permutation matrix.

2.5 Ambisonics

Introduced by Gerzon in the 1970s [58, 59], the ambisonic approach describes a multichannel system for encoding and rendering a 3D sound field. It relies on the decomposition of the sound field into the so-called spherical harmonic functions and is represented by its order N. Originally, ambisonics was introduced for processing 1st-order sound scenes recorded with four cardioid microphones in a tetrahedral formation and yielding a 4-channel WXYZ configuration called the B-format. W, the monopole channel, and XYZ, three dipoles channels form the ambisonics channels. More recently, ambisonics have been extended to higher orders, reaching presently practical realisations of between 3rd and 7th-order. These require more microphones and $(N+1)^2$ Higher Order Ambisonics (HOA) signals, with each order containing $2N+1$ signals, which can be represented in the time- or frequency-domain. The general process of ambisonics is divided into 3 blocks:

1. Encoding a signal or recording a sound scene
2. Manipulating of the sound scene in space
3. Decoding and reproduction over a loudspeaker setup or headphones.

The operations within and between these blocks are defined in the spatial transform domain. Complete mathematical details of spherical array, ambisonics and spherical microphone array processing can be found in [60], [61] and [62]. The following describes the fundamentals of the three blocks mentioned above.

2.5.1 Spherical Harmonics

The acoustic pressure in every point of a source-free sphere that is centred at the origin of a chosen referential can be expressed by a Fourier-Bessel decomposition:

$$p(kr,\theta,\phi) = \sum_{m=0}^{\infty} i^m j_m(kr) \sum_{n=0}^{m} \sum_{\sigma=\pm 1} B_{mn}^{\sigma} Y_{mn}^{\sigma}(\theta,\phi) \tag{2.24}$$

where (kr,θ,ϕ) are the spherical coordinates of a measured point, with k representing the wave number, r the radius and, θ and ϕ the azimuth and elevation respectively. The spherical Bessel functions $j_m(kr)$ are dependent on the distance between the origin and the point, and the frequency. The functions $Y_{mn}^{\sigma}(\theta,\phi)$ are the spherical harmonics, which define an orthonormal basis, and, together with B_{mn}^{σ} they represent the projection of the acoustic pressure on this basis [63]. Theoretically, the exact sound field can be represented by an infinite number of harmonics, but in practice, the series are truncated to a finite order N. The sound field is described by a limited number of spherical Fourier coefficients and can only be partially reconstructed.

Spherical harmonics can be defined by their real-valued representation [61]:

$$Y_n^m(\theta,\phi) = \sqrt{(2n+1)(2-\delta_{0m})\frac{(n-m)!}{(n+m)!}} P_n^m(cos\theta) \times \begin{cases} cos(m\phi) & m > 0 \\ 1 & m = 0 \\ sin(m\phi) & m < 0 \end{cases} \tag{2.25}$$

where δ_{mn} is the Kronecker delta function, $P_n^m(x)$ the associated Legendre functions and $0 \leq n \leq N$, $n \leq m \leq n$ are the order and degree, respectively. The directivity patterns of the real spherical harmonics up to order 4 are shown in Figure 2.8. The first two levels (row) of harmonics represent the WXYZ of the B-format mentioned above. The higher the order, the more the spatial information increases.

2.5.2 Encoding a Virtual Source

The ambisonic theory is based on the principle of plane wave (far-field) reproduction of amplitude S (the source signal), coming from a direction (θ_S, ϕ_S). The encoding process consists in decomposing the sound field on the spherical harmonics Y_n^m. Defining $\mathbf{Y}$ as the matrix of spherical harmonics and $\mathbf{B}$ as the vector of ambisonics signals B_n^m the encoding process is expressed by:

$$\mathbf{B} = S\mathbf{Y} \tag{2.26}$$

2.5.3 Encoding a Soundfield Recording

The sound field can also directly be recorded in multichannel and encoded to HOA signals. This is achieved by placing microphones that sample the sound field. Over

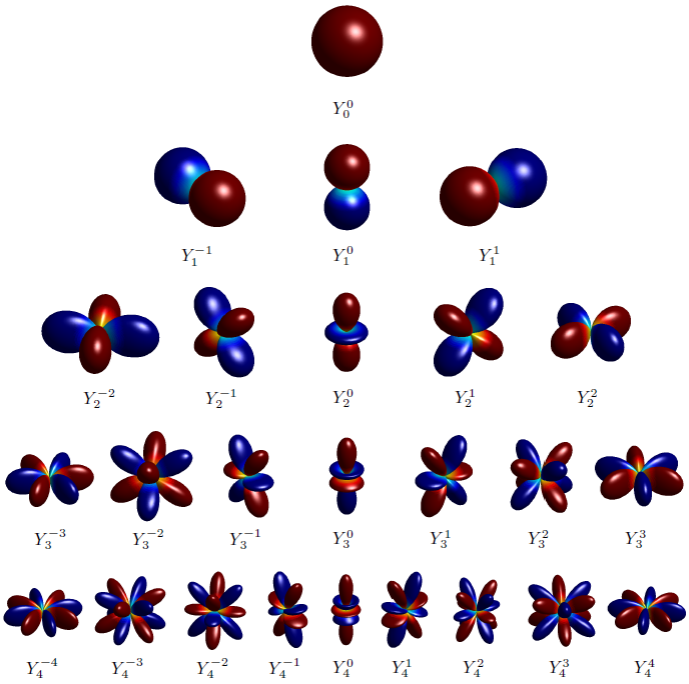

Figure 2.8: Directivity patterns of real spherical harmonics up to 4rd-order [61].

the past decade, spherical microphone arrays (microphones mounted on a hard baffle) have gained in popularity [64, 65], due to the convenient configuration and known sound pressure characteristics on a rigid sphere. Assuming a spherical array of radius R with $Q > (N+1)^2$ sensors, it is possible to obtain an estimation of ambisonics signals up to order N. A common method is the least square method. The sound field is sampled at the surface of the sphere and the pressure on a sensor q is described in a matrix form as [61]:

$$\boldsymbol{p} = \boldsymbol{Y}\boldsymbol{W}\boldsymbol{B} \tag{2.27}$$

where $\boldsymbol{p}$ is the sensor pressures $(N \times 1)$ vector, $\boldsymbol{Y}$ the spherical harmonics $(N \times N)$ matrix and $\boldsymbol{B}$ the ambisonics signals $(N \times 1)$ matrix. $\boldsymbol{W}$ is a "pseudo diagonal" $(N \times N)$ matrix of radial functions (also called equalisation weights), $\boldsymbol{W} = pdiag\,[W_n(kR)]$. These can be considered as filters applied to the microphone signals; filters that are dependent on the geometry of the array, the ambisonic order, and the frequency.

The estimation of the ambisonics signals $\widetilde{\boldsymbol{B}}$ is determined by the following least square solution [61]:

$$\widetilde{\boldsymbol{B}} = \boldsymbol{W}^{-1}(\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{Y})^{-1}\boldsymbol{Y}^{\boldsymbol{T}}\boldsymbol{p} = pdiag[1/W_n(kR)]\boldsymbol{E}\boldsymbol{p} \tag{2.28}$$

where $\boldsymbol{E} = pinv(\boldsymbol{Y}) = (\boldsymbol{Y}^T\boldsymbol{Y})^{-1}\boldsymbol{Y}^T$ (pseudo-inverse matrix). Therefore, the microphone signals are first encoded by $\boldsymbol{E}$, to then be filtered by the inverse radial functions. Figure 2.9 shows a graphical representation of the encoding process of a microphone array of Q sensors encoded into $(N+1)^2$ HOA channels.

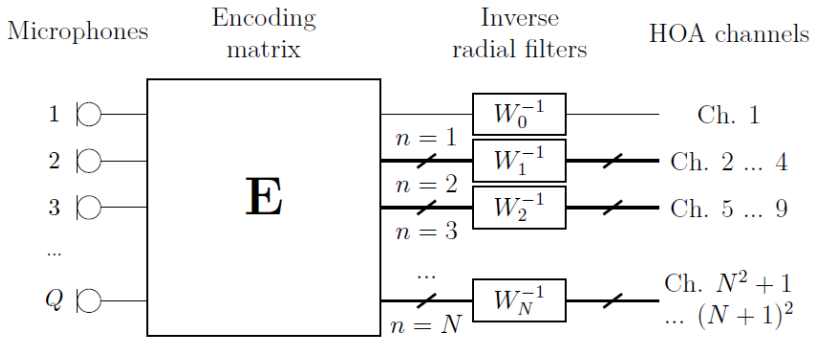

Figure 2.9: HOA encoding process of a microphone array. [61].

In the case of a hard sphere, the equalisation weights filters are given by [61]:

$$W_n(kR) = \frac{i^{n+1}}{(kR)^2 h_n'(kR)} \tag{2.29}$$

where $h'(x)$ is the first derivative of the Hankel function.

A problem occurring when inverting these filters is that internal noise of the sensors are amplified, mainly at low frequencies and higher orders [66]. A popular approach to compensate for this low frequency "boost" is to apply a regularised inversion with Tikhonov filter [65]. These filters would be added after each inverse radial filter.

2.5.4 Sound Field Rotation

Sound field rotation is particularly attractive for binaural reproduction with head-tracking, but also serves for combining multiple HOA sound fields together or repositioning sound sources. The common approach is based on rotating an object that has 3 degrees of freedom. For a HOA sound field, this is achieved by multiplying the ambisonics signals (in matrix form) to rotation matrix which is a block diagonal matrix of $(2n+1) \times (2n+1)$ matrices for each order n = 0,...,N [61]:

$$\boldsymbol{B_{rot}} = \boldsymbol{RB} \tag{2.30}$$

A common rotation matrix convention is the Yaw-Pitch-Roll (YPR) convention. Details of rotation matrix computation is out of the scope of this project, but the reader is referred to [67].

2.5.5 Binaural Decoding

To reproduce the sound field at a listener's position the HOA signal of $(N+1)^2$ channels shall be decoded to an array of $L > (N+1)^2$ loudspeakers, either with a real setup or, virtually with a set of virtual loudspeakers. The loudspeakers must be distributed on a sphere at directions (θ_l, ϕ_l) and radiating towards the listening point, and placed far enough such that the wavefront is sufficiently plane at the origin. A common loudspeaker configuration is the t-design, consisting of a nearly-uniform distribution of points on a sphere [68]. The advantage of a uniform distribution is that it provides an orthogonal sampling of spherical harmonics, thus facilitating the decoding process [69]. The matrix $\mathbf{S}$, composed of the loudspeaker signals $S_1, ..., S_L$, can be found by the equation:

$$\mathbf{S} = \mathbf{DB} \tag{2.31}$$

where $\mathbf{D}$ is the decoding matrix composed of elements $D^m_{n,l}$. Optimal decoder design has been one of the biggest challenge in ambisonics [70]. The most simplest decoder is the Sampling Decoder (SAD) which correspond essentially to a plane-wave decomposition at the direction of each loudspeaker. The method consists in calculating the spherical harmonics of order N at the L loudspeaker directions, then normalising the transpose of this matrix by the mean angular segment of each loudspeaker [70]:

$$\mathbf{D}_{SAD} = \frac{\pi}{L}\mathbf{Y}_L^T \tag{2.32}$$

A binaural reproduction technique consists in decoding the HOA signal to a set of virtual loudspeaker and assign a pair of HRTFs to each loudspeaker. "*The signal at each ear is the sum of L loudspeaker signals $S_l(\omega)$ filtered with the corresponding HRTFs $H_{l,Left}(\omega)$ and $H_{l,Right}(\omega)$*" [61]:

$$S_{ear}(\omega) = \sum_{l=1}^{L} H_l(\omega) S_l(\omega) \tag{2.33}$$

Moreover, since the sound field is only reproduced partially due to the truncation of the Fourier-Bessel series to an order N, sound sources rendered by the HOA decoding system appear with a blur which is proportional to the ambisonic order [63]. An objective rough angular estimate of this blur width is shown in Table 2.1. Also, studies have evaluated the perceptual localisation in HOA rendering systems, either recorded with microphone arrays [63] or synthesised [71], with results showing a better performance with higher order systems. A third order is recommended in [71].

Order	1	2	3	4
Blur	45°	30°	22.5°	18°

Table 2.1: Angular blur as a function of ambisonic order. Values taken from [63]

.

Note that binaural rendering with the ambisonics approach do not require the incorporation of a real-time interpolation between HRTFs. Only the sound field is rotating, with the virtual loudspeaker at fixed positions.

2.5.6 Spatial Room Impulse Response Rendering

A Spatial Room Impulse Response (SRIR) is a multichannel RIR captured with a microphone array. A SRIR contains a much finer acoustic information of the space as it capture the directionality information of reflections. In order to render a sound source from which the SRIR was measured, the SRIR is encoded to ambisonics and a dry signal is convolved to each of the ambisoncs signals [72, 73]. More advanced parametric techniques such as SIRR and DirAC take into account the statistics and dependencies of the signal recorded and render the sound field with psychoacoustic and perceptual motivations [74, 75].

Chapter 3

Room Impulse Response Measurements

All room impulse responses employed in this research were measured in two rectangular cuboid class rooms at our University (A-2.0.028 and D-0.108). The first is an (8.8 x 8.8 x 3.4) m room, considered of medium size, and the second is a (19.3 x 15.2 x 3.2) m large class room. Binaural and spatial room impulse responses were measured in both rooms from 6 different directions around the recording point positioned at the centre of the room in a seated height (1.23 m). The angular directions of the measurements, relative to the listening point, are listed in Table 3.1 and illustrated in Figure 3.1. The directions were the same for both rooms.

Table 3.1: Room impulse response measurement angles.

ID	Azimuth	Elevation
P1	0°	0.9°
P2	30°	17.9°
P3	120°	-36.8°
P4	180°	43.7°
P5	-150°	-27.8°
P6	-60°	28.9°

3.1 Hardware

The following hardware was used for the RIR recordings:

1. A dummy head B&K 4100 Head-and-Torso simulator (HATS)
2. An Eigenmike em32 spherical array microphone

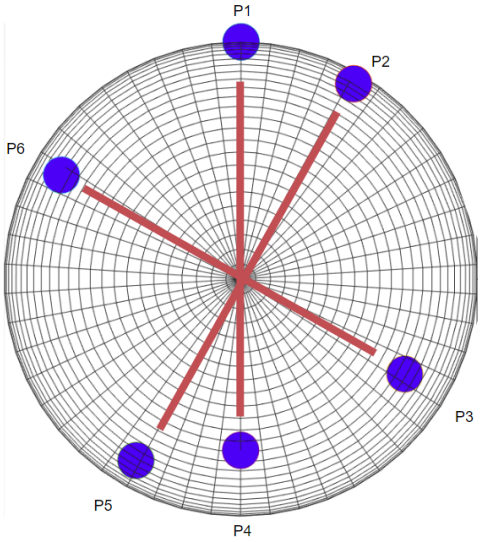

Figure 3.1: Top view of room impulse response measurement positions. The recording point is the centre of the sphere.

3. A Fireface 800 audio interface
4. Six Dynaudio MK-II loudspeakers

3.1.1 The Eigenmike

The em32 Eigenmike from mh Acoustics[1] is a rigid sphere of radius 4.2 cm. It is composed of 32 half-inch microphones mounted on the surface of the sphere, nearly uniformly. The microphones are directly connected to preamplifiers and AD converters placed inside the sphere, so that only one CAT digital cable connects the Eigenmike to the interface. The array is first connected to its own interface box (EMIB) which converts the proprietary multiplexed microphone signals into a Firewire audio stream to appear as a 32 channel audio driver in the PC. Note that with its 32 microphones, the Eigemike signals can be encoded to ambisonics up to 4th-order.

3.2 Procedure

The rooms were partially emptied from the chairs and tables, the remaining were placed against the walls. Both room had a background noise of approximately 27 dBA. The B&K and the Eigenmike em-32 were placed alternately at the centre of the room at seated height (1.23 m) surrounded by the 6 loudspeakers placed at 1.2 m

[1] `https://mhacoustics.com/products`

distance and tilted towards the recording point. The source centre was considered to be the point on the loudspeaker between the membrane and the tweeter. Figure 3.2 shows the measurement setup. Microphone signals were amplified to obtain a peak signal at -6 dB.

Impulse responses were generated, recorded and processed with the ITA Matlab toolbox [76], using the exponential sine sweep (ESS) method [77], with an FFT length of 18 (= 5.5 seconds), at sampling rate 48 kHz and covering the frequency range 20 - 20000 Hz. IRs were then low-passed to 20 kHz and cropped to a length of 1 s for the medium room and 1.5 s for the large room.

(a) B&K 4100 HATS in the medium-size room.

(b) Eigenmike in the large-size room.

Figure 3.2: Room impulse response measurement setup. P1 is the loudspeaker directly in front of the microphone. The next position are following clockwise.

Chapter 4

Implementation

4.1 HRTF Personalisation

4.1.1 Design

The HRTF personalisation follows the same concept as Zotkin [40], described in Section 2.2.5. It is designed as a combination of a selection of measured HRTFs from a database, with the spherical head model with optimised head radius and ear offsets (see Section 2.2.2 and Section 2.2.3 for theoretical details). The HRTF dataset is selected from the CIPIC database with the image-guided "ear-contours" selection method presented in Section 2.2.4.

Three anthropometric dimensions are needed for the proposed personalisation:

1. The pinna contour (C1), for the HRTF dataset selection.
2. The head width (X_1), for the spherical head model.
3. The head depth (X_3), idem as (2).

Figure 4.1 shows the HRTF personalisation procedure. First, according to the C1 ear contour, the best match from the CIPIC database is selected. Then, the ITD and low-frequency head-shadow is corrected for each point of the dataset according the optimised spherical head model.

4.1.2 HRTF Database

The CIPIC database was used in this project. It contains datasets of 1250 measurement points for 25 different azimuths angles ranging from -80 and 80 degrees (azimuths = [-80, -65, -55, -45:5:45, 55, 65, 80]) and 50 different elevations angles ranging from -45 to 230.625 degrees (in steps of 5.626), defined in the interaural-polar coordinates system. Each impulse response contain 200 samples, sampled at 44.1 kHz.

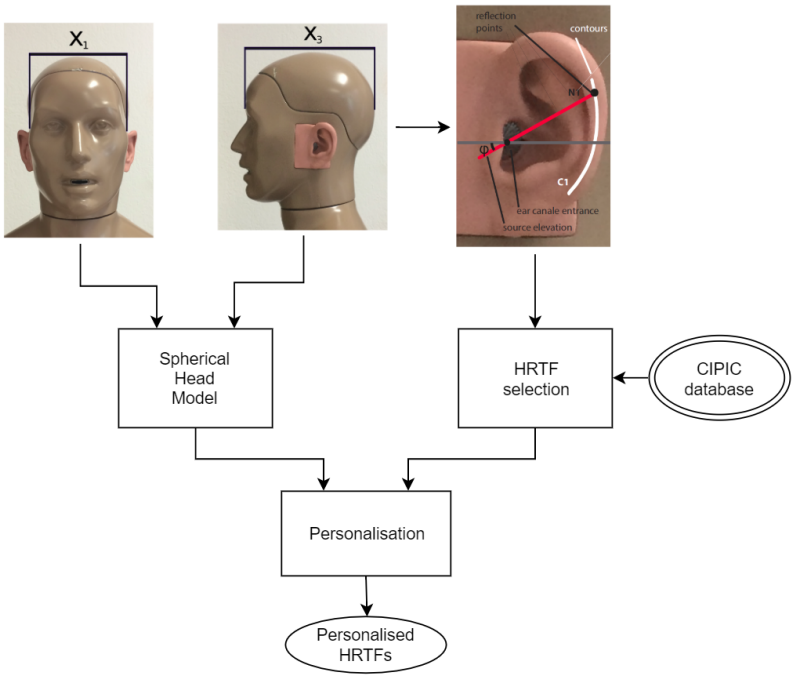

Figure 4.1: HRTF personalisation procedure. The Head width and head depth are parameters for the spherical head model HRTF. The pinna is for the measured HRTF selection process. The synthesised and measured HRTFs are then combined.

Extrapolation

Given that the CIPIC measurement grid is not a full sphere around the listener and that reflections coming from the ground are most likely to have an elevation lower than -45 degrees, HRTF extrapolation was needed. Moreover, it important for the interpolation algorithm to have points that span the entire sphere. The most important points to extrapolate are the points at +/-90 degrees azimuth and the points below the listener, i.e. elevation -90 and 270 degrees. For practical reasons, points at +/-90 azimuth were obtained by simple arithmetic mean of points of azimuth +/-80 degrees azimuth, respectively. On the other hand, points of elevation -90 and 270 degrees were obtained by a weighted mean of points of elevation -45 and 230.625 degrees. Therefore, the dataset was extended to a total of 1404 points.

4.1.3 HRTF Selection Software

The authors of the pinna-based selection method have developed a mouse-driven Matlab software implementing the algorithm [41]. Figure 4.2 shows a screenshot of the main application window. After loading a picture of a pinna and adjusting it (scaling, rotation, and pixel-to-meter ratio adjustments), N pinna contours and K

ear canal entrances are to be estimated by tracing them directly on the image. The software then computes the mismatch (Equation 2.12) between the notch frequencies estimates and the pre-computed notch frequencies from the CIPIC subjects and outputs the subjects scores according to the three metrics described in Section 2.2.4. The selected dataset is the one from the CIPIC subject appearing the most in the "top-3 rank" of the C1 contour with negative reflection coefficients condition [41].

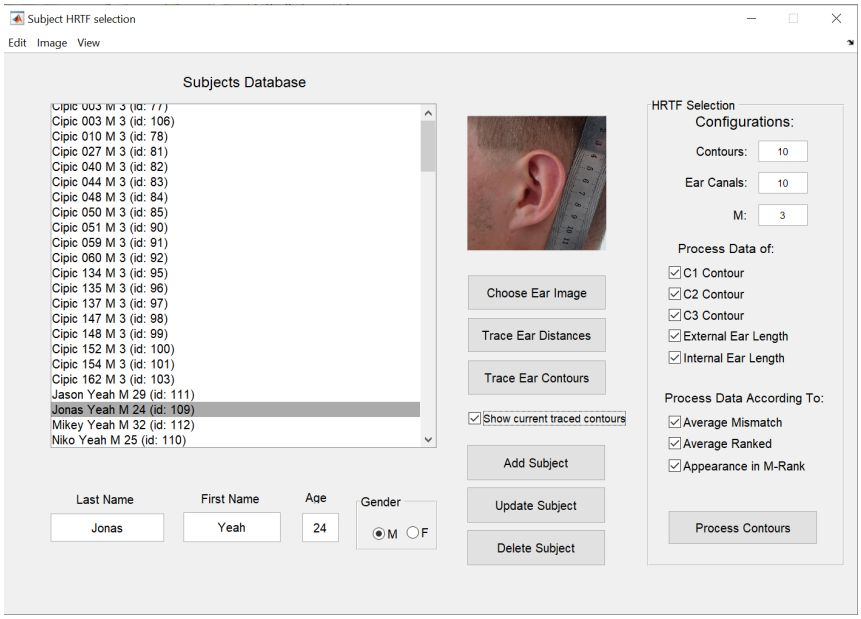

Figure 4.2: Screenshot of the HRTF selection software.

4.1.4 ITD and head-shadow personalisation

ITDs and low frequencies head shadow are modified for each point in the selected dataset with the methods presented in Section 2.2.2, 2.2.3 and 2.2.5, i.e.:

1. ITD estimations of the spherical head model with optimal head radius (Eq. (2.7)) and ear offsets (the ear offsets of $e_b = -0.94$ and $e_d = 2.10$ cm are fixed).

2. Magnitude response combination and ITD reinsertion by minimum-phase plus delay reconstruction method (Eq. (2.14)).

Note that Equation (2.2) was easily discretised by a bilinear transform (see Appendix A).

A Matlab script automating the whole process was created for this purpose.

4.1.5 Anthropometric Data Acquisition

Anthropometric data was acquired from a front and lateral picture of the subject. The subject was asked to seat with his/her back and head straight while the frontal and side pictures were carefully taken from a distance of 70 cm. Also, a ruler was placed in order to have a reference for the digital pixel-to-meter ratio adjustment (see Figure 4.3). Images were then uploaded in the HRTF selection software and post-processed (aligned, cropped, pixel-to-meter ratio adjusted) in order to be ready for the anthropometric data extraction. While the software was the tool for the pinna image HRTF selection, it also allowed for distance measurements and was found as the most practical and precise method for measuring the head width X_1 and head depth X_3 metrics.

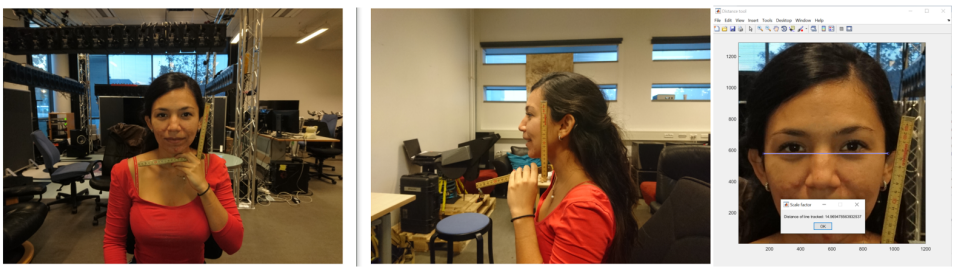

Figure 4.3: Image acquisition for anthropometric data extraction.

4.2 The Binaural SDN

This section describes the implementation of the SDN auralization system which was built upon *ScatAR* [78], an AR application implementing the SDN algorithm (description in the next section). The binaural SDN is developed in Unity[1] C# as a fully customisable system in terms of room geometry, wall properties, number of sources and personalised listener with head tracking (Oculus Rift). Figure 4.4 shows a general overview of the auralization system. A dry signal associated with a source object is rendered by the SDN representing the room which is parameterised based on RIR measurements. The listener object, defined by its position and orientation from the headtracking data, influences both the SDN mechanism and the human hearing cues. The latter are produced by HRTF filtering, personalised according to the user's anthropometric data.

[1]Unity 3D: `https://unity3d.com/`

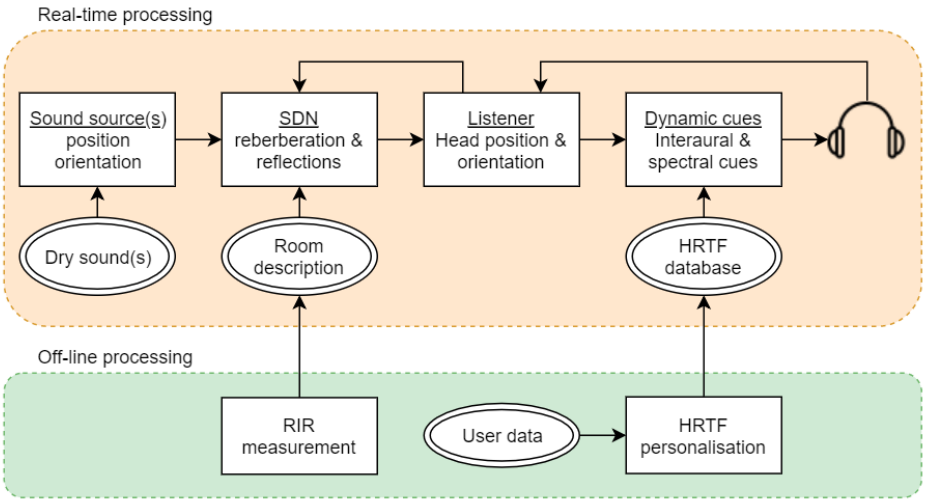

Figure 4.4: Overview of the binaural SDN.

4.2.1 ScatAR

ScatAR, developed by Alex Baldwin, is an open source augmented reality application[2] that has implemented the SDN based on a scan of the room with an AR device [78]. The application was developed for the Lenovo Phab 2 Pro 1, with Google's Tango AR framework[3] in Unity with the C# programming language. The depth camera sensor of the device scans the room, which is then triangulated into a mesh. From this mesh, first order reflection points are found according to the position of the listener (the device itself) and an omnidirectional sound source object placed in the AR scene (see Figure 4.5, left image). *ScatAR* implements the SDN dynamically, so that the source and the listener can move in space while the sound is rendered in real-time. Second order IIR filters were implemented as wall-frequency dependent filters. The application is rendered binaurally with Google VR spatial API [4] which spatialises the final output with the ambisonics method according uniquely to the sound source and listener position [5]. Therefore, *ScatAR* do not implement any binaural rendering based on the direction of reflections. *ScatAR*'s open source project contains a prototype "shoebox" model that will be used as a basis for the current work (see Figure 4.5, right image).

ScatAR's implementation includes:

[2]ScatAR Github: `https://github.com/rampartisan/scatAR`

[3]Presently Google Tango is deprecated and replaced by ARCore platform.

[4]`https://developers.google.com/vr/reference/ios-ndk/group/audio`

[5]Spatialising a sound source with Google VR API consitsts simply in associating the listener's object with the API plugin and change some parameters. It is not coded directly into *ScatAR*'s implementation

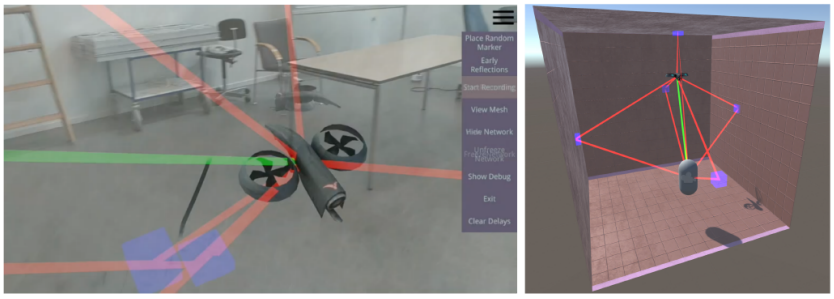

Figure 4.5: Illustration of ScatAR. **Left**: augmented reality environment, with a virtual drone as the sound source. **Right**: prototype scene of scatAR modeling a "shoebox" room in Unity 3D [78].

- A reflection finder algorithm based on the image source method, finding the first order reflection points and keeping track of the paths' information in an organised structure passed on to the main SDN script.
- A main SDN script attached to the source(s) and handling the propagation of the network in a sample-by-sample processing system. It implements omnidirectional sources, second order lowpass and highpass filter as wall absorption properties, adjustable wall absorption coefficient and omnidirectional microphone at the listener's position.
- An efficient triple thread handling, between the main Unity game thread (environment, positions, etc.), the Unity audio processing thread (a buffer based processor) and the SDN audio processing thread (the thread in which the sample-by-sample processing is handled).

4.2.2 Binaural Rendering of ScatAR

The binaural rendering implementation of the SDN further developed the efficient structure of *ScatAR* into a dynamic auralization system. Its implementation structure is depicted in Figure 4.6, and shows a simplified version of the system dependencies within the 3 processing threads.

Unity Game Thread

The thread in which reflection paths are computed, validated, updated and stored to be passed onto the main SDN process. An additional structure is calculating the azimuth and elevation of each nodes and source relative to the listener's head position and orientation. This information is passed to the **HRTF manager** which updates and interpolates each corresponding HRTF pairs, retrieved from **HRTF container**. The latter contains the personalised and zero-padded HRTF database loaded during the initialisation phase of the program.

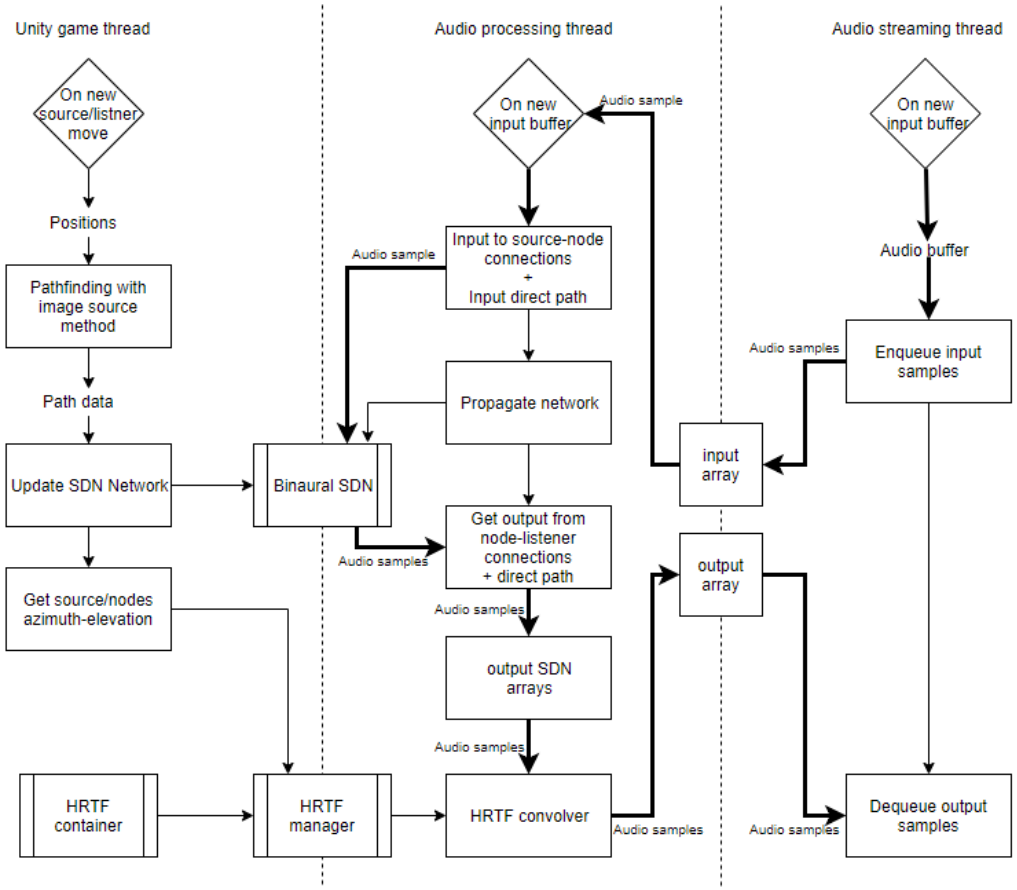

Figure 4.6: The binaural SDN implementation in Unity. Thicker lines represent the signal path of audio sample(s). Flowchart design derived from [79].

Audio Streaming Thread

The audio streaming thread is trivial. It enqueues the dry signal samples from its buffer in a First-In-First-Out queue to be passed onto the SDN thread. It also retrieves the 2-channel binaural processed outputs to be played over headphones.

Audio Processing Thread

The entire SDN processing is done in this thread. Incoming samples are pushed one by one into the source-node and source-listener connections. At this same rate, the rest of the network continues to propagate and the 7 outputs (6 node-listener plus the direct path) are accumulated in arrays of same size as the audio streaming thread buffer. When these arrays are full, they are ready to be convolved to the HRTFs in **HRTF convolver**. The convolution is processed by multiplication in the frequency domain, with FFT and IFFT operations using the *AForge.net* framework

[80]. Because convolution results in a signal that exceeds the buffer size[6], the convolution must be processed with the overlap-add method [81][7] over frames of double the buffer size. Finally, the 7 outputs of each channel (left and right) are respectively summed and sent back to the audio streaming thread.

Remarks

- The sampling rate and buffer size were fixed to 48 kHz and 1024 samples, respectively. Therefore, the OLA convolution were processed over 2048 samples.
- **HRTF container**'s task is to import the processed and personalised HRTFs from a text file, zero-pad and store them in a 3-D matrix (azimuth x elevation x samples) in their frequency representation to be ready for the OLA convolution process. Additionally, it contains a matrix of the Delaunay triangulation points for the interpolation process, implemented following Gamper's method [51] described in Section 2.3.2. Figure 4.7 shows the result of the Delaunay triangulation, with points and vertices, also showing the more condensed grid for azimuth angles between -45 and 45 degrees.

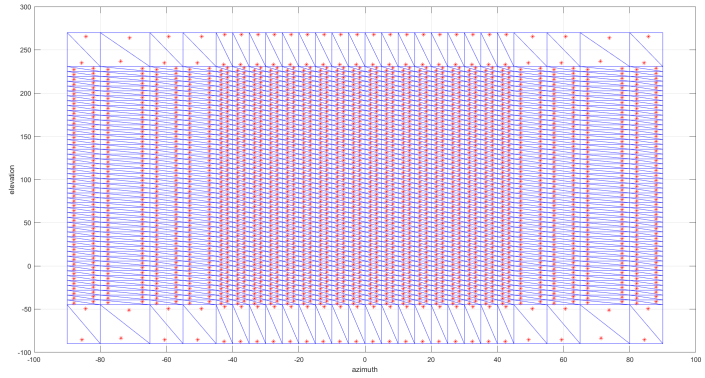

Figure 4.7: Delaunay triangulation obtained with the Matlab function *delaunayTriangulation* () after adding extrapolated points, for the interpolation process as described in [51]. The asterisk represent the points of (azimuth, elevation) pairs contained in their respective triangles.

[6] If N is the length of the first signal, M the length of the second signal, the convolution of these 2 signals result in a signal of length $L = N + M - 1$

[7] `https://ccrma.stanford.edu/~jos/sasp/Overlap_Add_OLA_STFT_Processing.html`

Material name	125 Hz	250 Hz	500 Hz	1000 Hz	2000 Hz	4000 Hz	8000 Hz
Concrete wall	0.02	0.02	0.03	0.03	0.04	0.05	0.05
Carpet Tile	0.27	0.26	0.52	0.43	0.51	0.58	-
Double glass	0.15	0.05	0.03	0.03	0.02	0.02	0.02
Vinyl floor	0.02	0.02	0.04	0.05	0.05	0.10	-
Perforated gypsum	0.45	0.55	0.60	0.90	0.86	0.75	-
Rockfon panel	0.30	0.70	0.85	0.90	0.90	0.85	0.60

Table 4.1: Absorption coefficients of employed materials in octave bands [84, p. 303]

4.2.3 Material Absorption Filter

An important part of room acoustic simulation is the modelling of material properties. Large databases of material acoustic properties exist [82]. Material absorption coefficients $\alpha\left(\omega_i\right)$ depend on frequency and direction, but in practice they are given independent of frequency in octave bands
$\omega = [125, 250, 500, 1000, 2000, 4000, 8000]$ Hz. Given these octave band values, low order digital filters can be designed to match the frequency response of a given material. Following the method in [83], we define the reflectance frequency magnitude response derived from the coefficients α as:

$$\mid R\left(j\omega\right) \mid = \sqrt{1 - \alpha\left(\omega\right)} \tag{4.1}$$

The magnitude response is then inter- and extrapolated and reduced to minimum-phase. Finally, a filter design is executed by least squares method (invfreq.m in Matlab).

Fourth order filters were designed and implemented in the SDN system. Figure 4.8 shows the magnitude frequency responses of the filter implemented in the current work, with octave band absorption coefficients listed in Table 4.1.

4.2.4 Room Parameterisation

The Binaural SDN was parameterised according to the two study case class rooms presented in Chapter 3. Wall surface materials were assigned according to an estimation of where the first order reflection point would occur and according to the availability in the material databases. Figure 4.9 shows VR modelling of the medium size class room.

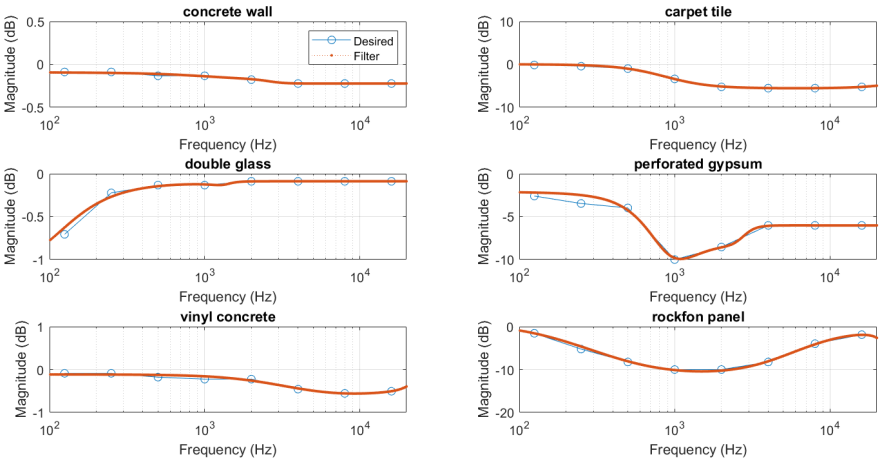

Figure 4.8: Magnitude response of absorption filters for materials in Table 4.1.

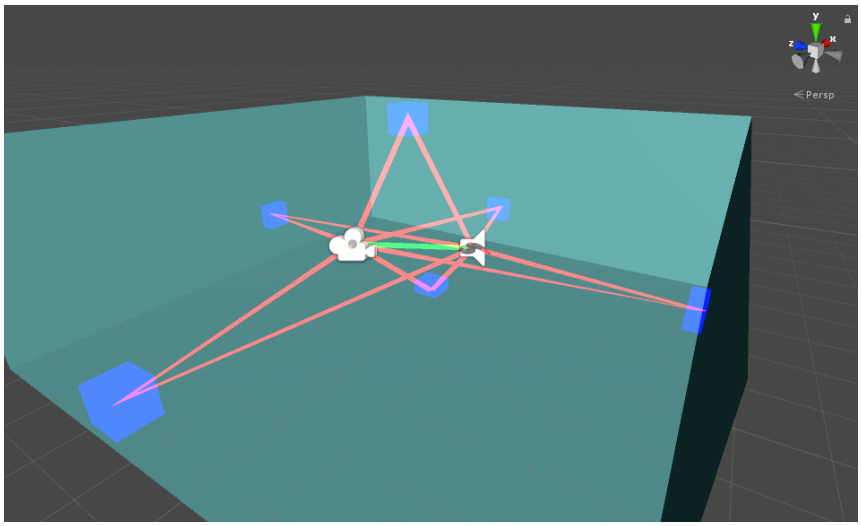

Figure 4.9: A screenshot of the modelled (9x9x3.4) m virtual room in Unity. The dimensions and wall materials of the room are parameterise to the case study "medium" class room. The source is represented by the speaker icon and the camera icon represents the listener. The red rays are the first order reflection paths connected to the listener via the nodes, the green ray is the direct sound path.

4.3 Ambisonics Processing

Within this section, the ambisonics auralization system is described.

4.3.1 SPARTA plugins

SPARTA (Spatial Audio Real-time Applications) is a collection of VSTs (Virtual Studio Technology) incorporating state-of-the-art techniques in spatial audio production, analysis and reproduction. These plugins are developed by the members of the Acoustics Lab at Aalto University[8]. The developers tested these plugins in the Digital Audio Workstation (DAW) Reaper[9].

Array2SH

This plugin converts microphone array signals to ambisonics signals. Its implementation details are described in [66], but the underlying principles are described in Section 2.5.3. It implements a regularised least-square encoder. The plugin offers control over the type of microphone, its geometry and the position of the sensors, as well the low-frequency regularisation method. Presets are already prepared for popular microphone array such as the Eigenmike. Figure 4.10 shows a screenshot of the graphical user interface. Observe the graphical EQ showing the regularised low-frequency compensation for each order (Tikhonov method), with the full colour lines representing the compensation made on the transparent lines.

The plugin was used to encode the 32-channels measured SRIRs to 4th-order ambisonic signals.

AmbiBIN

AmbiBIN, shown in Figure 4.11, is a binaural ambisonic decoder for reproducing ambisonic signals over headphones. It decodes the ambisonic signal to a virtual set of loudspeakers positioned with a t-design distribution, and renders binaurally by HRTF filtering (see Section 2.5.5). The number of loudspeakers depend on the ambisonic order selected. For 4th-order rendering, a t-design of degree 10 resulting in 60 loudspeaker is implemented.

Moreover, the plugin incorporates a sound field rotator (see Section 2.5.4) and headtracking support via OSC control[11], as well as the possibility to import HRTF dataset in the SOFA format (Spatially Oriented Format for Acoustics), a universal format for spatial audio data[12].

[8] http://research.spa.aalto.fi/projects/sparta_vsts/

[9] https://www.reaper.fm/

[11] http://opensoundcontrol.org/spec-1_0

[12] SOFA: https://www.sofaconventions.org/mediawiki/index.php/SOFA_(Spatially_Oriented_Format_for_Acoustics)

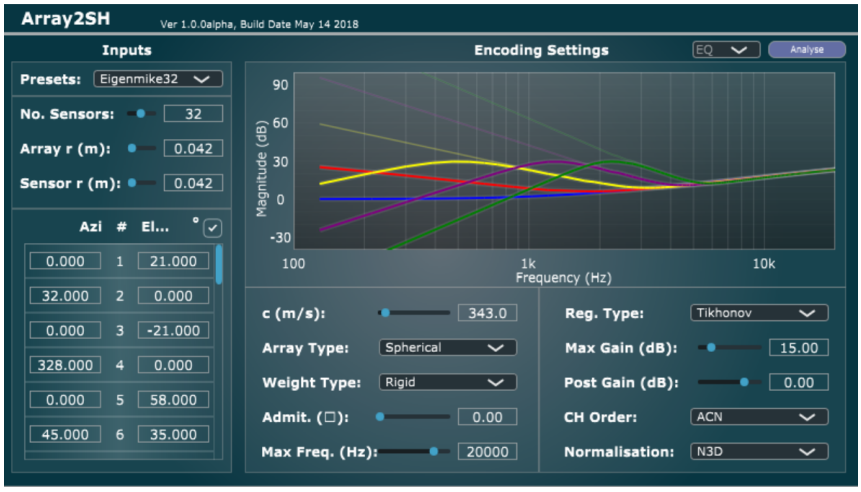

Figure 4.10: Screenshot of SPARTA's Array2SH. Left panel: microphone technical description with number of sensors, microphone radius and position (direction) of sensors. Bottom-middle panel: additional microphone information. Bottom-right pannel: regularisation parameters (Tikhonov) and ambisonics conventions (channel order and normalisation)[10]. Top panel: graphical EQ for regularisation filters effect, one line for each order (here, N = 4).

This plugin was used for the ambisonic rendering.

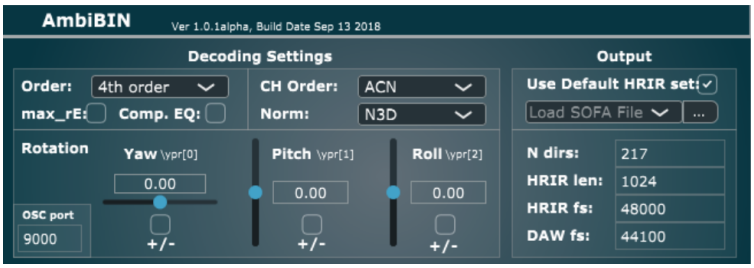

Figure 4.11: Screenshot of SPARTA's AmbiBIN.

4.3.2 Ambisonics System Design

The ambisonic auralization system is depicted in Figure 4.12. First, the 32-channel recorded SRIR is encoded into 4th-order ambisonics (25 channels) with Array2SH plugin in Reaper. Then, following the method presented in Section 2.5.6, each ambisonic channel is convolved with a mono dry audio signal, thus creating an acoustic scene with a sound source placed at the specific position where the loudspeaker was placed when the SRIR was recorded.

The real-time binaural rendering of the sound scene is performed in Reaper with the AmbiBIN plugin. Headtracking values from the Oculus headset are sent from Unity via OSC messages using Garcia's implementation [85].

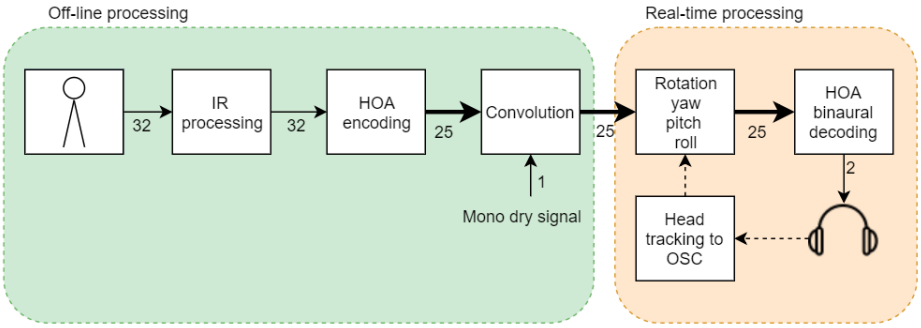

Figure 4.12: Overview of the work flow for the HOA rendering approach. The numbers next to the arrows represent the number of channels that line contains, with a thicker line for ambisonic signals.

Chapter 5

Evaluation

5.1 Objective Analysis

5.1.1 HRTF personalisation: a case study on the KEMAR

This first part of the objective evaluation presents a horizontal and median plane study of the personalisation of the KEMAR HATS with large pinnae from the CIPIC library which anthropometric dimensions and pictures are available in the library. The personalisation of the KEMAR resulted in the selection and customisation of CIPIC human subject 060.

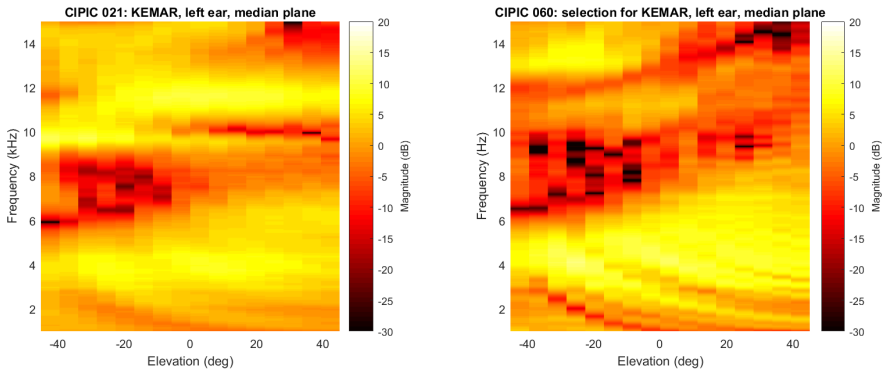

Figure 5.1: Frequency magnitude response comparison, median plane for elevation between $-45°$ and $45°$ and frequency range 1-15 kHz.

Figure 5.1 shows a comparison between the left ear frequency magnitude response of both the KEMAR and its personalised human version, in the median plane (azimuth $0°$) for elevations ranging from $-45°$ to $45°$ and frequency range 1-15 kHz. Similarities can be seen, in particular frequency notches appearing in the frequency range 6-10 kHz, a frequency region sensitive to the pinna shape, thus

supporting the success of the image-guided selection software.

Furthermore, Figure 5.2 shows the horizontal plane (elevation 0°) frequency magnitude responses of the left ear for all azimuths, starting from azimuth $-90°$ (the point at the left) and with a clockwise rotation. Here, the frequency range is set to 0-1 kHz to illustrate the effect of the spherical head shadow effect model, effective up to the cutoff frequency 1 kHz (see Eq. (2.14)). The head shadow model introduces a more important contrast between a source emitting from the ipsilateral and contralateral side of the ear than in the case of the measured HRTF. In particular, as can be seen on the right plot, low frequencies are attenuated up to 15 dB for a source situated in the contralateral side of the ear. Therefore, the low end of the spectrum of the measured (selected) HRTF dataset is appropriately corrected according to physical properties of diffraction mapped to the size of the head and in the same time prevents the potential lack of low frequencies from HRTF databases.

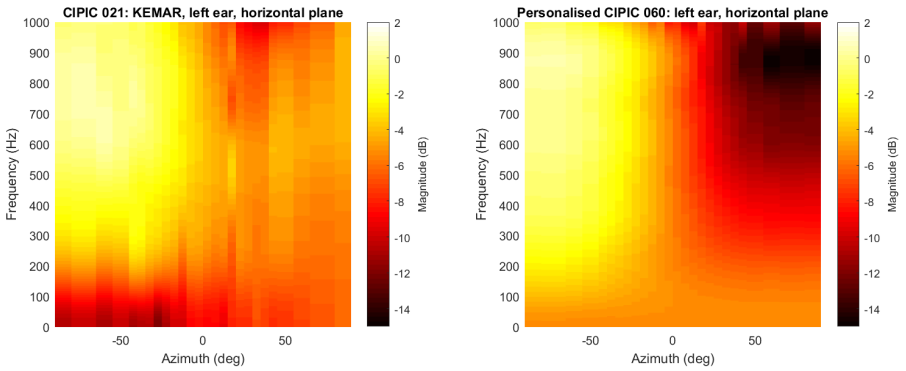

Figure 5.2: Frequency magnitude response comparison, horizontal plane for all azimuth and frequency range 0-1 kHz.

In addition to the frequency analysis above, a time-domain comparison is made between the original KEMAR dataset and its personalisation. Figure 5.3 shows the ITD curve between both datasets in the horizontal plane and for all azimuths. ITDs we obtained from the time difference between onsets of respective HRIR pairs. The figure illustrates an excellent match, therefore supporting the ITD estimation obtained from Woodworth's formula (Eq. (2.5)) and the regression formula for an optimal head radius (2.7).

5.1.2 Measured vs Modelled BRIR

This section presents a comparison between measured and modelled BRIRs on a signal-level. The analysis concerns uniquely the binaural SDN with a focus on

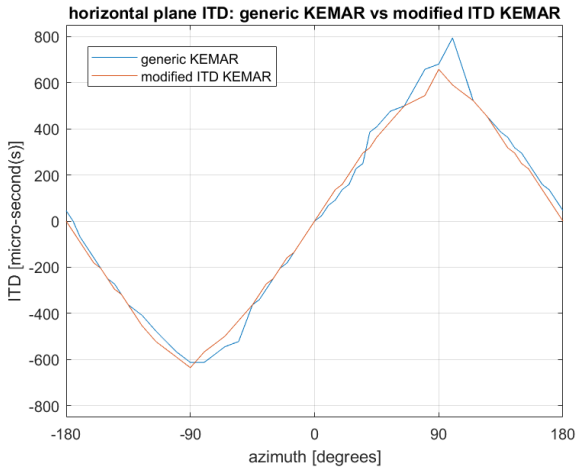

Figure 5.3: Original vs personalised ITD.

early reflections, overall room decay and spectral content of the impulse response. The binaural SDN was parameterised according to the 2 measured class rooms presented in Chapter 3. The room dimensions were straightforward parameters to enter in the SDN model. Wall surface materials were chosen according to the surfaces on which the first order reflections occurred. Individual wall absorption coefficients, influencing the room decay time, were adjusted manually to best fit the target room decay time. The final parameter values for both rooms are listed in Table 5.1, where each wall is assigned a wall id, wall name, surface material and absorption coefficient. The HRTF dataset used for this analysis is a full sphere dataset from the B&K HATS 4100 model which is the model corresponding to the one utilised for the BRIR measurements of this thesis project [1]. It contains HRTFs measured in a 2 degrees steps.

For convenience, the following analysis is restricted to position P1, P2 and P5. In the figures, these 3 positions are associated to a colour, i.e. blue for P1, red for P2 and green for P5.

Large view BRIR: Figure 5.4 and Figure 5.5

These figures show the first 150 ms (medium room) and 200 ms (large room) of the time-domain responses of the left and right ears. The top panels of each positions represent the measured BRIRs, whereas the bottom panels represent the modelled ones. In these plots, it can be observed clear differences in reflection density between the measured and the modelled BRIRs. However, the SDN shows some conformity with the measured responses of this room, notably in terms of

[1] The dataset was provided by B&K.

room	wall id	wall name	material	abso. coeff.
medium	1	right	concrete	0.96
	2	front	concrete	0.96
	3	left	window	0.96
	4	back	concrete	0.96
	5	ceiling	gypsum	0.96
	6	floor	vinyl	0.96
large	1	right	concrete	0.90
	2	front	concrete	0.90
	3	left	window	0.91
	4	back	concrete	0.93
	5	ceiling	rockfon	0.97
	6	floor	carpet	0.97

Table 5.1: SDN parameters calibrated to the 2 measured rooms: wall IDs mapped to wall names (direction of the walls relative to the listener), wall surface materials and wall absorption coefficients.

early reflection timing and overall decay. Note that the dynamic range of the reflections are small compared to the direct sound because of the short distance (1.2 m) between the sound sources and the listening point relative to the room dimensions.

dB scale BRIR: Figure 5.6 and Figure 5.7

These figures show the first 300 ms (medium room) and 500 ms (large room) of the time-domain dB-scale responses for the left and right ears. The modelled BRIRs are superimposed to the measured ones in order to observe the linear overall time decay match between both.

Zoom BRIR: Figure 5.8 and Figure 5.9

These figures show a zoom on the first 35 ms (medium room) and 65 ms (large room) of the time-domain responses for the left and right ears. In these plots, the simulated BRIRs are once again superimposed on the measured ones. It is interesting to observe the accuracy in time of arrivals of early reflections which are numbered from 1 to 6 according to the wall mapping of table 5.1. The time of arrival of reflections 5 and 6, representing the ceiling and floor, respectively, are very precisely modelled, whereas the reflections coming from lateral walls can be slightly shifted. This can be due to a difference in positions of source/listener between the measured and modelled scenarios or, simply because of the non-homogeneous wall surfaces from the measured class rooms. For instance, the wall on the left (number 3) of both rooms contains cavities for windows that are deeper than the

concrete wall and that are not modelled by our SDN system. This can be observed on reflection 3 of the medium room.

Furthermore, it can be observed a difference in amplitude between the reflections of measured and modelled BRIRs. One reason could be a lack of source directivity model in our SDN system. In fact, the measurement loudspeaker has its own directivity pattern which is direction and frequency dependent. Therefore, amplitudes of signals/waves leaving from the back of the loudspeaker are attenuated relative the ones emitted to the front. Another reason could be a difference in wall absorption coefficients between the reality and the simulation.

Spectrogram: Figure 5.10 and Figure 5.11

In these figures are plotted the left ear spectrograms for the 3 selected positions (left column: measured, right column: modelled). It can be observed that frequencies are more pronounced over the whole spectrum for the simulated BRIRs, notably the high and low frequencies present for longer period. This shows the difficulty in matching the frequency response of a room with a minimum amount of modelled elements (a network of 6 nodes) with only a few low order filters. Moreover, the measured class rooms were not empty, thus modifying the acoustic response both temporally and spectrally compared to a totally empty room (as modelled by our SDN system).

T30: Figure 5.10 and Figure 5.11

The last 2 figures concern the room acoustical criteria reverberation time T30, defined in ISO3382 [86]. The reverberation times are plotted by octave bands for the 6 positions and for both ears of measured and modelled BRIRs. For the medium room, these plots show a relatively similar reverberation time behaviour for frequencies above 2 kHz. However, the reverberation time is longer for low frequencies for the modelled BRIRs, which was also seen in the spectrogram plots of Figure 5.10. On the other hand, the large room shows differences up to 7 kHz, with longer reverberation times in the low end of the spectrum, below 1.5 kHz and, shorter ones above that frequency. This shows the difficulty in calibrating this large room, which had a longer and more damped reverberation tale.

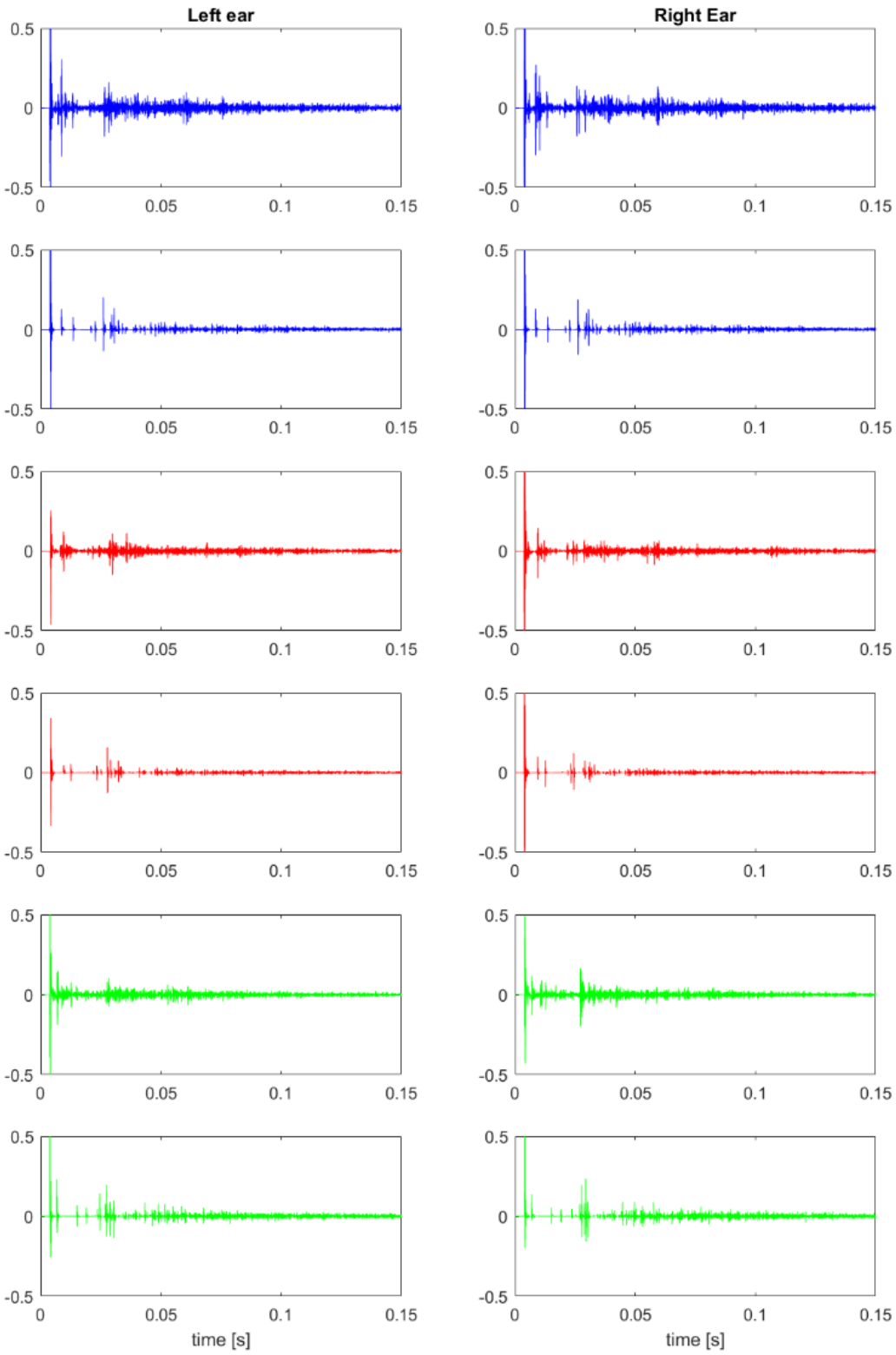

Figure 5.4: medium room: First 150 ms of Measured and Modelled BRIRs for position P1 (blue), P2 (red), P5 (green). Measured on the top panel, Modelled on the bottom panel. Note that the signals were normalised to the maximum value of both channels, respectively, thus keeping the ratio between left and right. Note also that the y-axis was cut to $[-0.5; 0.5]$ to obtain more details on the plot.

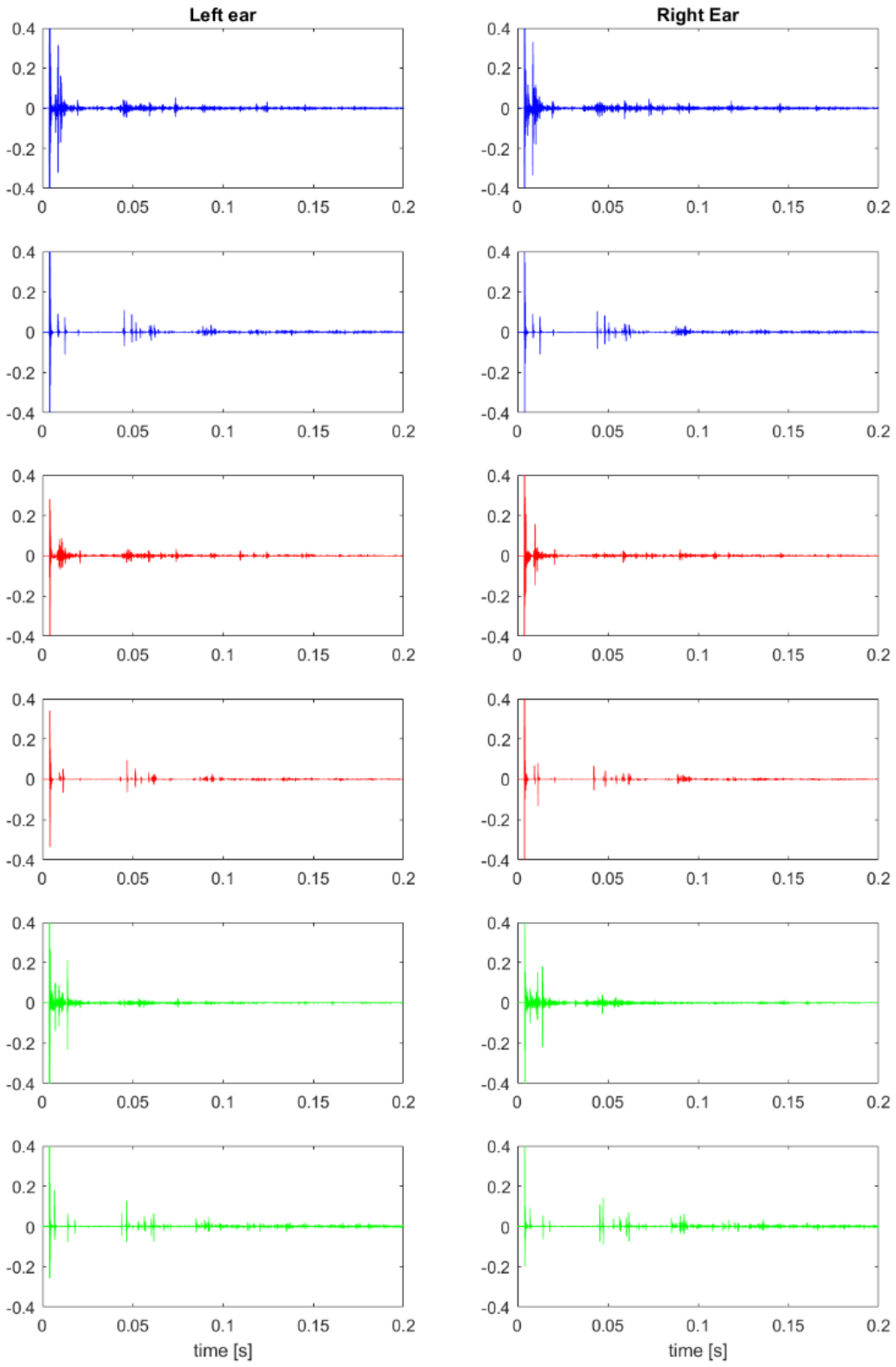

Figure 5.5: large room: First 200 ms of Measured and Modelled BRIRs for position P1 (blue), P2 (red), P5 (green). Note that the signals were normalised to the maximum value of both channels, respectively, thus keeping the ratio between left and right. Note also that the y-axis was cut to $[-0.4; 0.4]$ to obtain more details on the plot.

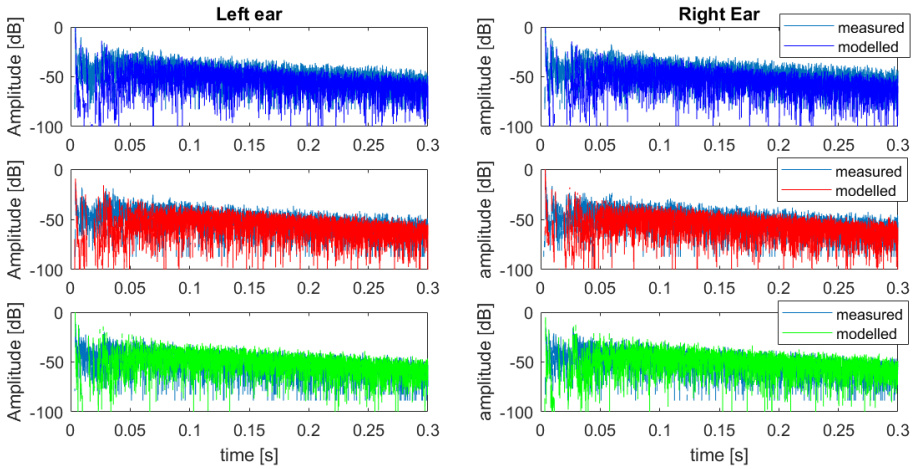

Figure 5.6: medium room: First 300 ms of Measured and Modelled BRIRs in dB for position P1 (blue), P2 (red), P5 (green).

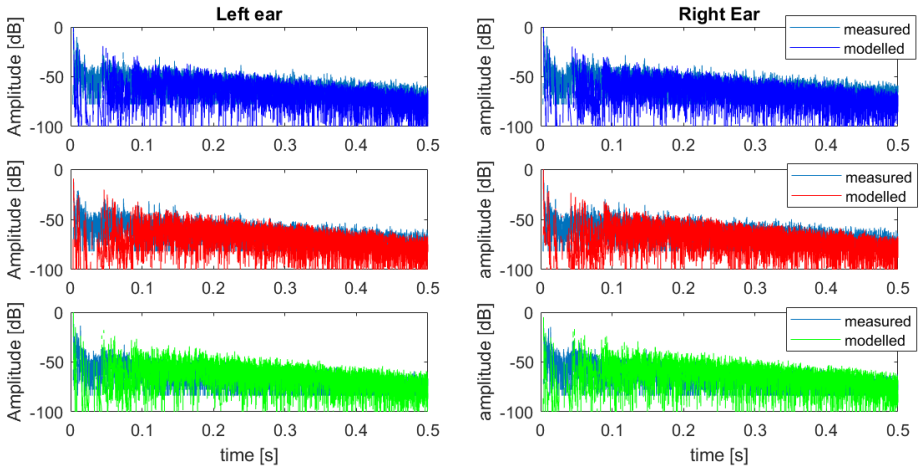

Figure 5.7: large room: First 500 ms of Measured and Modelled BRIRs in dB for position P1 (blue), P2 (red), P5 (green).

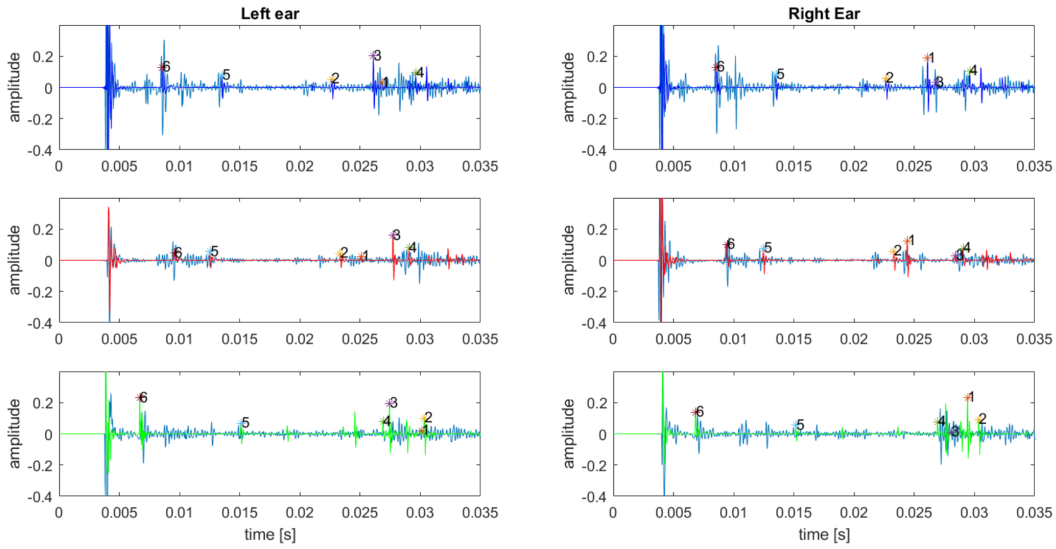

Figure 5.8: medium room: Early reflections of Measured and Modelled BRIRs for position P1 (blue), P2 (red), P5 (green).

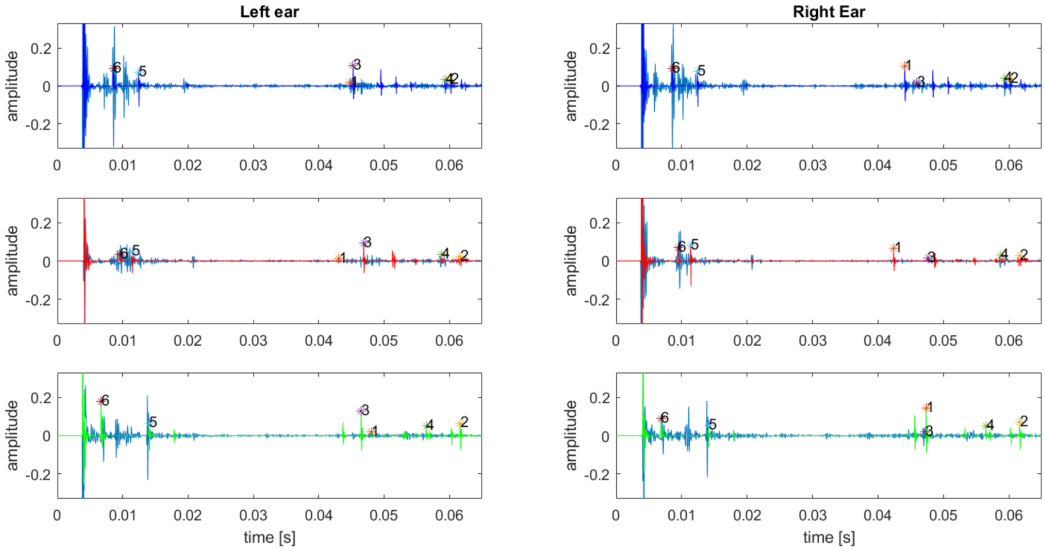

Figure 5.9: large room: Early reflections of Measured and Modelled BRIRs for position P1 (blue), P2 (red), P5 (green).

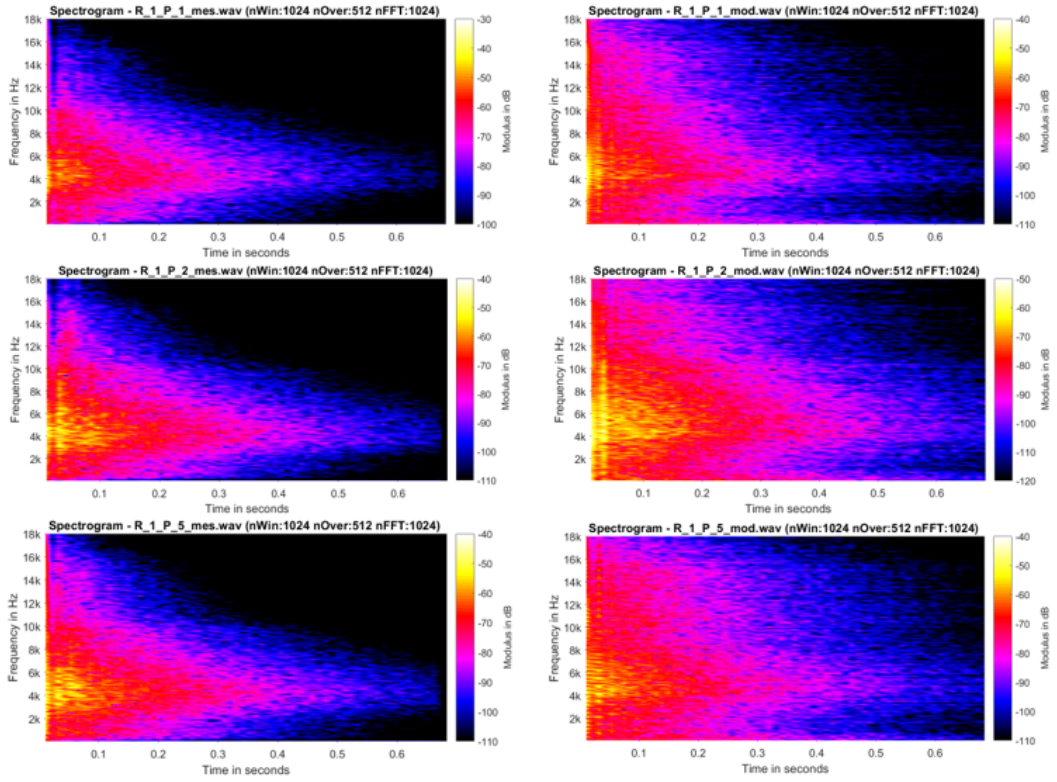

Figure 5.10: medium room: Spectrogram comparison of left ear for position P1 (top), P2 (middle) and P5 (bottom). Left column: measured, Right column: simulated

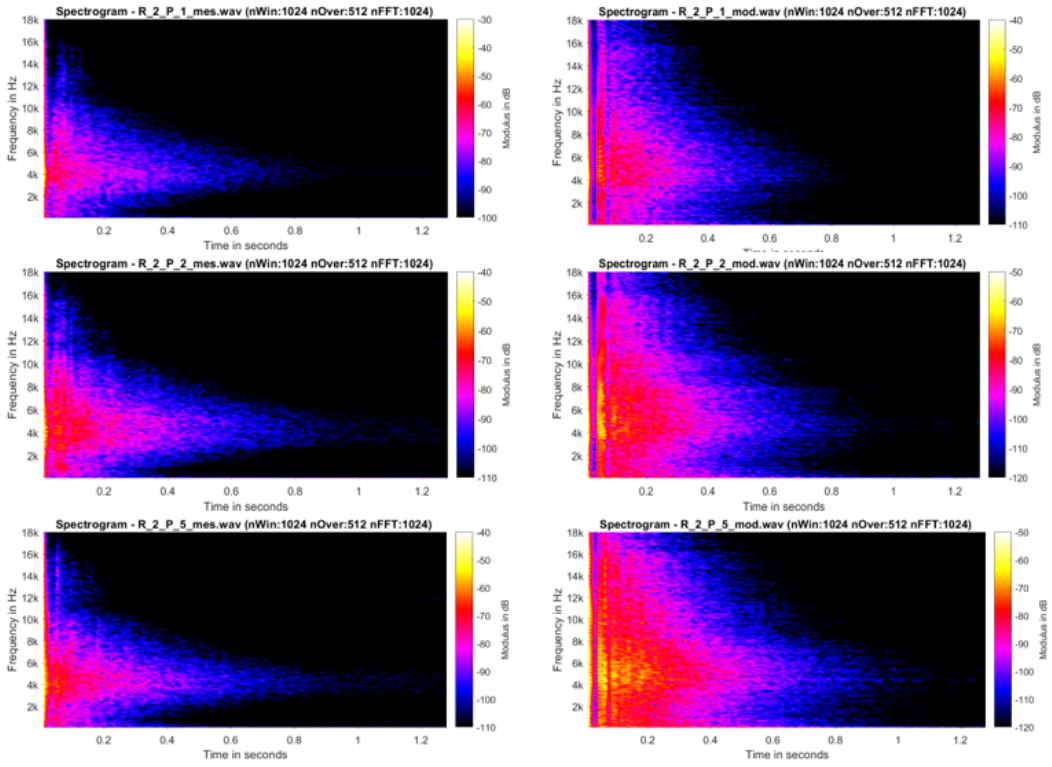

Figure 5.11: large room: Spectrogram comparison of left ear for position P1 (top), P2 (middle) and P5 (bottom). Left column: measured, Right column: simulated.

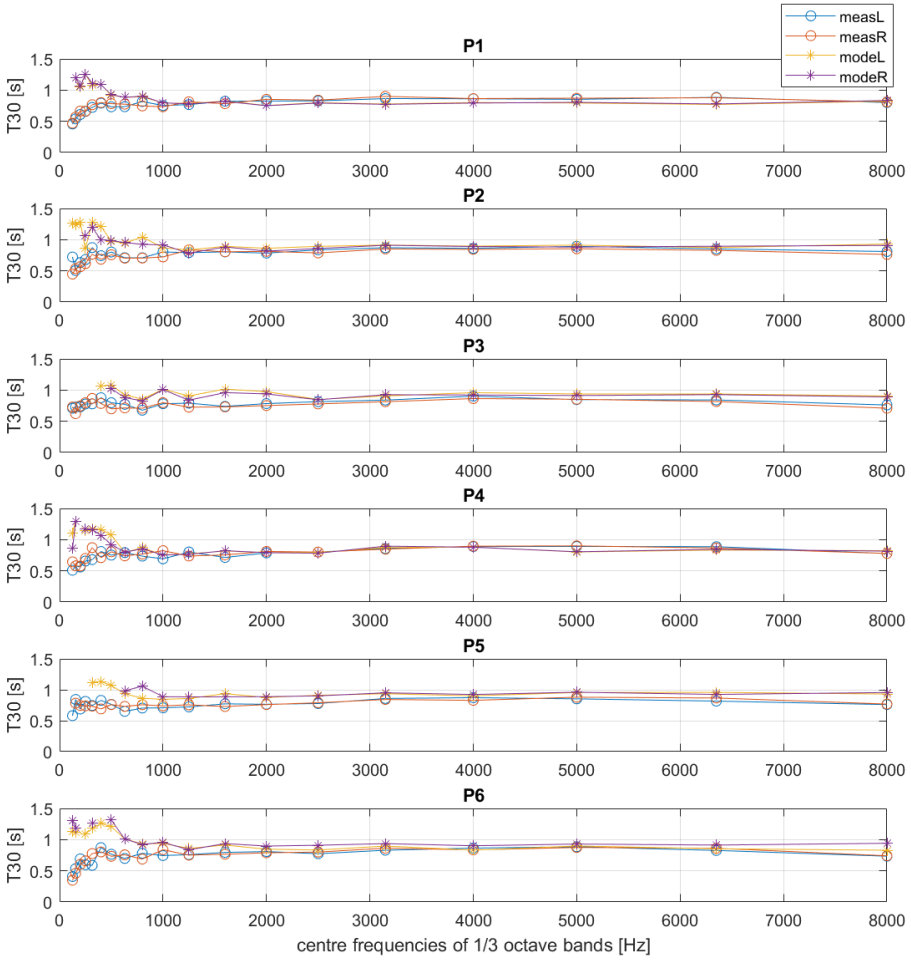

Figure 5.12: medium room: T30 comparison by octave band for the 6 positions.

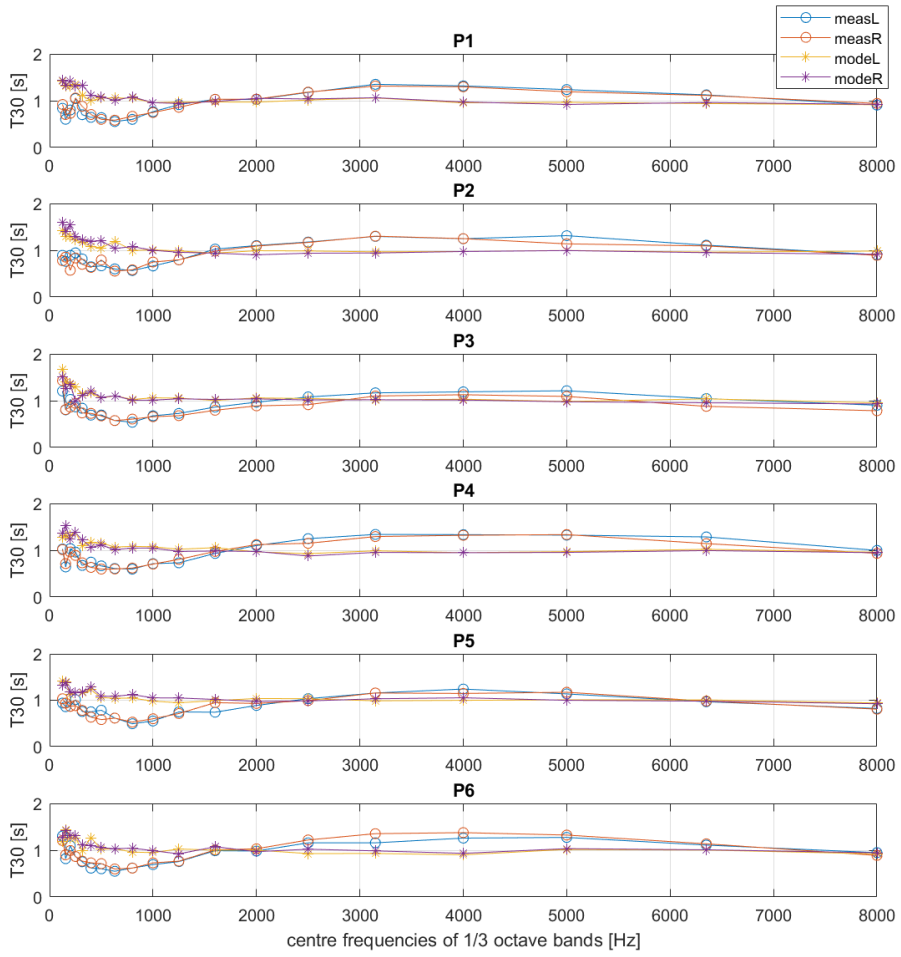

Figure 5.13: large room: T30 comparison by octave band for the 6 positions.

5.2 Localisation Test

5.2.1 Experimental Design

A localisation test serves as an efficient first approach to subjectively evaluate a binaural rendering system utilising different sets of HRTFs. The aim of this experiment is to evaluate the accuracy in sound localisation (azimuth, elevation and reversals errors) and externalisation judgements for the personalisation process implemented in the SDN and HOA auralization system.

To evaluate all combinations of conditions and variables to be studied the experimental design shown in Table 5.2 was developed. The SDN and the HOA auralization systems were both tested with a generic dataset of HRTF and the subject's personalised one. Furthermore, the 3 room conditions, "anechoic", "medium" and "large" were also included as testing conditions. Therefore, the experimental variables are challenged by 12 (2 x 2 X 3) conditions which were evaluated on each subjects.

Our 6 measured directions, with azimuth and elevation from Table 3.1, were each rendered by the 12 experimental conditions, resulting in a total of 72 trials, with four experimental blocks on audio rendering conditions presented as follow:

- The four main blocks, presented in a Latin square order were:
    1. The SDN system with Generic HRTF (**SDN+G**)
    2. The SDN system with Personalised HRTF (**SDN+P**)
    3. The HOA system with Generic HRTF (**HOA+G**)
    4. The HOA system with Personalised HRTF (**HOA+P**)
- For each block, 3 room acoustic conditions also presented in a Latin square order:
    1. Anechoic (**A**)
    2. Medium(**M**)
    3. Large (**L**)
- Finally, 6 directions presented in a random order.

5.2.2 Subjects

Nine subjects, 8 males and 1 female, average age 27.8 (SD: 3.2) years old, with self-reported normal hearing and no motor impairments, participated in the experiment. All males are part of the Sound and Music Computing community, therefore all having general knowledge in the field of Audio. One of them is currently a 3-D Audio developer. The female participant had no experience.

5.2.3 Apparatus

Unity Environment

Based on the implementation of Chapter 4, the experiment was set up in a VR environment, shown in Figure 5.14. In order to have a minimum visual influence on the auditory cues, a minimalistic scene was created where the listener was placed in the middle of a dark room where only the floor was illuminated by a directional light spot on top of the listener (illuminated radius of 3 meters). On the ground, under the listener, a cross was placed to help the user's orientation in the VR space: a blue and a red line for the north-south and east-west directions, respectively. Head orientation was visually rendered in the scene through a virtual laser pointer located at the centre of the head. The initial position of the listener was instructed to be seated with head straight so that the head would go through the blue circle placed ahead.

The right Oculus controller, from which a laser was coming out, served as the tool for saving the coordinates of the perceived sound source location by pointing to the direction and press the trigger button. The coordinates correspond to the point of contact between the laser and an invisible 2-meters radius sphere centred at the listeners head position (see Figure 5.14). When the trigger was pressed, the point of contact was converted from Cartesian to spherical coordinate system to save the azimuth and elevation of the point on a text file. Note determining the direction of the sound source by "head pointing" or "hand pointing" have been investigated in the literature, with results showing similar outcomes with both methods [87].

The second utility of the controller was to save the perceived externalisation: the externalisation was rated using the controller joystick to move the slider placed in front of the listener (see Figure 5.14). The small grey transparent sphere was representing the head of the listener. The slider, of length 3 times longer than the sphere radius and starting from inside the sphere, was defined as the perceived distance of the sound source. This continuous scale externalisation rating system prevents the bias of directly asking to the subject the "Boolean" type question "in-head or out-of-the-head localisation?" [21].

Table 5.2: Localisation test experimental conditions and variables

Rendering condition	SDN						HOA					
HRTF condition	Generic			Personalised			Generic			Personalised		
Room condition	Anechoic	Medium	Large	Anechoic	Medium	Large	Anechoic	Medium	Large	Anechoic	Medium	Large
Variables												
Azimuth error	x	x	x	x	x	x	x	x	x	x	x	x
Elevation error	x	x	x	x	x	x	x	x	x	x	x	x
Reversals	x	x	x	x	x	x	x	x	x	x	x	x
Externalisation	x	x	x	x	x	x	x	x	x	x	x	x

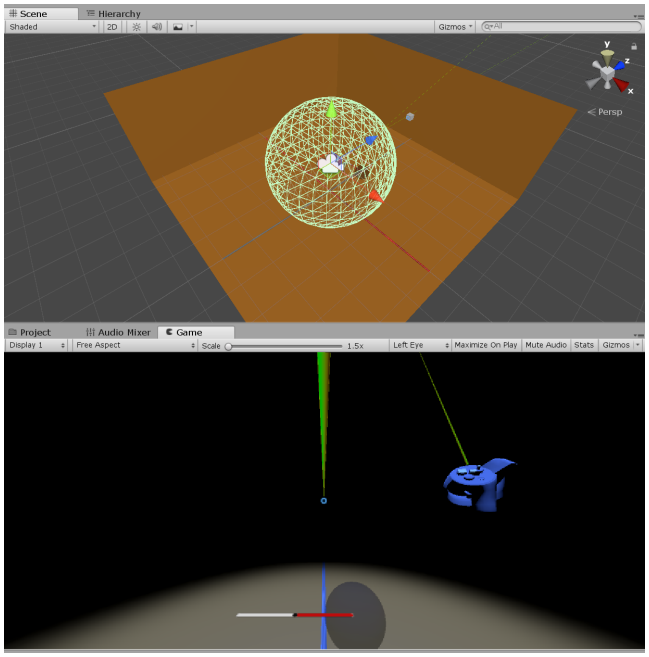

Figure 5.14: Localisation test VR environment. Top figure: scene view. Bottom figure: game camera view.

Hardware and software

Based on the implementation of Chapter 4, both auralization systems were using the Oculus Rift headset running in Unity for the headtracking information. The binaural SDN was as well run in Unity, whereas the HOA system was run through the SPARTA ambiBIN plugin in Reaper.

Stimuli were played back through Sennheiser HD-600 headphones which were equalised using the APO-Equaliser software[2], an equaliser capable of low-latency convolution with custom frequency-response. The inverse of an average of Sennheiser HD-600 headphone transfer functions measured on more than a 100 human subjects[3] was fed to the equaliser.

5.2.4 Stimuli

Audio stimuli

Stimuli were composed of a train of 3 consecutive 40 ms white Gaussian noise (WGN) bursts, with 30 ms of silence between each bursts, repeated 6 times for a total of 3 seconds. This type of audio stimuli has proven to be effective in localisation tests [43]. The presentation level of the stimuli was approximately 60 dBA for

[2]https://sourceforge.net/projects/equalizerapo/
[3]http://sofacoustics.org/data/headphones/ari

the frontal position, a level corresponding to normal speech at 1 m. The level was measured with a sound level meter placed in one of the headphone's ear cap.

SDN condition

The medium and the large rooms of the binaural SDN system were calibrated to the measured BRIRs with the parameters from the objective analysis of Section 5.1.2. The anechoic stimuli were simply the rendering of the direct sound path only, disabling the incoming wall reflections from the network. The dry WGN bursts signal was fed into the system to be rendered according to the experimental conditions and source positions.

HOA condition

For the HOA system, the stimuli were prepared offline in Matlab. The room stimuli were generated by convolving the dry WGN bursts signals to our 12 (2 rooms, 6 positions) 4th-order Ambisonics-RIRs. The anechoic stimuli were generated by simply encoding the dry WGN bursts into 4th-order Ambisonic signals at the specific experimental positions. Signals were then exported to WAV format to be rendered by the AmbiBIN plugin in Reaper.

HRTF conditions

The HRTF database used for this evaluation was the CIPIC database. The generic HRTF dataset was the one of the KEMAR HATS with large pinnae (CIPIC subject 165), with added extrapolated points for the interpolation algorithm. The personalised dataset was constructed for each experimental participants from the proposed personalisation procedure presented in Section 4.1. The datasets were then exported in their appropriate format, i.e. text file format for the SDN system and SOFA file format for the HOA rendering.

5.2.5 Experimental Procedure

During the localisation test, subjects were seated on an office chair with 360° rotation. The chair was set so that the head position would be at 1.3 m height (same height as the RIR measurement setup). The task consisted in locating the audio stimuli, pointing the laser with a straight arm, triggering to save the point, adjusting the distance (externalisation) slider and come back to the initial positioning to wait for the next stimuli. Participants were free to rotated their head, their body and the office chair.

The test was preceded by a VR scene exploration and a training session of 6 trials in a arbitrary room rendered by the SDN with an arbitrary HRTF dataset. The training room, rendered by the SDN, and HRTF dataset was the same for all

subjects. The training stimuli and procedure were the same used for the experiment. The goal of the training session was to ensure that the participant correctly understood the experimental procedure, and get acquainted with the VR scene and controller.

The localisation test followed the 8 block structure described in section 5.2.1, with a short break after the completion of 2 blocks (18 trials), allowing the experimenter to load the next HRTF set if needed. Operations were controlled and triggered manually by the experimenter, due to the many different platforms and condition changes during this short localisation test (HRTF datasets, Rooms, SDN in Unity, HOA in Reaper). The test duration was on average of 1 hour and 10 minutes, including the personalisation process and training session.

5.2.6 Results

A 2 x 2 x 3 (rendering type: SDN/HOA; HRTF type: generic/personalised; room: anechoic/medium/large) 3-way repeated measures analysis of variance (ANOVA) was conducted for each of the four dependent measures of Table 5.2: azimuth error, elevation error, front–back/back-front reversal rate and externalisation. An alpha level of 0.05 was used for all statistical tests. If pair wise post-hoc analysis was performed, a Bonferroni correction was applied on p-values. A preliminary analysis on data distributions were subjected to Levene's test for homoscedasticity, and inspections of normality in linear model residuals according to Shapiro-Wilk test showed that the main assumptions were not violated.

Azimuth Error

Azimuth error was defined as the unsigned error between the target and judged azimuth direction. Judged azimuth were corrected for front-back and back-front reversal errors, i.e. if the judged azimuth direction was opposite the target response relative to the interaural axis, the judged azimuth was flipped on the other side of this axis. The azimuth error for each experimental conditions was calculated from the average of unsigned errors across all evaluated directions and for all subjects. Figure 5.15 shows the mean azimuth error for all conditions.

The main effect for rendering type was non-significant, but a significant main effect was found for HRTF type ($F(1,8) = 83$, $p < 0.001$) and room type ($F(2,16) = 9$, $p < 0.002$). Therefore, the personalisation had a positive effect on azimuth localisation (see Figure 5.17); and large rooms produced a general degradation of performance (post-hoc: $p < 0.05$). The analysis also revealed a significant 2-way interaction between rendering and HRTF type ($F(1,8) = 7.3$, $p = 0.027$), with a significant advantage of SDN+P over SDN+G ($p < 0.001$), HOA+P over HOA+G ($p < 0.01$) and HOA+G over SDN+G ($p < 0.05$) (see Figure 5.16). This confirms that the personalisation is efficient in terms of azimuth for both rendering systems.

Figure 5.15: Average azimuth error for all conditions

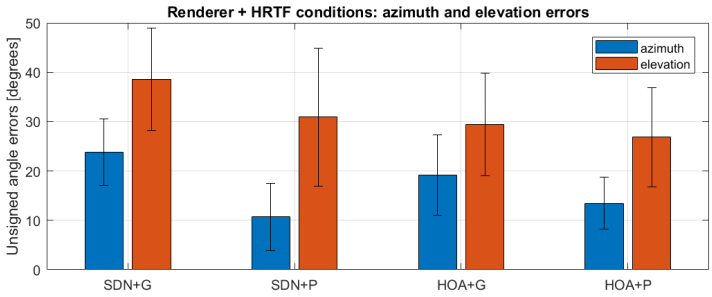

Figure 5.16: Mean azimuth and elevation error for rendering + HRTF condition

Table from Figure 5.18 shows the mean values for unsigned azimuth errors for individual subjects in comparison to the overall mean, grouped by experimental condition. It can be noticed that the statistically significant trends from the overall analysis is constant for all subjects that performed better in azimuth localisation with the personalised HRTF condition and with less accuracy when room reverberation was added. Furthermore, the SDN with personalised HRTF condition (in green) outperforms the HOA with generic HRTF condition for all the subjects. These results can be observed even for subject ZC which had the worst performance in azimuth localisation.

Elevation Error

Elevation error was defined as the unsigned error between the target and judged elevation direction. The elevation error for each experimental conditions was calculated from the average of unsigned errors across all evaluated directions and for all subjects. Figure 5.19 shows the mean elevation error for all conditions.

A significant main effect was found for rendering type ($F(1,8) = 7.2$, $p = 0.027$) with a better performance of HOA rendering (mean = 28.1° ; SD = 1.8°) over the SDN rendering (mean = 34.7°; SD = 1.7°) condition. However, in contrast to the

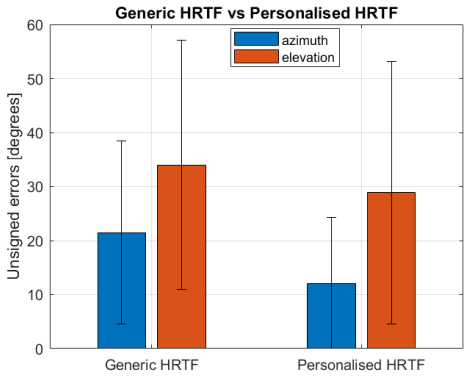

Figure 5.17: Average azimuth and elevation error for HRTF condition

azimuth errors, the room condition and the 2-way interaction between rendering and HRTF type were statistically non significant.

Table from Figure 5.20 shows the mean values for unsigned elevation errors for individual subjects in comparison to the overall mean, grouped by experimental condition. In contrast to the azimuth performances, accuracy in elevation judgement was prone to a high variability across and within participants which is reflected in the plot of Figure 5.19 and in the statistical analysis. Five subjects were bad elevation localiser (ZA, ZC, ZD, ZE and ZF), with an average error above 30° across all conditions. In particular, ZE appeared as a particular case of performing worse with the personalised HRTF condition with an average of 34.7° error for personalised and 29.6° for generic condition. Also, Subject ZB's personalisation was effective for the SDN condition, with an improvement of 8.5° on average, but induced more errors to the subject in the HOA condition, with a decrease of 8.5° precision on average. Note that ZB started the experiment with the HOA+personalised condition and was perhaps still in the learning process. On the other hand, ZG, ZH and ZI were good elevation localisers with also a personalisation that enhanced accuracy in their judgement with average errors of 38.7 (SDN+G), 24.7 (SDN+P), 27.1 (HOA+G) and 21.7 (HOA+P). Note also that for these 3 subjects, SDN+P outperforms HOA+G.

Reversals Error

Reversal errors are defined as the total of front-back and back front-error. Average reversal percentage rates for all conditions are shown in Figure 5.21. Results of the ANOVA test reveal a significant main effect for the rendering condition ($F(1,8) = 7.6$, $p = 0.024$), HRTF condition ($F = (1,8) = 91$, $p < 0.001$) and a 2-way interaction between rendering and HRTF type ($F(1,8) = 22.3$, $p = 0.002$). Therefore indicating significantly:

	SDN	Generic		SDN	Perso		HOA	Generic		HOA	Perso		
	A	M	L	A	M	L	A	M	L	A	M	L	
ZA	18.82	21.88	16.63	5.76	7.08	11.65	25.41	10.45	27.66	10.15	10.90	11.86	Mean
	17.59	18.95	11.77	4.41	5.37	8.34	30.26	14.75	24.21	10.94	9.31	9.36	Std
ZB	24.55	29.04	26.56	8.81	14.17	35.75	30.01	30.97	27.57	20.31	11.43	17.50	Mean
	23.66	18.76	10.28	6.95	18.86	26.04	19.90	22.92	12.89	15.62	16.65	18.77	Std
ZC	19.98	27.96	30.93	8.06	5.47	19.05	5.23	15.97	12.18	4.56	13.39	10.25	Mean
	16.43	14.73	19.61	2.64	4.08	16.14	4.75	12.97	11.05	2.55	11.40	9.61	Std
ZD	17.11	19.09	40.26	2.93	7.70	17.13	11.62	23.86	26.84	15.62	10.36	25.02	Mean
	9.06	15.33	26.68	3.36	5.09	12.97	16.24	9.87	19.87	14.52	9.87	29.20	Std
ZE	23.86	13.75	34.35	5.11	9.34	10.63	20.54	38.86	27.74	10.35	15.80	10.55	Mean
	19.07	11.51	20.10	3.83	6.23	8.13	7.95	26.18	21.88	13.66	6.36	3.54	Std
ZF	20.44	39.89	15.08	16.67	7.68	20.01	9.90	20.48	19.82	9.18	17.58	26.36	Mean
	16.90	28.93	10.40	12.96	10.71	21.90	10.21	22.87	13.69	6.62	12.63	22.21	Std
ZG	24.22	23.67	22.85	2.42	6.76	9.54	10.64	19.55	11.22	7.83	15.22	21.08	Mean
	21.14	13.65	14.72	2.07	4.52	5.98	9.92	20.03	12.44	5.62	14.52	11.54	Std
ZH	24.90	27.26	19.15	4.11	10.93	11.33	10.28	12.57	17.15	18.26	8.59	9.73	Mean
	14.35	10.71	15.54	2.59	14.11	8.86	5.42	7.87	10.83	7.97	9.96	7.43	Std
ZI	18.87	19.82	22.31	8.22	11.62	10.97	15.23	18.60	16.69	9.77	11.87	10.50	Mean
	3.72	12.03	14.26	1.31	3.29	1.93	13.56	9.38	7.43	4.33	7.48	2.35	Std
MEAN	21.42	24.71	25.35	6.9	8.97	16.23	15.43	21.26	20.76	11.78	12.79	15.87	
STD	2.97	7.51	8.45	4.33	2.78	8.31	8.19	8.96	6.84	5.16	2.92	6.76	

Figure 5.18: AZIMUTH errors and standard deviations (std) for all conditions and for each subjects. Each subject is denoted by two letters (ZA, ZB, ...). Overall mean and std values are in the the red box.

- more reversals for the SDN (mean = 12.3, SD = 1.2) over HOA (mean = 7.7, SD = 1.3) condition,
- more reversals for generic HRTF (mean = 17.2, SD = 1.3) over personalised (mean = 2.7, SD = 1.0) condition,
- a reversal rate reduction with SDN compared to HOA in the generic HRTF condition (post-hoc $p < 0.001$), and a further reduction from generic to personalised in SDN (post-hoc $p < 0.001$)

Externalisation

A sound is externalised if it was perceived out of the head. The externalisation factor was normalised so that the edge of the sphere was set to 0.25. Considering a verged-cranial localisation, i.e. in close proximity of the head, as an in-head localisation, the threshold for externalisation was set to 0.35.

Mean externalisation judgement for all conditions are shown in Figure 5.22. On average, non of the anechoic stimuli was perceived as externalised. Statistically, a significant main effect was found for the rendering type ($F(1,8) = 8.2$, $p = 0.021$) and the room type ($F(2,16) = 21.2$, $p < 0.001$), with a mean of 0.543 (SD = 0.025) for the SDN and 0.488 (SD = 0.026) for the HOA condition and, medium and large rooms stimuli perceived as more externalised over anechoic stimuli ($p < 0.01$).

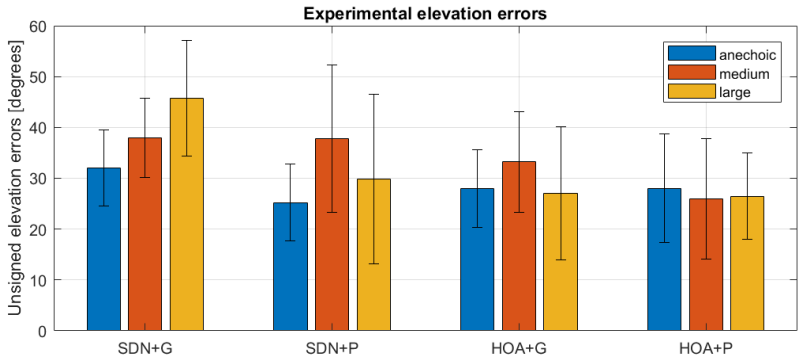

Figure 5.19: Average elevation error for all conditions

	SDN	Generic		SDN	Perso		HOA	Generic		HOA	Perso		
	A	M	L	A	M	L	A	M	L	A	M	L	
ZA	23.49	32.92	36.29	33.63	12.74	56.73	46.05	35.90	57.61	47.40	16.01	37.61	Mean
	20.80	24.32	20.46	25.60	7.38	34.17	33.85	22.59	20.43	42.03	15.55	33.11	Std
ZB	27.12	33.73	36.40	29.02	25.48	17.25	26.55	21.59	13.09	37.78	16.92	31.46	Mean
	15.35	24.63	22.04	14.67	11.53	22.35	15.32	10.41	9.50	37.94	10.30	23.33	Std
ZC	44.37	34.23	62.45	20.32	53.66	52.99	23.06	43.00	27.38	25.84	25.57	19.32	Mean
	18.95	18.24	30.79	14.36	43.56	37.32	13.15	10.93	19.64	22.53	25.16	12.46	Std
ZD	30.21	48.75	57.50	25.56	43.48	40.89	22.29	30.53	26.62	26.96	28.64	24.03	Mean
	17.58	20.33	37.91	13.50	37.05	40.12	15.71	22.46	14.31	15.78	8.54	20.32	Std
ZE	34.95	33.91	42.43	39.07	40.93	27.20	23.89	27.28	15.44	37.45	27.45	36.34	Mean
	26.24	24.88	20.41	16.98	26.34	17.51	19.43	11.25	8.39	20.91	29.85	25.80	Std
ZF	43.44	35.80	35.40	23.85	46.26	22.77	31.52	53.44	24.62	15.37	53.64	20.55	Mean
	28.42	26.72	14.99	18.10	20.36	22.11	26.88	28.03	27.29	10.93	31.02	15.37	Std
ZG	28.16	29.69	33.40	21.08	52.99	25.47	29.21	26.83	31.66	24.34	26.66	29.42	Mean
	19.51	28.26	31.90	16.11	38.60	23.09	20.75	14.45	16.69	18.34	16.47	23.99	Std
ZH	30.02	39.97	50.18	18.67	43.20	12.56	21.08	34.49	18.08	19.28	13.23	27.83	Mean
	20.88	20.99	28.77	10.46	23.68	8.14	13.13	34.56	12.17	19.92	13.28	13.50	Std
ZI	26.32	52.82	57.79	15.61	21.19	12.42	27.81	25.91	28.67	18.03	25.77	11.49	Mean
	18.96	22.88	28.61	8.23	7.80	3.99	21.87	23.30	15.31	10.11	28.62	2.05	Std
MEAN	32.01	37.98	45.76	25.2	37.77	29.81	27.94	33.22	27.02	28.05	25.99	26.45	
STD	7.44	7.81	11.34	7.54	14.49	16.65	7.61	9.91	13.11	10.7	11.83	8.47	

Figure 5.20: ELEVATION errors and standard deviations for all conditions and for each subjects. Each subject is denoted by two letters (ZA, ZB, ...). Overall mean and std values are in the the red box.

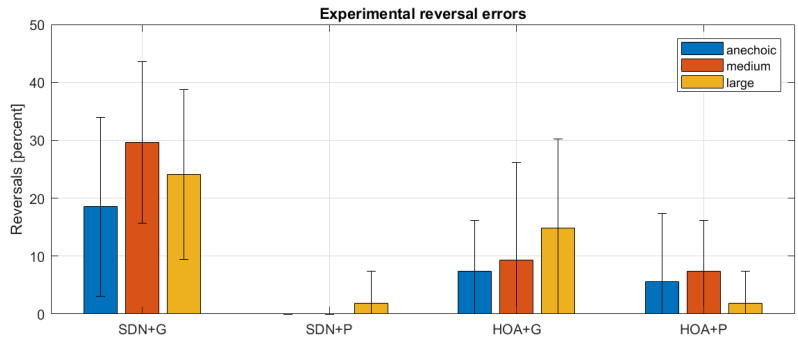

Figure 5.21: Percentage reversal errors for all conditions

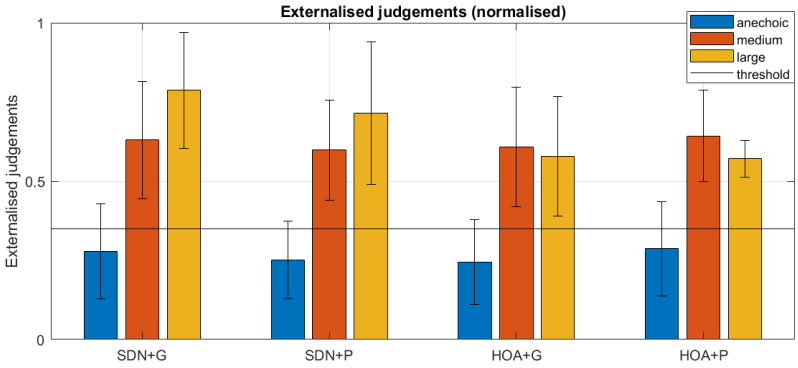

Figure 5.22: Normalised externalisation judgements for all conditions. The black horizontal represent the externalisation threshold.

Chapter 6

Discussion

6.1 Personalisation Wins

Results from the localisation test showed that our HRTF personalisation has an important impact on localisation performances. It has significantly reduced azimuth errors, in general and within both auralization systems. Since azimuth localisation is mainly influenced by ITD and ILD, it indicates that the head model implemented is effective. Furthermore, the increasing error from anechoic to large rooms could be due to a typical reverberation effect: sound reflections introduces a blurring effect that broadens the auditory image and makes it more difficult to localise. The more sparse are these reflections the more blurring this image can be.

Elevation-wise, results were more difficult to interpret. A high variability occurred between participants and no statistical conclusion could be brought regarding the HRTF condition. Nevertheless, the perception of elevation was present for all subjects and three of them achieved good results in the lines of expectations. In the literature, Begault also found some high variability between participants for elevation perception with reverberation conditions [21]. Shinn-Cunningham states that "*judgment of the up/down direction of a source may be biased because reverberant energy tends to alter the mean spectral shape of the signals reaching the listener*" [23]. In the current work, possible explanations for the high error variability could be that the testing VR environment lacked of spatial visual reference points. Indeed, the dark room with only a cross on the ground and no reference above head level could have introduced a disorientation to the subjects. Therefore, visual reference points such as dome above the head with median and horizontal marks is recommended [88]. Furthermore, the localisation test was relatively short in duration, with only a few stimuli per condition. A change of HRTF dataset was happening at least 2 times and at most 4 times in less than 45 minutes which could be misleading for the cortical processing. This could be interpreted that the listeners have a better adaptation to temporal (ITD) than frequency dependent cues.

In addition, results from the localisation test also showed an improvement in front-back and back-front reversal errors which were considerably reduced when sound was rendered with the personalisation. This difference is strong in the SDN condition with almost a zero percent rate with the personalised HRTFs.

Based on these results, our 2-D image-guided HRTF personalisation outperforms the use of a generic set of HRTFs.

6.2 Minimal, but External

In this work, the 4th-order ambisonic auralization system was rendering recorded sound fields. Although limited by the number of spherical harmonics it retains a substantial amount of the acoustic properties of the rooms. On the other hand, the SDN auralization system was simulating these rooms with a minimum set of room parameters and was rendering accurately only essential acoustic properties. Nevertheless, combined with an appropriate binaural system and room parameterisation, the SDN showed that it is capable of bringing perceptually relevant features to the user, notably in terms of externalisation of a sound source which is one of the main issues regarding headphone based 3D Audio. In the localisation test, results from the externalisation judgement revealed a significant improvement compared to anechoic stimuli, which proves that the SDN system is at the level of the measurement-based HOA system. The results also showed more externalisation for the SDN compared to the HOA system. Meaning that the subjects perceived the sound coming from farther for the SDN. Since it is known that distance perception is partly related to early reflections (mainly floor and ceiling reflections), this suggests that the reflections from the SDN were not appropriately calibrated in terms of amplitude. If we look back at the signal comparisons between measured and modelled BRIRs (Figure 5.8 and 5.9), the reflections have a relatively precise timing, but the amplitudes are variable. The most probable cause is the lack of source directivity in our SDN model (see Figure 10 in [52]).

6.3 Future Works

Numerous paths still needs to be explored, notably in terms of the quality of the rendering. Indeed, the localisation test gave a first subjective impression on the potentiality of the binaural SDN system. However, source position is only one of the many perceptual attributes of sound and gives little information of what a person perceives. The next logical step for this work is to evaluate the *quality of experience* (QoA). A state of the art framework is the Spatial Audio Quality Inventory (SAQI) from Lindau et al. proposing a list 48 descriptors of audititory qualities of a spatial audio reproduction system (binaural or loudspeaker based) grouped into 8 categories (timbre, tonalness, geometry, room, time behaviour, dynamics, artifacts, and

general impressions) [89]. Similarly, Simon et al. proposed a list of eight attributes aiming at evaluating the quality of a binaural rendering system with the use of non-individual HRTFs and to understand the perceptual variance involved [90]. In this way, realistic signals such as speech and music would be rendered through the systems, with potentially more than one source.

Secondly, the binaural SDN could be improved. For instance, source directivity is an important element in an auralization system. The authors of the SDN proposes a simple direction dependent gain, according to microphone directivity patterns such as cardioid or hypercardioid [52]. More advanced frequency-dependent directivity filters were developed by Lokki in his DIVA auralization program [91]. He proposes a diffuse field and equalised diffuse field filters which are direction, frequency and perceptually dependent. Another improvement could be the phenomena of thermal relaxation which can be observed as an increasing low-pass filtering as a function of distance relative to the source [92].

Chapter 7

Conclusion

This thesis presented the design, implementation and evaluation of an anthropometry-based HRTF personalisation integrated in two dynamic auralization systems, namely the Scattering Delay Network (SDN) and the common Ambisonics approach. The HRTF personalisation was designed as a straightforward 2-D image-guided process for the extraction of three anthropometric dimensions to create a personalised HRTF dataset from a combination of an optimised spherical head model and a HRTF dataset selected from a database. The systems were evaluated in a virtual reality environment in terms of localisation performance and externalisation. While results show significant improvements in localisation accuracy with the proposed personalisation method, its combination with the SDN system revealed a simple, flexible and effective fully customisable binaural reverberator that is promising for future development and integration in virtual and augmented reality systems.

Bibliography

[1] Durand R. Begault. *3D Sound for Virtual Reality and Multimedia*. San Diego, CA, USA: Academic Press Professional, Inc., 1994. ISBN: 0-12-084735-3.

[2] S. Serafin, M. Geronazzo, C. Erkut, N. C. Nilsson, and R. Nordahl. "Sonic Interactions in Virtual Reality: State of the Art, Current Challenges, and Future Directions". In: *IEEE Computer Graphics and Applications* 38.2 (2018), pp. 31–43. ISSN: 0272-1716. DOI: 10.1109/MCG.2018.193142628.

[3] Mendel Kleiner, Bengt-Inge Dalenbäck, and Peter Svensson. "Auralization-An Overview". In: *J. Audio Eng. Soc* 41.11 (1993), pp. 861–875. URL: http://www.aes.org/e-lib/browse.cfm?elib=6976.

[4] Vesa Välimäki, Julian Parker, Lauri Savioja, Julius O. Smith, and Jonathan Abel. "More Than 50 Years of Artificial Reverberation". In: *Audio Engineering Society Conference: 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)*. 2016. URL: http://www.aes.org/e-lib/browse.cfm?elib=18061.

[5] Jens Blauert. *The Technology of Binaural Listening*. 1st. Springer-Verlag Berlin Heidelberg, 2013. ISBN: 978-3-642-37761-7, 978-3-642-43946-9.

[6] Elizabeth Wenzel, Marianne Arruda, Doris J. Kistler, and Frederic L. Wightman. "Localization using nonindividualized head-related transfer functions". In: 94 (Aug. 1993), pp. 111–23.

[7] Jan-Gerrit Richter, Gottfried Behler, and Janina Fels. "Evaluation of a Fast HRTF Measurement System". In: *Audio Engineering Society Convention 140*. 2016. URL: http://www.aes.org/e-lib/browse.cfm?elib=18197.

[8] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. "The CIPIC HRTF database". In: *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*. 2001, pp. 99–102. DOI: 10.1109/ASPAA.2001.969552.

[9] O. Warusfel. *Listen HRTF database*. 2002. URL: http://recherche.ircam.fr/equipes/salles/listen/.

[10] Austrian Academy of Sciences Acoustics Research Institute (ARI). *ARI Database*. URL: https://www.kfs.oeaw.ac.at/index.php?option=com_content&view=article&id=608&Itemid=606&lang=en.

[11] Ramona Bomhardt. *Anthropometric Individualization of Head-Related Transfer Functions Analysis and Modeling*. Zenodo, Sept. 2017. DOI: 10.5281/zenodo.885037. URL: https://doi.org/10.5281/zenodo.885037.

[12] Carlile S. *The Physical and Psychophysical Basis of Sound Localization. In: Virtual Auditory Space: Generation and Applications. Neuroscience Intelligence Unit*. Berlin, Heidelberg: Springer, 1996.

[13] George F. Kuhn. "Model for the interaural time differences in the azimuthal plane". In: *The Journal of the Acoustical Society of America* 62.1 (1977), pp. 157–167. DOI: 10.1121/1.381498. eprint: https://doi.org/10.1121/1.381498. URL: https://doi.org/10.1121/1.381498.

[14] C. P. Brown and R. O. Duda. "A structural model for binaural sound synthesis". In: *IEEE Transactions on Speech and Audio Processing* 6.5 (1998), pp. 476–488. ISSN: 1063-6676. DOI: 10.1109/89.709673.

[15] Jens Blauert and Robert A. Butler. "Spatial Hearing: The Psychophysics of Human Sound Localization by Jens Blauert". In: *The Journal of the Acoustical Society of America* 77.1 (1985), pp. 334–335. DOI: 10.1121/1.392109. eprint: https://doi.org/10.1121/1.392109. URL: https://doi.org/10.1121/1.392109.

[16] Jack Hebrank and D. Wright. "Are two ears necessary for localization of sound sources on the median plane?" In: *The Journal of the Acoustical Society of America* 56.3 (1974), pp. 935–938. DOI: 10.1121/1.1903351. eprint: https://doi.org/10.1121/1.1903351. URL: https://doi.org/10.1121/1.1903351.

[17] Ruth Y. Litovsky, H. Steven Colburn, William A. Yost, and Sandra J. Guzman. "The precedence effect". In: *The Journal of the Acoustical Society of America* 106.4 (1999), pp. 1633–1654. DOI: 10.1121/1.427914. eprint: https://doi.org/10.1121/1.427914. URL: https://doi.org/10.1121/1.427914.

[18] Christof Faller and Juha Merimaa. "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence". In: *The Journal of the Acoustical Society of America* 116.5 (2004), pp. 3075–3089. DOI: 10.1121/1.1791872. eprint: https://doi.org/10.1121/1.1791872. URL: https://doi.org/10.1121/1.1791872.

[19] Donald H Mershon, William L Ballenger, Alex D Little, Patrick L McMurtry, and Judith L Buchanan. "Effects of Room Reflectance and Background Noise on Perceived Auditory Distance". In: *Perception* 18.3 (1989). PMID: 2798023, pp. 403–416. DOI: 10.1068/p180403. eprint: https://doi.org/10.1068/p180403. URL: https://doi.org/10.1068/p180403.

[20] M. Barron. "Late lateral energy fractions and the envelopment question in concert halls". In: *Applied Acoustics* 62.2 (2001), pp. 185 –202. ISSN: 0003-682X. DOI: https://doi.org/10.1016/S0003-682X(00)00055-4. URL: http://www.sciencedirect.com/science/article/pii/S0003682X00000554.

[21] Durand R. Begault, Elizabeth M. Wenzel, and Mark R. Anderson. "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source". In: *J. Audio Eng. Soc* 49.10 (2001), pp. 904–916. URL: http://www.aes.org/e-lib/browse.cfm?elib=10175.

[22] W. M. Hartmann. "Localization of sound in rooms". In: *The Journal of the Acoustical Society of America* 74.5 (1983), pp. 1380–1391. DOI: 10.1121/1.390163. eprint: https://doi.org/10.1121/1.390163. URL: https://doi.org/10.1121/1.390163.

[23] Barbara G. Shinn-Cunningham, Norbert Kopco, and Tara J. Martin. "Localizing nearby sound sources in a classroom: Binaural room impulse responses". In: *The Journal of the Acoustical Society of America* 117.5 (2005), pp. 3100–3115. DOI: 10.1121/1.1872572. eprint: https://doi.org/10.1121/1.1872572. URL: https://doi.org/10.1121/1.1872572.

[24] Harold Schlosberg Robert Sessions Woodworth. *Experimental psychology*. 1940.

[25] V. Ralph Algazi, Carlos Avendano, and Richard O. Duda. "Estimation of a Spherical-Head Model from Anthropometry". In: *J. Audio Eng. Soc* 49.6 (2001), pp. 472–479. URL: http://www.aes.org/e-lib/browse.cfm?elib=10188.

[26] R. O. Duda and W. L. Martens. "Range-dependence of the HRTF for a spherical head". In: *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*. 1997, 5 pp.–. DOI: 10.1109/ASPAA.1997.625597.

[27] Hélène Bahu and David Romblom. "Optimization and Prediction of the Spherical and Ellipsoidal ITD Model Parameters Using Offset Ears". In: *Audio Engineering Society Conference: 2018 AES International Conference on Spatial Reproduction - Aesthetics and Science*. 2018. URL: http://www.aes.org/e-lib/browse.cfm?elib=19599.

[28] Neil L. Aaronson and William M. Hartmann. "Testing, correcting, and extending the Woodworth model for interaural time difference". In: *The Journal of the Acoustical Society of America* 135.2 (2014), pp. 817–823. DOI: 10.1121/1.4861243. eprint: https://doi.org/10.1121/1.4861243. URL: https://doi.org/10.1121/1.4861243.

[29] R. O. Duda, C. Avendano, and V. R. Algazi. "An adaptable ellipsoidal head model for the interaural time difference". In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*. Vol. 2. 1999, 965–968 vol.2. DOI: 10.1109/ICASSP.1999.759855.

[30] Tomi Huttunen, Asta Kärkkäinen, Leo Kärkkäinen, Ole Kirkeby, and Eira T. Seppälä. "Some Effects of the Torso on Head-Related Transfer Functions". In: *Audio Engineering Society Convention 122*. 2007. URL: http://www.aes.org/e-lib/browse.cfm?elib=14015.

[31] Richard O. Duda, V. Ralph Algazi, and Dennis M. Thompson. "The Use of Head-and-Torso Models for Improved Spatial Sound Synthesis". In: *Audio Engineering Society Convention 113*. 2002. URL: http://www.aes.org/e-lib/browse.cfm?elib=11294.

[32] V. Ralph Algazi, Richard O. Duda, Ramani Duraiswami, Nail A. Gumerov, and Zhihui Tang. "Approximating the head-related transfer function using simple geometric models of the head and torso". In: *The Journal of the Acoustical Society of America* 112.5 (2002), pp. 2053–2064. DOI: 10.1121/1.1508780. eprint: https://doi.org/10.1121/1.1508780. URL: https://doi.org/10.1121/1.1508780.

[33] Michele Geronazzo, Simone Spagnol, and Federico Avanzini. "Estimation and Modeling of Pinna-Related Transfer Functions". In: *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10)*. Graz, Austria, Sept. 2010, pp. 431–438. ISBN: 978-3-200-01940-9.

[34] Simone Spagnol, Michele Geronazzo, and Federico Avanzini. "On the Relation Between Pinna Reflection Patterns and Head-Related Transfer Function Features". In: *Trans. Audio, Speech and Lang. Proc.* 21.3 (Mar. 2013), pp. 508–519. ISSN: 1558-7916. DOI: 10.1109/TASL.2012.2227730. URL: http://dx.doi.org/10.1109/TASL.2012.2227730.

[35] S. Spagnol, D. Rocchesso, M. Geronazzo, and F. Avanzini. "Automatic extraction of pinna edges for binaural audio customization". In: *2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)*. 2013, pp. 301–306. DOI: 10.1109/MMSP.2013.6659305.

[36] Michele Geronazzo, Jacopo Fantin, Giacomo Sorato, Guido Baldovino, and Federico Avanzini. "Acoustic Selfies for Extraction of External Ear Features in Mobile Audio Augmented Reality". In: *Proc. 22nd ACM Symposium on Virtual Reality Software and Technology (VRST 2016)*. Munich, Germany: ACM, Nov. 2016, pp. 23–26. ISBN: 978-1-4503-4491-3. DOI: 10.1145/2993369.2993376.

[37] Brian F. G. Katz. "Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation". In: *The Journal of the Acoustical Society of America* 110.5 (2001), pp. 2440–2448. DOI: 10.1121/1.1412440. eprint: https://doi.org/10.1121/1.1412440. URL: https://doi.org/10.1121/1.1412440.

[38] Hannes Gamper, David Johnston, and Ivan Tashev. "Interaural time delay personalisation using incomplete head scans". In: IEEE, 2017. URL: https://www.microsoft.com/en-us/research/publication/interaural-time-delay-personalisation-using-incomplete-head-scans/.

[39] H. Gamper, M. R. P. Thomas, and I. J. Tashev. "Estimation of multipath propagation delays and interaural time differences from 3-D head scans". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, pp. 499–503. DOI: 10.1109/ICASSP.2015.7178019.

[40] D. Y. N. Zotkin, J. Hwang, R. Duraiswaini, and L. S. Davis. "HRTF personalization using anthropometric measurements". In: *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)*. 2003, pp. 157–160. DOI: 10.1109/ASPAA.2003.1285855.

[41] Michele Geronazzo, Enrico Peruch, and Fabio Prandoni. "Improving Elevation Perception with a Tool for Image-guided Head-related Transfer Function Selection". In: *Proceedings of the 20th International Conference on Digital Audio Effects (DAFx-17)*. 2017.

[42] M. Geronazzo, S. Spagnol, and F. Avanzini. "Do We Need Individual Head-Related Transfer Functions for Vertical Localization? The Case Study of a Spectral Notch Distance Metric". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.7 (2018), pp. 1247–1260. ISSN: 2329-9290. DOI: 10.1109/TASLP.2018.2821846.

[43] M. Geronazzo, S. Spagnol, A. Bedin, and F. Avanzini. "Enhancing vertical localization with image-guided selection of non-individual head-related transfer functions". In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014, pp. 4463–4467. DOI: 10.1109/ICASSP.2014.6854446.

[44] Fabio Prandoni. "A virtual reality environment with personalized spatial audio rendering". MA thesis. Italy: UNIVERSITA DEGLI STUDI DI PADOVA, 2017.

[45] Michele Geronazzo, Erik Sikstroöm, Jari Kleimola, Federico Avanzini, Amalia de Götzen, and Stefania Serafin. "The impact of a good vertical localization with HRTFs in short explorations of immersive virtual reality scenarios". In: *Proceedings of the IEEE International Symposium for Mixed and Augmented Reality 2018 (To appear)*. 2018.

[46] Doris J. Kistler and Frederic L. Wightman. "A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction". In: *The Journal of the Acoustical Society of America* 91.3 (1992), pp. 1637–1647. DOI: 10.1121/1.402444. eprint: https://doi.org/10.1121/1.402444. URL: https://doi.org/10.1121/1.402444.

[47] Michele Geronazzo, Simone Spagnol, and Federico Avanzini. "Mixed Structural Modeling of Head-Related Transfer Functions for Customized Binaural Audio Delivery". In: *Proc. 18th Int. Conf. Digital Signal Process. (DSP 2013)*. Santorini, Greece, July 2013, pp. 1–8. ISBN: 978-1-4673-5805-7. DOI: 10.1109/ICDSP.2013.6622764.

[48] Alexander Lindau and Fabian Brinkmann. "Perceptual Evaluation of Headphone Compensation in Binaural Synthesis Based on Non-Individual Recordings". In: *J. Audio Eng. Soc* 60.1/2 (2012), pp. 54–62. URL: http://www.aes.org/e-lib/browse.cfm?elib=16166.

[49] Braxton Boren, Michele Geronazzo, Fabian Brinkmann, and Edga Choueiri. "Coloration Metrics for Headphone Equalization". In: *International Conference on Auditory Display*. 2015. URL: http://hdl.handle.net/1853/54097.

[50] Braxton B. Boren, Michele Geronazzo, Piotr Majdak, and Edgar Choueiri. "PHOnA: A Public Dataset of Measured Headphone Transfer Functions". In: *Audio Engineering Society Convention 137*. 2014. URL: http://www.aes.org/e-lib/browse.cfm?elib=17449.

[51] Hannes Gamper. "Head-related transfer function interpolation in azimuth, elevation, and distance". In: *The Journal of the Acoustical Society of America* 134.6 (2013), EL547–EL553. DOI: 10.1121/1.4828983. eprint: https://doi.org/10.1121/1.4828983. URL: https://doi.org/10.1121/1.4828983.

[52] H. Hacihabiboglu, E. De Sena, and Z. Cvetkovic. "Frequency-Domain Scattering Delay Networks for Simulating Room Acoustics in Virtual Environments". In: *2011 Seventh International Conference on Signal Image Technology Internet-Based Systems*. 2011, pp. 180–187. DOI: 10.1109/SITIS.2011.41.

[53] E. De Sena, H. Hachabiboğlu, Z. Cvetković, and J. O. Smith. "Efficient Synthesis of Room Acoustics via Scattering Delay Networks". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23.9 (2015), pp. 1478–1492. ISSN: 2329-9290. DOI: 10.1109/TASLP.2015.2438547.

[54] Patty Huang, Matti Karjalainen, and Julius O. Smith. "Digital Waveguide Networks for Room Response Modeling and Synthesis". In: *Audio Engineering Society Convention 118*. 2005. URL: http://www.aes.org/e-lib/browse.cfm?elib=13110.

[55] Jont B. Allen and David A. Berkley. "Image method for efficiently simulating small-room acoustics". In: *The Journal of the Acoustical Society of America* 65.4 (1979), pp. 943–950. DOI: 10.1121/1.382599. eprint: https://doi.org/10.1121/1.382599. URL: https://doi.org/10.1121/1.382599.

[56] Jean-Marc Jot and Antoine Chaigne. "Digital Delay Networks for Designing Artificial Reverberators". In: *Audio Engineering Society Convention 90*. 1991. URL: http://www.aes.org/e-lib/browse.cfm?elib=5663.

[57] F. Stevens, D. T. Murphy, L. Savioja, and V. Välimäki. "Modeling Sparsely Reflecting Outdoor Acoustic Scenes Using the Waveguide Web". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.8 (2017), pp. 1566–1578. ISSN: 2329-9290. DOI: 10.1109/TASLP.2017.2699424.

[58] M. A. Gerzon. "Periphony: Width-height sound reproduction". In: *Journal of the Audio Engineering Society* 21.1 (1973), pp. 2–10.

[59] M. A. Gerzon. "Ambisonics in multichannel broadcasting and video". In: *Journal of the Audio Engineering Society* 33.11 (1985), pp. 859–871.

[60] Boaz Rafaely. *Fundamentals of Spherical Array Processing*. Springer-Verlag Berlin Heidelberg, 2015.

[61] Jakob Vennerød. "Binaural Reproduction of Higher Order Ambisonics A Real-Time Implementation and Perceptual Improvements". MA thesis. Norway: NTNU - Trondheim, Norwegian University of Science and Technology, 2014.

[62] Archontis Politis. *Microphone array processing for parametric spatial audio techniques*. en. G5 Artikkeliväitöskirja. 2016. URL: http://urn.fi/URN:ISBN:978-952-60-7037-7.

[63] Stéphanie Bertet, Jérôme Daniel, Etienne Parizet, and O. Warusfel. "Investigation on Localisation Accuracy for First and Higher Order Ambisonics Reproduced Sound Sources". In: *Acta Acustica united with Acustica* 99 (2013), pp. 642–657. URL: https://hal.archives-ouvertes.fr/hal-00848764.

[64] Stéphanie Bertet, Jérôme Daniel, and Sébastien Moreau. "3D Sound Field Recording with Higher Order Ambisonics - Objective Measurements and Validation of Spherical Microphone". In: *Audio Engineering Society Convention 120*. 2006. URL: http://www.aes.org/e-lib/browse.cfm?elib=13661.

[65] S. Bertet S. Moreau J. Daniel. "3D Sound Field Recording with Higher Order Ambisonics - Objective Measurements and Validation of a 4th Order Spherical Microphone". In: *120th AES Convention*. 2006.

[66] Leo McCormack, Symeon Delikaris-Manias, Angelo Farina, Daniel Pinardi, and Ville Pulkki. "Real-Time Conversion of Sensor Array Signals into Spherical Harmonic Signals with Applications to Spatially Localized Sub-Band Sound-Field Analysis". In: *Audio Engineering Society Convention 144*. 2018. URL: http://www.aes.org/e-lib/browse.cfm?elib=19456.

[67] Cheol Ho Choi, Joseph Ivanic, Mark S. Gordon, and Klaus Ruedenberg. "Rapid and stable determination of rotation matrices between spherical harmonics by direct recursion". In: *The Journal of Chemical Physics* 111.19 (1999), pp. 8825–8831. DOI: 10.1063/1.480229. eprint: https://doi.org/10.1063/1.480229. URL: https://doi.org/10.1063/1.480229.

[68] R. H. Hardin and N. J. A. Sloane. "McLaren's improved snub cube and other new spherical designs in three dimensions". In: *Discrete & Computational Geometry* 15.4 (1996), pp. 429–441. ISSN: 1432-0444. DOI: 10.1007/BF02711518. URL: https://doi.org/10.1007/BF02711518.

[69] Franz Zotter, Matthias Frank, and Alois Sontacchi. *The Virtual T-Design Ambisonics-Rig Using VBAP*. 2010.

[70] Franz Zotter, Matthias Frank, and Hannes Pomberger. *Comparison of energy-preserving and all-round Ambisonic decoders*. 2013.

[71] Alois Sontacchi, Markus Noisternig, Piotr Majdak, and Robert Holdrich. "Subjective Validation of Perception Properties in Binaural Sound Reproduction Systems". In: *Audio Engineering Society Conference: 21st International Conference: Architectural Acoustics and Sound Reinforcement*. 2002. URL: http://www.aes.org/e-lib/browse.cfm?elib=11174.

[72] Michael A. Gerzon. "Recording Concert Hall Acoustics for Posterity". In: *J. Audio Eng. Soc* 23.7 (1975), pp. 569, 571. URL: http://www.aes.org/e-lib/browse.cfm?elib=2669.

[73] Angelo Farina and Regev Ayalon. "Recording Concert Hall Acoustics for Posterity". In: *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*. 2003. URL: http://www.aes.org/e-lib/browse.cfm?elib=12277.

[74] Juha Merimaa and Ville Pulkki. "Spatial Impulse Response Rendering I: Analysis and Synthesis". In: *J. Audio Eng. Soc* 53.12 (2005), pp. 1115–1127. URL: http://www.aes.org/e-lib/browse.cfm?elib=13401.

[75] Ville Pulkki. "Spatial Sound Reproduction with Directional Audio Coding". In: *J. Audio Eng. Soc* 55.6 (2007), pp. 503–516. URL: http://www.aes.org/e-lib/browse.cfm?elib=14170.

[76] Ramona Bomhardt, Marco Berzborn, Johannes Klein, Jan-Gerrit Richter, and Michael Vorlaender. "The ITA-Toolbox: An Open Source MATLAB Toolbox for Acoustic Measurements and Signal Processing". In: *DAGA 2017, Vol. 43*. 2017. URL: http://www.ita-toolbox.org/publications/ITA-Toolbox_paper2017.pdf.

[77] Angelo Farina. "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique". In: *Audio Engineering Society Convention 108*. 2000. URL: http://www.aes.org/e-lib/browse.cfm?elib=10211.

[78] Alex Baldwin, Stefania Serafin, and Cumhur Erkut. "ScatAR: A Mobile Augmented Reality Application That Uses Scattering Delay Networks for Room Acoustic Synthesis". In: *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*. VRST '17. Gothenburg, Sweden: ACM, 2017, 73:1–73:2. ISBN: 978-1-4503-5548-3. DOI: 10.1145/3139131.3141201. URL: http://doi.acm.org/10.1145/3139131.3141201.

[79] Jonas Holfelt. "ScatAR-WGW: Implementation and Evaluation of the Waveguide Web in an Application for Artificial Reverberation in a Virtual Environme". MA thesis. Denmark: Aalborg University, 2018.

[80] AForge.net. *Online*. URL: http://www.aforgenet.com/framework/.

[81] Julius O. Smith. *Spectral Audio Signal Processing*. online book, 2011 edition. http://ccrma.stanford.edu/~jos/sasp/, accessed <date>.

[82] Michael Vorlnder. *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. 1st. Springer Publishing Company, Incorporated, 2007. ISBN: 3540488294, 9783540488293.

[83] J. Huopaniemi, L. Savioja, and M. Karjalainen. "Modeling of reflections and air absorption in acoustical spaces a digital filter design approach". In: *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*. 1997, 4 pp.–. DOI: 10.1109/ASPAA.1997.625594.

[84] Michael Vorlander. *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. 1st Edition. Springer-Verlag Berlin Heidelberg, 2008.

[85] Jorge Garcia. *UnityOSC (Github)*. URL: https://github.com/jorgegarcia/UnityOSC.

[86] *Acoustics – Measurement of the reverberation time of rooms with reference to other acoustical parameters*. The document can be consulted by contacting: TIS-GS : Ana-Paula Bernardes. Geneva: ISO, 1997. URL: https://cds.cern.ch/record/442096.

[87] Piotr Majdak, Matthew J. Goupell, and Bernhard Laback. "3-D localization of virtual sound sources: Effects of visual environment, pointing method, and training". In: *Attention, Perception, & Psychophysics* 72.2 (2010), pp. 454–469. ISSN: 1943-393X. DOI: 10.3758/APP.72.2.454. URL: https://doi.org/10.3758/APP.72.2.454.

[88] Michele Geronazzo, Erik Sikström, Jari Kleimola, Federico Avanzini, Amalia De Götzen, and Stefania Serafin. "The impact of an accurate vertical localization with HRTFs on short explorations of immersive virtual reality scenarios". In: *Proc. 17th IEEE/ACM Int. Symposium on Mixed and Augmented Reality (ISMAR)*. Munich, Germany, Oct. 2018, pp. 90–97.

[89] Alexander Lindau, Vera Erbes, Steffen Lepa, Hans-Joachim Maempel, Fabian Brinkmann, and Stefan Weinzierl. "A spatial audio quality inventory (SAQI)". In: 100 (Oct. 2014).

[90] Laurent S. R. Simon, Nick Zacharov, and Brian F. G. Katz. "Perceptual attributes for the comparison of head-related transfer functions". In: *The Journal of the Acoustical Society of America* 140.5 (2016), pp. 3623–3632. DOI: 10.1121/1.4966115. eprint: https://doi.org/10.1121/1.4966115. URL: https://doi.org/10.1121/1.4966115.

[91] Tapio Lokki. *Physically-based auralization : design, implementation, and evaluation*. en. G5 Artikkeliväitöskirja. 2002-11-08. URL: http://urn.fi/urn:nbn:fi:tkk-001990.

[92] Lauri Savioja, Jyri Huopaniemi, Tapio Lokki, and Ritta Väänänen. "Creating Interactive Virtual Acoustic Environments". In: *J. Audio Eng. Soc* 47.9 (1999), pp. 675–705. URL: http://www.aes.org/e-lib/browse.cfm?elib=12095.

Appendix A

Bilinear Transform of the Head Shadow Effect Filter

A bilinear transform is applied to an analog transfer function in order to obtain a discrete-time transfer function of it. It is defined as follow:

$$H(z) = H(s)\big|_{s=\frac{2(z-1)}{T(z+1)}}.$$

Therefore:

Let $\alpha(\gamma) = \alpha$,

$$\begin{aligned} H(z,\gamma) &= \frac{\alpha\frac{2}{T}\frac{z-1}{z+1} + \beta}{\frac{2}{T}\frac{z-1}{z+1} + \beta} \\ &= \frac{\frac{2\alpha z - 2\alpha}{Tz+T} + \beta}{\frac{2z-2}{Tz+T} + \beta} \\ &= \frac{\frac{2\alpha z - 2\alpha + \beta(Tz+T)}{Tz+T}}{\frac{2z-2+\beta(Tz+T)}{Tz+T}} \\ &= \frac{2\alpha z - 2\alpha + \beta T z + \beta T}{2z - 2 + \beta T z + \beta T} \\ &= \frac{(2\alpha + \beta T)z + (\beta T - 2\alpha)}{(2 + \beta T)z + (\beta T - 2)}\frac{z^{-1}}{z^{-1}} \\ &= \frac{(2\alpha + \beta T) + (\beta T - 2\alpha)z^{-1}}{(2 + \beta T) + (\beta T - 2)z^{-1}}. \end{aligned} \tag{A.1}$$

T is the sampling period in seconds.