UNIVERSITAT POLITÈCNICA DE VALÈNCIA

_ TELECOM ESCUELA TÉCNICA VLC SUPERIOR DE INGENIERÍA DE TELECOMUNICACIÓN

INTERPOLATION TECHNIQUES IN ADVANCED SPATIAL AUDIO SYSTEMS

Victor García Gómez

Tutor: José Javier López Monfort

Trabajo Fin de Máster presentado en la Escuela Técnica Superior de Ingenieros de Telecomunicación de la Universitat Politècnica de València, para la obtención del Título de Máster en Ingeniería Telecomunicación

Curso 2017-18

Valencia, 24 de julio de 2018

Escuela Técnica Superior de Ingeniería de Telecomunicación
Universitat Politècnica de València
Edificio 4D. Camino de Vera, s/n, 46022 Valencia
Tel. +34 96 387 71 90, ext. 77190
www.etsit.upv.es

VLC/ CAMPUS VALENCIA, INTERNATIONAL CAMPUS OF EXCELLENCE

EMAS Gestión ambiental verificada REG.NO. ES-CV- 000030

Acknowledgements

Special appreciation to my tutor, José Javier López Monfort, who gave me the opportunity to carry out this work, guided me throughout it and made me discover a thrilling field as audio spatialization is.

Also thanks to Pablo, for his invaluable help, sharing his knowledge and for bearing all my questions always I need it.

Abstract

Increasing popularity of virtual reality systems and applications, such as videoconferencing and gaming, is demanding vast efforts in real-time audio virtualization, contributing powerfully to the user's sense of immersion. Sound positioning is achieved by means of convolving audio sources with Head Related Impulse Responses collections, whose measurement can be laborious and costly. Interpolation gives the opportunity to save time and efforts by allowing us to obtain faithfully new samples from few ones.The aim of this thesis is to design and implement interpolation algorithms for HRIRs, Binaural Room Impulse Responses and a more advanced audio format, Ambisonics. The developed algorithms will have a strong practical orientation onto Virtual reality in sound spatialization. Thus, algorithms are conceived with the focus on real-time implementations with low computational cost, promoting their deployment and exploitation in the professional market.

Resumen

La creciente popularidad de los sistemas y aplicaciones de realidad virtual, así como las videoconferencias y los videojuegos, exigen grandes esfuerzos en la virtualización de audio en tiempo real, contribuyendo poderosamente al sentido de inmersión del usuario. El posicionamiento del sonido se logra mediante la convolución de las fuentes de audio con las colecciones de Head Related Impulse Responses, cuya medición puede ser tediosa y costosa. La interpolación brinda la oportunidad de ahorrar tiempo y esfuerzo al permitirnos obtener muestras nuevas fielmente a partir de unas pocas. El objetivo de esta tesis es diseñar e implementar algoritmos de interpolación para HRIR, Binaural Room Impulse Responses y un formato de audio más avanzado, Ambisonics. Los algoritmos desarrollados tendrán una fuerte orientación práctica en la espacialización del sonido en la realidad virtual. Por lo tanto, el desarrollo de los algoritmos estará enfocado en implementaciones en tiempo real con bajo coste computacional, promoviendo su implementación y explotación en el mercado profesional.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context and Motivation

Increasing popularity of virtual reality systems and applications, such as videoconferencing and gaming, is demanding vast efforts in real-time audio virtualization so as to contribute powerfully to the user's sense of immersion. Therefore, to create a full convincing 3D experience, highly realistic audio spatialization is required to accompany 3D image.

Regarding audio spatialization, binaural reproduction over headphones is of high interest, since penetration of mobile devices has fomented the use of headphones, today ubiquitous in all multimedia systems.

Sound positioning upon binaural technologies is achieved by means of convolving audio sources with Head Related Impulse Responses (HRIRs), which represent interaural time and level difference, of signals arriving to both ears, as well as directional dependent filtering of the head and pinnae [1]. HRIRs, or their frequency-domain equivalent Head Related Transfer Function (HRTFs), can be measured at the ear canals at different distance and directions from the head, usually in anechoic chambers [2]. All these theoretical concepts regarding audio spatialization, amongst others, will be explained in further detail in Chapter 2. Hence the author encourages the reader to refer to this chapter if needed.

The HRTF models the localization in free field, however sources are normally located in real environments. The acoustic environment contributes to the sense of immersion, and also improves the sources localization. The simulation of the acoustic environment is achieved by means of convolving audio sources with finite Room Impulse Responses (RIRs), which consist of the direct path, early reflections and reverberation taking place in a given room.

Binaural Room Impulse Responses (BRIRs), also known as Combined Head and Room

Impulse Response (CHRIR) in the literature, are thought of as consisting of these two parts, HRTFs and RIRs.

Benefits from using BRIRs are clear for applications such as audioconferencing, as it allows to situate the participants of the conference at any place of the room, thus contributing significantly to the improvement of the intelligibility and sense of immersion [3]. Is such the interest in this technology that even large companies as Youtube/Google and Facebook (with Facebook Audio 360) have invested to start enabling reproduction of videos using binaural technologies to simulate 3D environments.

However, there is still a lot of work to do in the binaural field in order to achieve natural sounding scenarios. For a 3D sound application, HRIRs between source and receiver must be known at each moment in order to convolve them with the audio source in real time and produce the sensation of localization. There are computational methods for acoustic room modeling, which obtain the RIRs from the geometry and materials of the room. Despite that, these methods are not as accurate as the in-situ measurements and sound artificial. Typically, to measure in-situ a RIR or BRIRS at specific points of a room, Sweep or MLS measurement techniques can be used to record the impulses with two microphones capsules placed in a dummy head or in a real person ear canals.

Nonetheless, for real-time applications, the problem consists of that, if the source or the listener were to be moving, the BRIR needs to change with the movement of the source or the listener too, so as to perceive a smooth realistic movement. For that, infinite combinations of source-listener positions should be measured in a room. Due to the high effort that this would entail and in order to optimize storage and reproduction, in recent years, interpolation of BRIRs arose to aid to overcome this problem. With interpolation, it is intended to reduce the number of measurements of BRIRs in a single room, while maintaining perceptually correct spatialization.

There are two main ways of approaching the measurement of the BRIRs for later interpolation: fix one source position and move the listener, or reversely, fix the listener and move the source. The first approach allows one to simulate the movement of the listener for several fixed source positions meanwhile the second allows one to simulate the movement of the sources for a fixed listener position. Ideally, combining both approaches, very complex convincing sound environments could be reproduced by tracking the movement of the sources and the listener and combining the two filters resulting from both methods. However, it implies a massive amount of real time operations, measures and complexity, including tracking. Hence, in this thesis, the research of interpolation has been limited to the second approach, which is illustrated in Figure 1.1.

Further on, there are also other advanced audio recording systems that can be synthetized

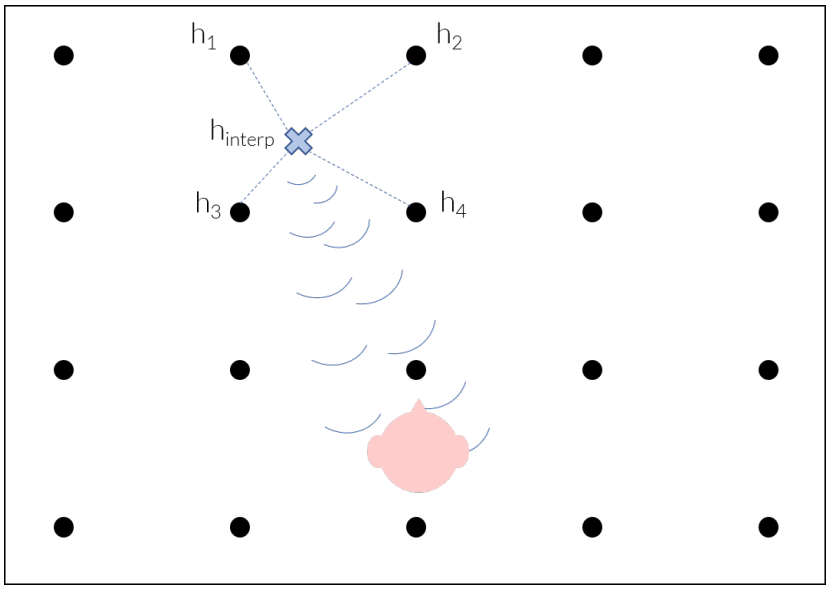

Figure 1.1: Interpolating BRIR scenario from measured BRIRs (black dots) in a room with a fixed listener and moving sources.

into BRIRs and that provide more flexibility, whose interpolation could be of high interest as well. Ambisonics is a popular 3D spatial audio encoding/recording and decoding/reproduction system that is based on the mathematical description and representation of the sound field as a decomposition into spherical harmonics. To put it simpler, one can thought of spherical harmonics decomposition as virtual microphones with directional polar patterns pointing to a specific directions, being this direction vector expressed in a spherical coordinate system.

Ambisonics was early introduced in 1970 but it has not been until now that it has become popular thanks to the advances in digital signal processing. Currently, Ambisonics is becoming the standard for 3D audio productions and Virtual Reality in the industry and almost every plugin or 3D audio software supports one or more of its formats. For example, Google chose Ambisonics Ambix format [4] for enconding 360º sound scenes in Youtube. Therefore, addressing the interpolation of this format is an added value that would be appreciated in the professional market. Actually, so far, it does not exist any known technique to interpolate Ambisonics.

The potential and interest of Ambisonics rely on the fact that the encoded audio format is independent on the reproduction set up. That means, provided that the decoding stage is designed properly, one can choose the reproduction format that best suits ones' needs [5], with the added benefit of being capable to apply extra processing in between (e.g. rotate the sound field with a tracking of the orientation of the head). To sum up, with Ambisonics, is as possible to reproduce the recorded sound field with several loudspeaker layout arrays, as to reproduce it with headphones, turning it to BRIRs.

In this thesis, interpolation of Ambisonics, specifically, in a format called B-Format,

measured in the same fashion as commented before, is also addressed. A deep review about Ambisonics will be presented in Chapter 2. Several methods of interpolation have been proposed, either for HRTFs and RIRs, but quite less for BRIR and none for Ambisonics. To point out some, in [6], Haneda et al. proposed a common acoustical pole and residue model, which can describe room transfer function variations using simple residue functions. Therefore, interpolation of RIRs can boil down to interpolating residue functions. This seems to be effective for the low frequency component of the room transfer function.

Other interpolation method proposed by Kearney et al. [7] involves a recent algorithm, known as Dynamic Time Warping (DTW), for automatic temporal alignment of direct and sparse early reflection components of the RIRs. They propose to split the measured impulse responses into two regions, namely the early reflection and diffuse decay regions. Then, interpolate linearly in time only in the first region after alignment, whereas the tail is synthesized according to the method presented in [8]. This method, avoids smearing distortions that occur in linear interpolation processes but seems to be not suited for interpolation in small rooms, since early reflections are not as sparse.

Frequency domain techniques have been also studied in [9] and it is claimed that they result in better performance than time domain interpolation.

Nevertheless, interpolation entails a drawback for real-time applications: its computational cost. Usually there will be a tradeoff between computational cost and perceptual faithfully to the equivalent measured BRIR.

Hence, the motivation of this thesis is to review in detail some of these techniques and propose new ones for both Binaural Room Impulse Responses and Ambisonics B-Format signals, which could work in real time to best suit the needs of the industry.

As a result of this work, the publication called *Binaural Room Impulse Responses Interpolation for Multimedia Real-Time Applications* [10] has been published. In this paper, it is proposed a novel algorithm for interpolation and temporal alignment of the BRIR, aiming to provide an efficient alternative to other interpolation methods while keeping a realistic perception of immersion, which, in turn contributes to a significant measurement reduction. This algorithm is described in Chapter 3. Besides, there is an incoming journal article about Ambisonics B-Format interpolation which will be published in the next weeks. In this article, a completely new algorithm to interpolate Ambisonics signals in real time is presented and evaluated, showing the results of this work.

1.2 Objectives

The concrete objectives of this thesis are:

- To analyze the state-of-the-art techniques for spatial audio interpolation.
- To understand the underlying principles of the localization of sound sources for different recording and reproduction methods.
- To learn new techniques which could be useful in the professional industry, regarding Binaural Room Impulse Responses and Ambisonics, and that could be easily deployed in a real system.
- To implement new interpolation algorithms, with Matlab, suitable for audio spatialization in real-time applications, evaluate and compare them with the existing techniques.

1.3 Structure of the text

The following text is structured as follows:

In chapter 2 theoretical key concepts that will appear continually throughout this memory will be introduced, regarding audio spatialization and interpolation.

In chapter 3, the two developed algorithms will be described, whose implementation is explained step by step.

In chapter 4, the results obtained from evaluating objectively and subjectively both algorithms in terms of faithfulness of the interpolated samples, perception of location and timber, will be shown.

Finally, in chapter 5, the main conclusions of the thesis will be stated and future possibilities of improvement will be proposed.

Chapter 2

Theoretical Basis

The aim of this chapter is merely to introduce the reader to the basic principles of audio spatialization, two of the most common audio formats to encode spatial information and interpolation, in order to provide the global vision needed to understand the following chapters.

2.1 Introduction to audio spatialization

The purpose of a spatial sound system is to recreate as accurate as possible the acoustic sensations that a listener would perceive inside an environment with a series of acoustic characteristics. Spatial sound systems have evolved in the latest years due to their great interest in fields like music production, virtual reality, immersive experiences, cinema, etc... From stereophony, considered as the simplest approximation to spatial sound, through surround sound systems with multiple reproduction channels (5.1, 7.1, 22.2 and more), to more advances systems as Wave Field Synthesis (WFS), all these systems work under the principles that govern the perception of the localization for the human hear. In the next point, the mechanisms used by the human hear to locate sound sources are described.

2.1.1 Spatial Sound Systems

Together with the vision, human hearing performs a major role in the way our environment is perceived. Sound sources, as well as real objects, are perceived in 3D and can be attributed with a width, height and depth. Sound scenarios are made of a complex combination of multiple sound sources, which can be either perceived individually or grouped according to some of the hearing mechanisms related to the spatial localization cues.

Recording and reproduction techniques capable of storing and transmit this spatial information have been a matter of research for many years. In these techniques two or more microphones and loudspeakers are used to capture different spatial cues. For example, in surround sound systems, it is a common approach to use the upper channels to reproduce the non-direct sound field (reverb) to enhance the sensation of immersion by trying to reproduce the reflections of the ceiling where the sound was recorded.

However, spatial sound reproduction systems based in loudspeakers layouts have a limitation: they can only reproduce accurately the acoustic field in an optimal listening area know as sweet spot. This area is usually limited to the central point of the set-up. Moving outwards that zone, breaks the balance of the mix and degrades the spatial sensation.

Therefore, another popular strategy to reproduce the spatial sensations it to do so directly in the ears of the listener, via headphones. This is what is known as binaural reproduction. The signals to be reproduced with headphones are usually recorded with an acoustic dummy head or, on the other hand, they can be synthesized using a measured Head-Related Transfer Function (HRTF). Note that those signals are not the same that are reproduced by the loudspeakers in surround systems as they also include the filtering effect of the head, torso and pinnae. Reproducing this signals with loudspeakers is likely to introduce crosstalk between left and right channel, although this non desired effect can be eliminated by prefiltering the signals.

Other recent strategies to achieve a faithful reproduction of the sound field are the Vector Base Amplitude panning (VBAP), which is an extension of the stereophony principle to complex loudspeaker setups, Ambisonics and Wave Field Synthesis. Later on, Ambisonics will be discussed in more detail.

2.1.2 Spatial Hearing

The human auditory system is very sophisticated and uses multiple cues to analyze and extract most spatial information pertaining to a sound source. In this section, the main localization cues are presented.

Interaural Differences

One of the basic binaural processing mechanisms involves comparing the information that arrives to each of the ears. On the one hand, the first difference is the time of arrival (Figure . This is commonly referred as Interaural Time Difference (ITD). The ITD provides

good information about the direction of a sound source, but introduces ambiguity with respect to the front and back hemispheres. On the other hand, the second difference is the level of the signals. This is commonly referred as Interaural Level Difference (ILD). The presence of the head provokes an attenuation at high frequencies as they are shadowed by the head. Hence, sound is perceived from the side with the highest amplitude.

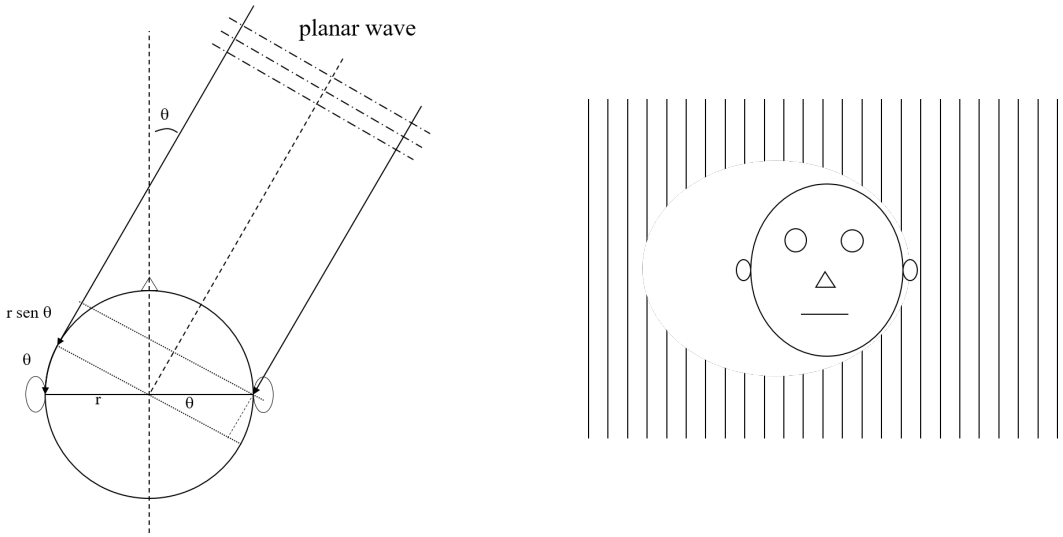

Figure 2.1: Interaural cues representation. On the right,a planar wave arrives to the ears with a path difference that results in the ITD. On the left, diffraction of the head causes the ILD specially for high frequencies.

The ITD and ILD are primary cues for the perceived azimuth angle of a sound source , being the ILD the predominant mechanism at high frequencies and the ITD the predominant mechanism at low frequencies (Duplex Theory). Therefore, using these cues, the estimated azimuth angle can be constrained to a particular cone of confusion, where ILD and ITD are equals between each pair or points.

Spectral Cues

Assuming a symmetrical head, interaural differences should not apply for the median plane. Nevertheless, people are still able to localize the sound due to what is known as monoaural cues. These cues are related to the spectral changes introduced by the pinnae, at high frequencies, and the torso and head, for low frequencies, and also help to discriminate the front from the back for sounds with sufficient high-frequency energy. This may occur because of the front/back asymmetry of the pinnae.

Distance and Dynamic Cues

The process of localizing a source is dynamic and is usually aided by other sensory inputs. Dynamic cues are very useful to overcome the ambiguities introduced by the static cues. For example, little tumble of the head, allows people to save the front/back asymmetry easier as well as enchance the perception of the angle of elevation.

Finally, at large distances, the above cues lack of information to estimate the distance of the source. Hence, range estimation is rather based on loudness. Loudness of a sound source changes with the distance due to the attenuation caused by propagation. However, the effectiveness of this cue usually relies on the familiarity of the listener with the source.

Another cue for distance perception is directly related to the environment: the ratio between the direct sound and the reflections. This provides a relative estimation of the distance between both listener and the source, regardless of the familiarity with that sound.

It is important to notice that to produce a realistic perception of the direction of the sound, all cues need to be consistent and usually need to be aligned with other non-auditory cues, such the visual ones.

2.1.3 Head Related Transfer Functions

As commented previously, for binaural reproduction, HRTFs can be used to synthesize the signals at the entrance of the ears. HRTFs, capture the effects of the different structures (head, torso and pinnae) that change the sound before it reaches the ear drums. Hence it is a function of direction, distance and frequency as follows:

$$HRTF(\theta, \phi, f) \tag{2.1}$$

being θ the azimuth, ϕ the elevation and f the frequency.

By definition, they are the transfer function between the sound pressure that is present at the center of the listeners head, when the listener is absent, and the sound pressure developed at the listeners ear. The HRTF expressed in time domain is called Head Related Impulse Response (HRIR).

Spatial localization cues such as ITD and ILD are encoded as the time of arrival for the first impulse on the HRIRs and the level difference of the HRTF magnitude response, respectively.

Figure 2.2: Dummy head mannequin to measure HRTFs.

Recall that HRTFs can be measured by using a dummy head as the one shown in Figure 2.2 or by placing two microphone capsules at the ear canals of a person. Then, usually a logarithmic sweep is employed to excite all frequencies in the room an obtain a high quality impulse response by a process of decorrelation.

It should be noted that each human has their own unique HRTFs that share many similarities to the ones measured with a dummy head or other person. Indeed, these subtle differences play a major role for precise localization. Several experiments [11] evidenced the fact that, when using ones own HRTFs, the perception is much more realistic.

2.1.4 Binaural Room Impulse Responses

Structure of the Binaural Room Impulse Responses

Likewise HRTFs, for binaural audio spatialization including room effect, it is necessary to convolve a mono audio source with the left and right ear impulse responses of the BRIR. As commented in [12], BRIR consist of two parts:

- HRTF: is a function of the listener, its relative location with the sound source and its orientation.
- RIR: is a function of the room characteristics and listener and source location and orientation.

Combined together, they provide a measure which is function of the actual (not relative) location and orientation and the audio source and listener. In Figure 2.6 the left and right channels of a single BRIR are illustrated.

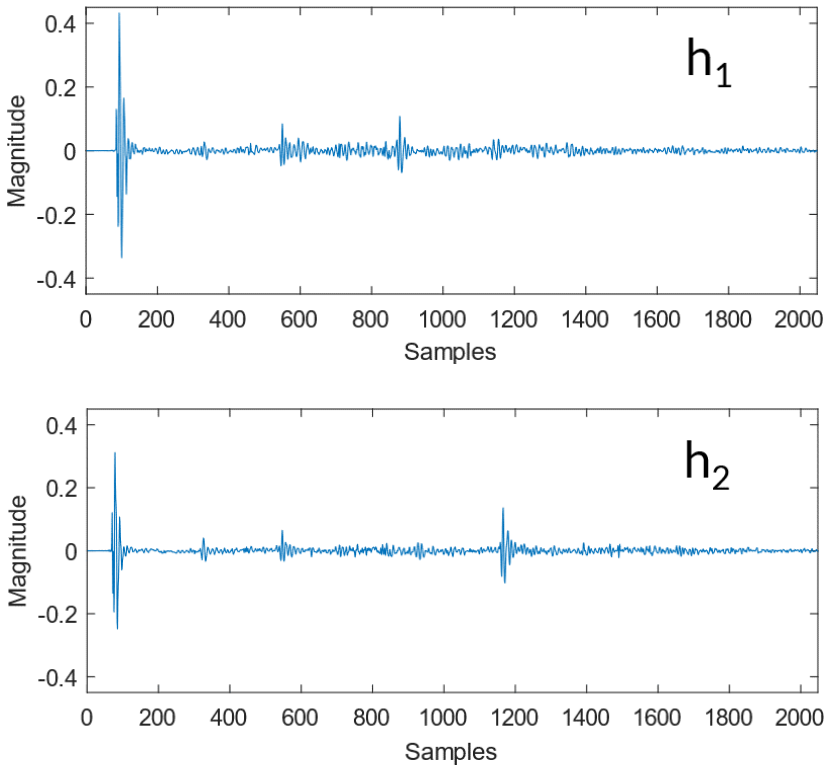

Figure 2.3: One channel of measured BRIRs h_1 (up) and h_2 (down)

As seen in the figure, BRIRs consist of a series of impulses, being the first one the direct sound, and, the rest, the impulses reflected on the walls, floor and ceiling. The first reflections, also known as early reflections or echoes, are sparse and can be distinguished

clearly, meanwhile the late reflections are grouped, forming what is known as the reverberant tail. It should be noted that for the human hear, the main impulse and the early reflections are the ones containing the most useful information for localization of sound sources in a room. Despite the reverberant tail also contributes to localization, it is more invariant throughout a room and rather contributes more to the timbre of the perceived sound.

Indeed, this fact can be exploited for interpolation. In [7], it was already proposed to split BRIRs, distinguishing between direct and early reflection region and reverberant tail by setting a transition point, which depends on the room volume and density of reflections. Therefore, the interpolation problem is split in two: first part, is the one containing the direct sound and the early reflections and which requires more dedicated and complex algorithms; second, is the reverberation tail that can be easily managed by simpler techniques.

2.1.5 Ambisonics

Ambisonics is a full-sphere surround sound technique early introduced in 1970 that has become popular quite recently due to the advances in digital signal processing. Unlike other surround sound formats, it does not transmit the signals to be reproduced on the loudspeaker setup. Instead it contains an speaker independent uninterrupted 360º representation of the sound field that can be decoded to any loudspeaker layout, including headphones, providing great flexibility in reproduction.

Ambisonics signals are usually given in the so-called B-Format, which is based on a truncated spherical harmonic decomposition of the sound field. For first order Ambisonics (FOA), the sound field is encoded in four signals. They correspond to the sound pressure W, and the three components of the pressure gradient X, Y and Z. Together, these approximate the sound field on a sphere around the microphone.

These spherical harmonics can be thought as of different virtual microphones with different polar patterns pointing to a specific direction, with the four being coincident. W is an omni-directional polar pattern, containing all sounds in the sphere, coming from all directions at equal gain and phase. X, Y and Z are three figure-8 bi-directional polar patterns pointing towards the three Cartesian axis.

As shown in Figure 2.4, the number and shape of the spherical harmonics varies according to the Ambisonics order, or number of required microphone capsules to capture the sound field. Ambisonics of higher order provide better audio localization, with the inconvenience that requires costly and complex microphone arrays.

Indeed, in practice, it is not possible to have coincident microphones capsules with those

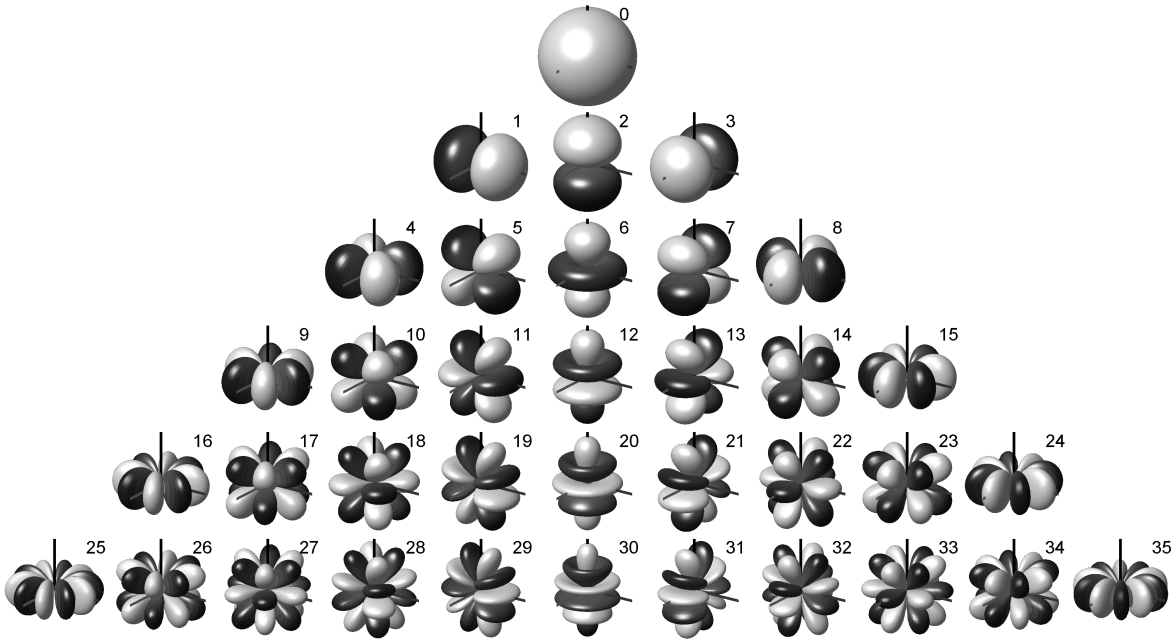

Figure 2.4: Spherical Harmonics directivity patterns representations from order 0 (top) to 5 (bottom). Each order is composed by each row plus all the patterns above.

patterns, so these signals are derived by applying a processing to the raw signals captured by the microphones, which does not need to have those patterns (in fact, cardioid are employed), usually referred as A-Format signals. Typical arrangement of microphones is a tetrahedron, as it minimizes the spatial error between capsules, which later derives in spatial aliasing, limiting the highest frequency that can be faithfully reconstructed. In format A, the signals preserve phase information which stores the localization information. In format B, phase information is lost as phase differences are corrected by making the capsules virtually coincident. Then localization information is encoded as a the energy differences between the resulting signals. Different decoding techniques exists for both formats. This methods are explained in further detail in [3], [13], [14] and [15].

Mathematically, for a sound signal S being encoded in B-Format FOA, WXYZ signals are described as follows:

$$W = \frac{S}{\sqrt{2}} \tag{2.2}$$

$$X = S \cdot cos(\theta)cos(\phi) \tag{2.3}$$

$$Y = S \cdot sin(\theta)cos(\phi) \tag{2.4}$$

$$Z = S \cdot sin(\phi) \tag{2.5}$$

being θ the azimuth angle and ϕ the elevation. Note that W is corrected by a 3dB factor to

obtain the same average energy than X, Y and Z.

There is evidence enough that proves that FOA provides poor spatial resolution, leading to a blurry perception of the direction of the sound sources, also due to the small area of the sweet spot, which is smaller than the human head for frequencies above 600 kHz [14]. This fact, steered the industry onto the search for higher order representations of the sound field. The spatial resolution can be increased and the sweet spot enlarged by adding groups of more selective directional spherical harmonic components to the B-format, that no longer corresponded to conventional polar patterns. They are what is called high order Ambisonics (HOA). In practice, HOA require more channels for reproduction and recording, $(n+1)^2$, being n the Ambisonics order, and the frequency up to which the sound field can be perfectly reproduced in the sweet spot is given by

$$f_{max} = \frac{nc}{2\pi r_{min}} \tag{2.6}$$

being c the speed of the sound, and r_{min} the minimum allowed radius of the sweet spot, which is approximately the size of the head (~18 cm).

Finally, decoding of the Ambisonics is carried out by a linear combination of the B-Format signals. However, in practice, a real Ambisonic decoder requires a number of psycho-acoustic optimizations to work properly, as highlighted in [14], following the Gerzon definition of Ambisonics (1992). Typical structure of an optimized Ambisonic decoder involves using different decoding matrices for each frequency range, followed by a near-field compensation that corrects the reactive component of the sound field when the listening position is within a few meters of the loudspeakers. There are several methods to calculate the decoding matrices for a specific loudspeaker layout, according to physical or psychoacoustical criteria. According to [16], some straightforward designs for the decoding matrix are the following:

$$SAMPLING: \quad D = \frac{1}{L} Y_L^T \tag{2.7}$$

$$ModeMatching: \quad D = (Y_L^T Y_L + \beta^2 I)^{-1} Y_L^T \tag{2.8}$$

$$ALLRAD: \quad D = \frac{1}{N_{td}} G_{td} Y_{td}^T \tag{2.9}$$

where L is the number of loudspeakers of the reproduction setup, Y_L is the $(n+1)^2 \times L$

matrix of spherical harmonics at loudspeakers directions. In the mode matching solution (MMD), β is a regularization value. In the ALLRAD method, matrix Y is the spherical harmonics matrix at the N_{td} directions in a uniform spherical t-design, where G_{td} are the vector-amplitude panning gains (VBAP) at the loudspeaker positions. By multiplying the B-Format signals with any of these matrices, it gives the signals to be reproduced by the loudspeakers.

Binauralization of these signals for reproduction over headphones, requires one extra final step. The signals of each loudspeaker are convolved with the HRTF for that loudspeaker direction and are added to form the left and right channel. This can be described as:

$$x_{bin}(t) = \begin{bmatrix} \sum_{i=1}^{(n+1)^2} LS_i(t) * HRIR_i^L(t) \\ \sum_{i=1}^{(n+1)^2} LS_i(t) * HRIR_i^R(t) \end{bmatrix} \tag{2.10}$$

where LS_i is the signal for the loudspeaker i and $HRIR_i^L$ and $HRIR_i^R$ are the left and right channel HRIR, respectively, for the direction of the loudspeaker i.

Finally, notice that here it may also be necessary to interpolate the HRTFs in those exact directions, highlighting the importance and application of interpolation.

2.2 Interpolation

In this section, the problem of interpolations is stated and several methods for BRIRs interpolation are introduced. As commented in the introduction, we cannot store BRIRs at infinite measurement positions. Instead, a reasonable number of measurement points can be selected on a room and, then, somehow, try to guess the missing measurements. Indeed, what interpolation does is to take as a reference surrounding BRIR measures to reconstruct the BRIR at the interpolation point.

There are several ways to interpolate but the most straightforward approach is what is known as linear interpolation. One dimension linear interpolation can be described by

$$h_{int} = h_1 + (h_2 - h_1)\frac{x_{int} - x_1}{x_2 - x_1} \tag{2.11}$$

where h_1 and h_2 represent the signal magnitude of BRIRs measured in positions x_1 and x_2, and h_{int} the interpolated BRIR at position x_{int} on X axis. Note that at the intermediate

position between x_1 and x_2, Equation 2.11 corresponds to the average formula. In the end, linear interpolation consist of carrying out a linear weighed average of the two signals.

Besides linear, polynomial interpolation could also be used, i.e. use equations of higher order. However, this would make sense if interpolation is carried out on a non-linear surface as the one of a sphere. For linear regular grids as the one used, linear interpolation is simpler and will perform well.

As was shown in Figure 1.1, measured grid conforms a 2D plane where interpolation can take place. Within each quadrant, a single linear interpolation cannot be sufficient for achieving 2D interpolation, since 4 measures contribute to the final result. To overcome this problem, bilinear interpolation is employed. As Figure 2.5 depicts, it consists of carrying out two linear interpolation in one axis, and use the resulting signals to interpolate in the remaining axis. Thus bilinear interpolation is achieved by means of three 1D linear interpolations. Similarly, moving to a 3D approach, it can be achieved by means of seven linear interpolations, but, in that case, measurements on Z axis have to be taken.

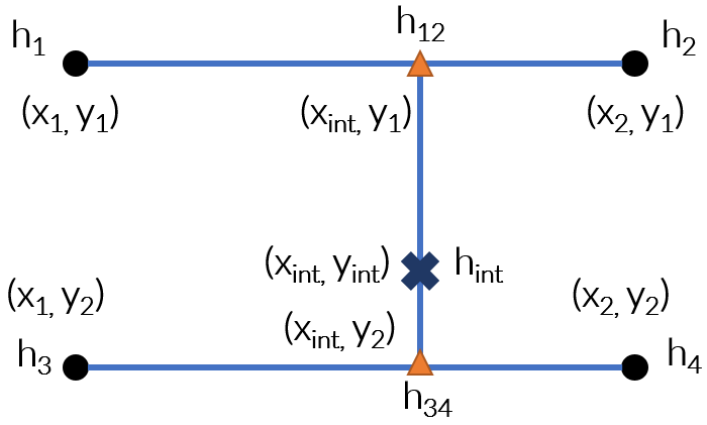

Figure 2.5: Bilinear interpolation

Of course, increasing the number of dimensions where to interpolate, increases the computation cost by a factor of $2^{nD} - 1$, being nD the number of dimensions. Besides, note that for BRIRs, right and left channels should be interpolated separately so the number of operations at the end are doubled.

However, for BRIRs, interpolation is not as simple as performing linear interpolation sample by sample between the two signals. In the following points, a review of the most common methods to interpolate BRIRs, with increasing complexity, will be analyzed, highlighting their strengths and drawbacks.

2.2.1 Linear interpolation in time domain

Starting with the most straightforward solution, lets suppose one has two measured BRIR in two different positions, h_1 and h_2, and only one channel is considered for sake of simplicity, as shown in Figure 2.6. Notice that neither the direct impulse nor the reflections are aligned in time. In addition, early reflections do not even have a clear correspondence between the two signals, i.e. each reflection in h_1 does not necessarily have a match in h_2 .

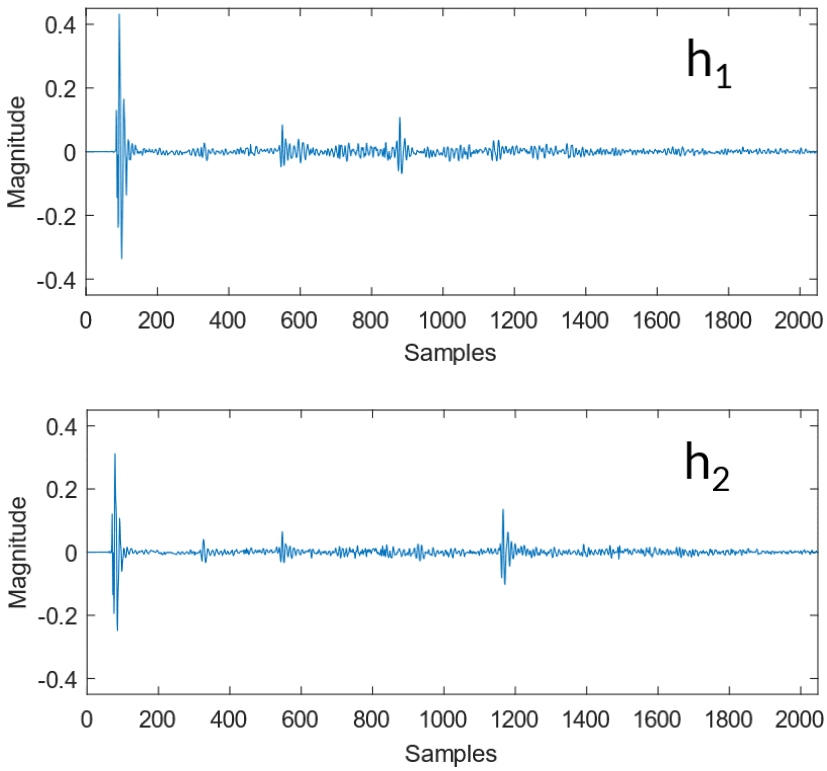

Figure 2.6: One channel of measured BRIRs h_1 (up) and h_2 (down). Only main impulse and early reflections are shown.

If sample-by-sample linear interpolation were applied to this signals, the resulting signal would present replicated and attenuated impulses of each of the impulses of h_1 and h_2 due to these misalignments, as illustrated in Figure 2.7. Therefore, proper alignment of the signals is required previously to interpolation so that, not only the magnitude but the position of the echoes is interpolated.

2.2.2 Linear interpolation in frequency domain

The second step, would be to consider linear interpolation in frequency domain. The position of the echoes is sort of determined by the phase, so it makes sense that if the modulus

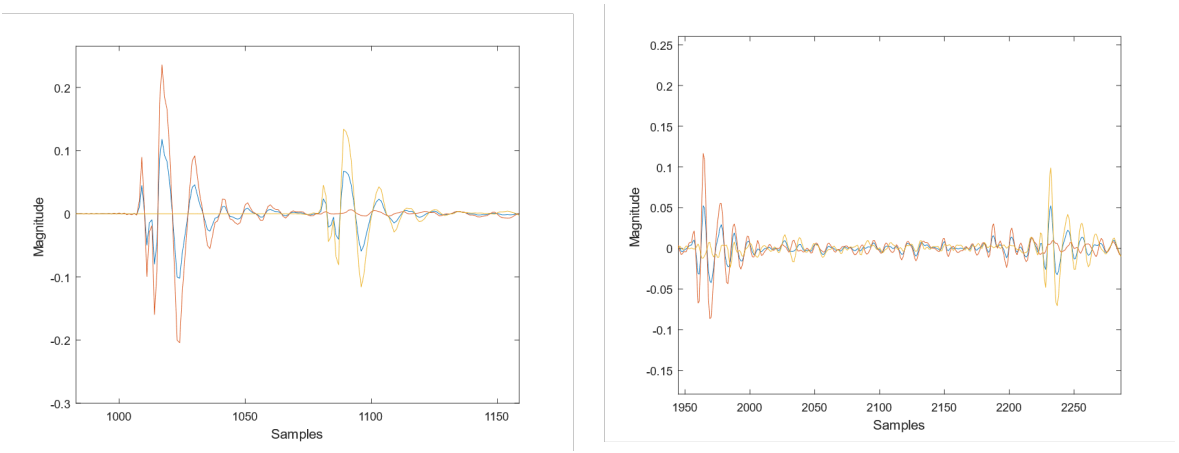

Figure 2.7: One channel linearly interpolated BRIR. In red, h_1, in yellow h_2 and in blue, h_{int}. On the left, the interpolated main impulse is shown. On the right, the interpolated early reflections.

of the spectrum is interpolated separately to the unwrapped phase (to avoid interpolating in the folds), this will achieve a similar effect as interpolating the signal's position as well as the magnitude.

Indeed, as shown in Figure 2.8, main impulse position is interpolated quite well. Nevertheless, the energy of the impulse is spread in the surrounding samples leading to a widening of the impulse. This well-known effect is called smearing. As can be guessed, this effect is not beneficial for localization. Also, early reflections have been displaced but still further away from the expected location and also present smearing.

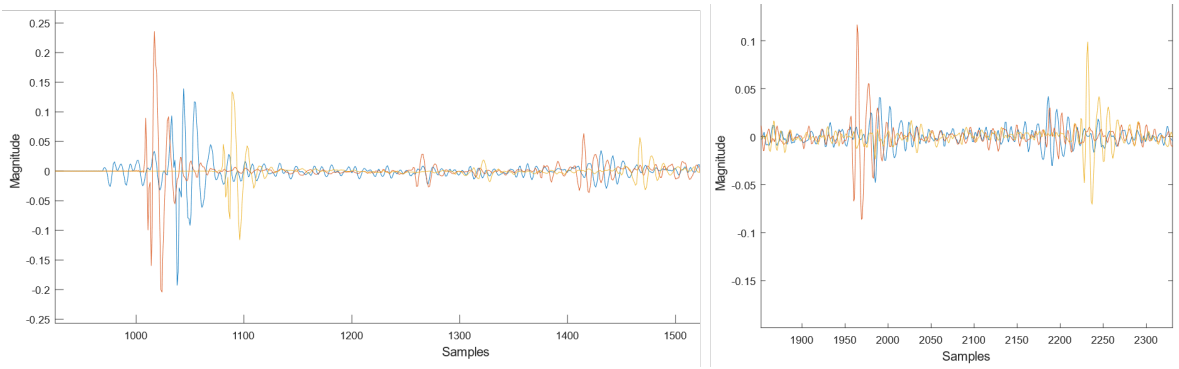

Figure 2.8: One channel linearly interpolated BRIR in the frequency domain. In red, h_1, in yellow h_2 and in blue, h_{int}. On the left, the interpolated main impulse is shown. On the right, the interpolated early reflections.

2.2.3 Interpolation with Dynamic Time Warping

Next approach uses one of the state-of-the-art algorithms to previously align the signals in time domain. When aligning two signals in time, the problem is that, even if one applies a delay to align the main direct impulse, this does not guarantee that subsequent reflections will match up (Figure 2.9), due to the different reflection paths lengths at different positions in the room. Kearney et al. proposed, in [7] and [1], to employ the well-known general Dynamic Time Warping (DTW) algorithm for successfully align BRIRs for interpolation.

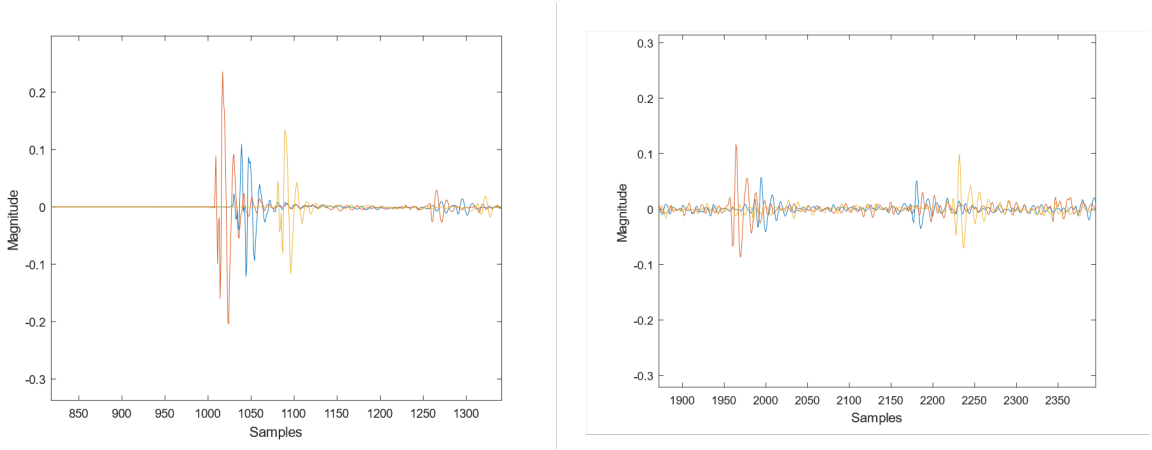

Figure 2.9: One channel linearly interpolated BRIR in the time domain aligning the main impulse. In red, h_1, in yellow h_2 and in blue, h_{int}. On the left, the interpolated main impulse is shown. On the right, the interpolated early reflections.

Dynamic Time Warping solves the interpolation problem by warping (stretching) non-linearly the signals by repeating samples in each time series, according to a distance criteria (Figure 2.10). In fact, DTW is based on the calculation of a minimum distance warp path through an accumulated distance matrix (Figure 2.10). The distance is the Euclidean distance between each of the samples of one signal to each of the samples of the other.

The optimal warp path is therefore given by

$$D(W) = \sum_{k=1}^{k=K} D(w_{ki}, w_{kj}) \tag{2.12}$$

where D(w) denotes the distance of the warp path and $D(w_{ki}, w_{kj})$ represents the distances between sample indexes at the k^{th} element of the warp path.

Next, applying the resulting warping vectors to h_1 and h_2, the warped versions of them, h_{w1} and h_{w2}, are obtained and are aligned. Once aligned, linear interpolation is applied to both signals and, also, to the warping vectors w_1 and w_2. Final step is to map the warped

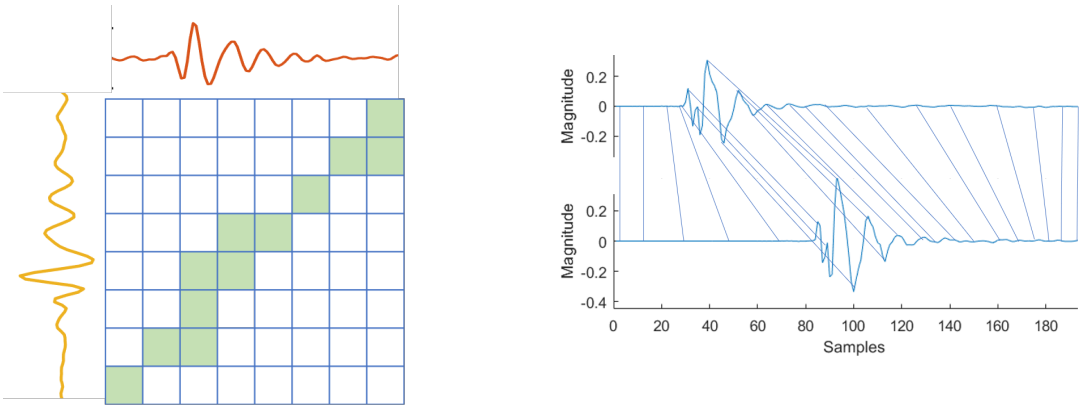

Figure 2.10: Dynamic Time Warping schema. On the left, the distance matrix with the minimum warp path in green. On the right, association between samples of h_1 and h_2 according to the minimum distance warp path

interpolate vector back in to the unwarped time domain using the interpolated warp vector. A more detailed description of the algorithm can be found in [7].

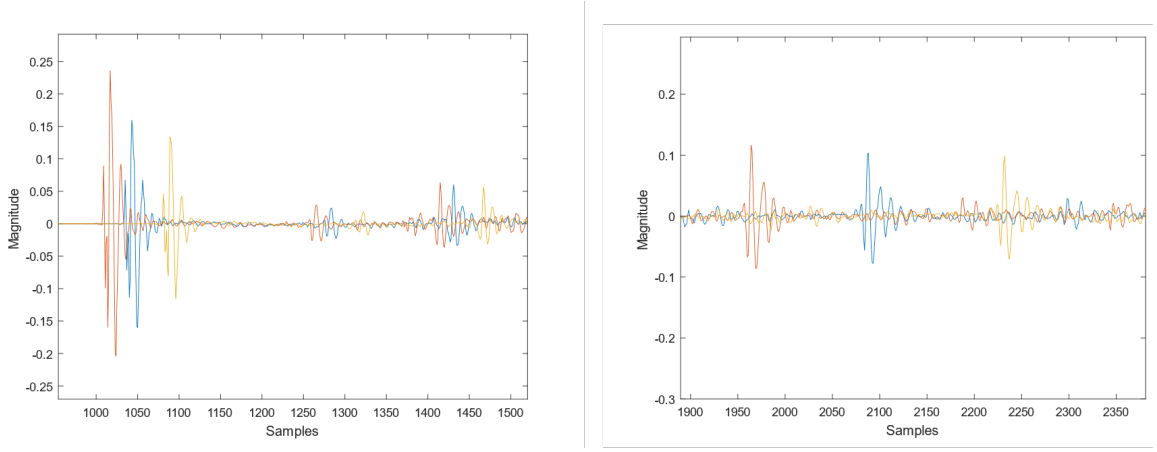

Figure 2.11: One channel linearly interpolated BRIR in the time domain using DTW. In red, h_1, in yellow h_2 and in blue, h_{int}. On the right, the interpolated main impulse is shown. On the left, the interpolated early reflections.

Figure 2.11 shows how the main impulse and the early reflections have been perfectly reconstructed in magnitude and location. Nonetheless, although DTW provides an effective interpolation solution, it should be noted that it does so at the expense of a highly computational cost. For example, for a typical early reflection part of a RIR of 3000 samples, it must compute a 3000x3000 matrix and perform several operations over it, which in the end represents a considerable amount of operations that cannot carried on in real-time multimedia applications, where multiple sources are constantly moving in the scene, and thus, requiring interpolations constantly. A fine mesh could be pre-computed offline and used in real time, implying that it could be necessary to store vast amount of data that for a single room it could surpass 1GB for high resolution. Anyway, even pre-computing the first warp, since bilinear

interpolation is employed, it is impossible to pre-compute the last warp as it does depend on the interpolation point.

2.2.4 Other experimental non-linear interpolation methods

Finally, there are also some other emerging techniques to interpolate BRIRs that still are in development and need further research, but with promising results. This techniques exploit non-linear approaches such as deep neuronal networks or machine learning polynomial regressions.

In [17], the authors use Multi-Layer Perceptron neuronal network using hyperbolic tangent functions as activation nodes to interpolate non-linearly the HRIRs over a sphere, from a set of points on the measured database which are selected by using an unsupervised clustering algorithm. In [18], the authors use instead Radial Basic Functions with minimum phase HRTFs. Both researches demonstrate good performance of the interpolated samples in high angular resolution meshes. However, frequency resolution of the interpolated samples is still poor due to the need of reducing the incoming dataset to the training model.

The potential of these methods is that they are extremely fast to compute once the model is trained, and that, putting it to the extreme, they eventually could even provide a way of obtaining semi-personalized BRIRs or HRTFs from just some physiological features of ourselves, if enough popular data is used to feed the network. This would be a great contribution to audio spatialiation as personal HRTFs sound more convincing than general ones for each individual.

Chapter 3

Developed Algorithms

As mentioned in the introduction, this thesis aims to develop new algorithms that could overcome the flaws of the current interpolation algorithms and make them suitable for real-time applications.

The first algorithm is a BRIR interpolation algorithm which is inspired by the Dynamic Time Warping but with several modifications that will be discussed in the next sections. The second algorithm is a First Order Ambisonics B-Format interpolation algorithm that is based on the first algorithm and introduces a totally new method for interpolating this kind of audio format.

3.1 BRIRs interpolation algorithm

The motivation of this algorithm is provide a interpolation as effective as DTW, but with a reasonable computational cost to be implemented in real time, thus avoiding to store huge look-up meshes. This algorithm also pursues some kind of non-linear alignment of the signals before interpolating linearly sample by sample. Nonetheless, it is based on the assumption that significant information, for correct synthesis of spatialization, is condensed into few energy blocks on the BRIRs, which correspond to the direct impulse and early reflections. Hence, instead of warp or stretch the whole signals, as DTW does, one could identify those blocks, match them accordingly to some relationship criteria and leave the rest of the signal, which has an aleatory nature, as it is. Thus, computational expense would be considerably reduced while maintaining the perceptual quality of the resulting interpolated BRIR.

The whole block diagram of the algorithm can be observed in Figure 3.1. In the following subsections, this full processing flow for interpolating BRIRs will be explained.

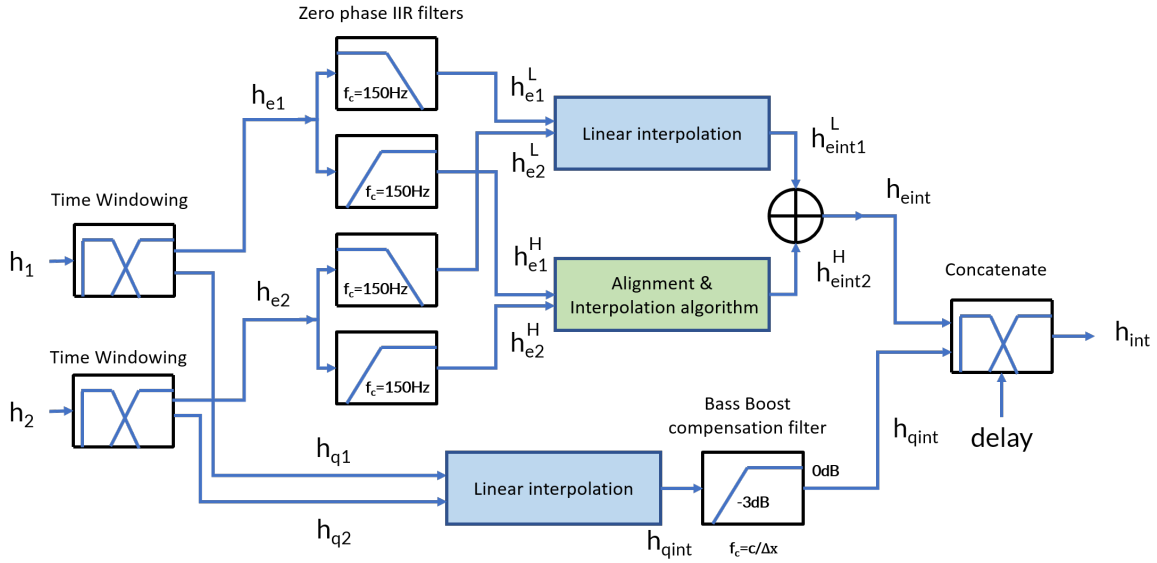

Figure 3.1: Full processing block diagram of the BRIRs interpolation algorithm. In green, it is highlighted the new developed alignment algorithm

First, cropping of the BRIR prior to interpolation is explained. Second it is described the steps followed by the algorithm to align and interpolate the early reflection components. Finally, a summary of the processing applied to the remaining parts of the BRIR and the reconstruction of the final interpolated BRIR is presented.

3.1.1 Windowing of the BRIR

Continuing with the reasoning of chapter 2, before interpolating, the BRIRs are split in two parts: first, the direct and early reflections and, second, the reverberant tail. Moreover, the initial delay due to travel distance between source and receiver has been subtracted from the BRIR, in order to avoid artifacts when aligning. Later, each part will be treated separately.

Then, lets denote h the left or right room impulse response measured at any position. It will be windowed as such:

$$delay = D \tag{3.1}$$

$$h_e = [h((D+1):(D+2049)] \tag{3.2}$$

$$h_r = [h((D+2049-ovlp):N)] \tag{3.3}$$

$$h = [zeros(D); h_e; h_r(ovlp+1:end)] \tag{3.4}$$

where D is the delay in samples until the first impulse with a margin of few samples, $ovlp$ is the number of samples of overlap between the reverberant tail (h_r) and the direct and early

reflections crop (h_e), and N the total number of samples of the full impulse response. Note that overlap is added because each region is treated differently and, so as to avoid discontinuity artifacts when recombining, crossover will be applied in those samples to provide a smoother union. Also, for convenience, a width of 2048 samples has been chosen for h_e, but it can be changed accordingly to the scenario. The following sections describe the processing applied to the second part containing the direct and early reflections.

3.1.2 Early echoes processing

Top flow of Figure 3.1 depicts the processing for the early echoes. As inputs it takes the two BRIRs cropped (Equation 3.2), h_{e1} and h_{e2}, measured at positions x_1 and x_2. Next, a sub-band processing is applied, splitting the signal into two frequency bands using low and high pass filters. From this point, two different lines of processing are applied. Low frequencies are simply interpolated linearly in time whereas mid and high frequencies are introduced to an alignment and interpolation algorithm.

Low band processing

The reason for dividing the processing in two parts is that low frequencies produce stationary modes in the room, as the wavelength is approximately of the order of the size of the room for that frequencies. Therefore, in the impulse response they are not perceived as peaks, because early reflection delays are even shorter than a period of these frequencies. In conclusion, there is not enough resolution in the cropped window for these frequencies.

In addition, time-domain interpolation involves a imbalance of the spectrum. It exists a general trend to overweight very low frequency components as they are more likely to be in phase, thanks to short distances between measuring positions, therefore this technique can help to balance the final mix. The cut-off frequency has been chosen on the basis of experimentation and perception around 150 Hz.

The filtering stage consists of two butterworth 3^{rd} order IIR filter, one low pass and other high pass, with cutoff frequency at 150 Hz. In order to preserve the phase linearity when filtering, so as to guarantee perfect synthesis later, bidirectional filtering has been applied. Basically, any IIR filter can be converted to a zero phase filter by filtering two times the signal with the same filter kernel, but one time in the reversed direction. This way, perfect reconstruction can be achieved by merely adding both signals in time. Low frequency signal is therefore interpolated linearly in time as shown in Equation 2.11.

High band processing

The high frequency signal will contain the blocks of energy we are aiming to align, so it will need further processing steps. Figure 3.2 depicts the steps taken by the alignment and interpolation algorithm for high frequency signals. It should be noted that left and right channels are processed and interpolated separately. Below, further details of each block in Fig. 3.2 are presented.

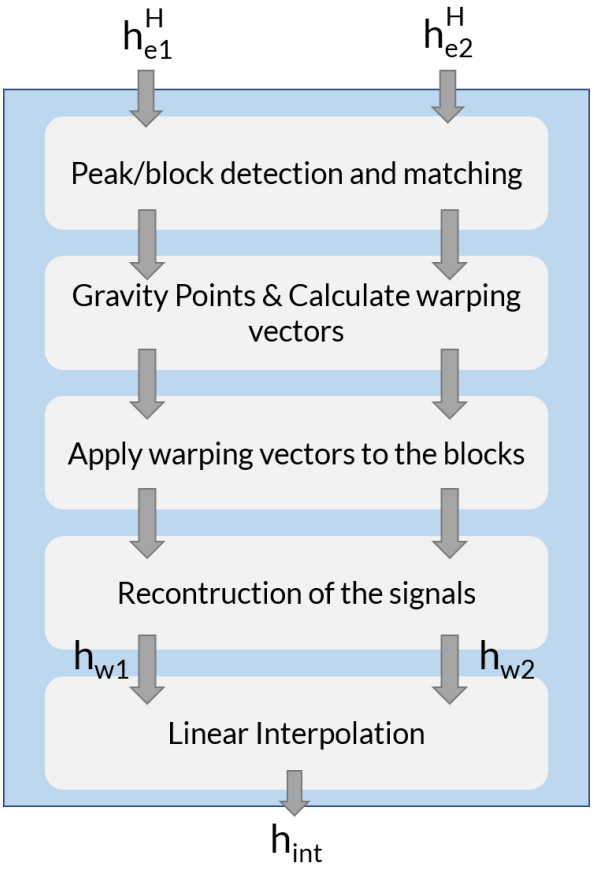

Figure 3.2: Alignment & Interpolation algorithm block diagram

A. Peak detection and matching

The first step of the alignment algorithm consist of identifying the blocks of energy that are related in both signals. To do so, M peaks are looked for in h_1 and h_2, with several constraints. Firstly, peaks must be spaced more than 100 samples, to discriminate peaks belonging to the same block of energy. Secondly, peaks under a weighted average threshold, described by

$$threshold = \sum_{i=1}^{i=M} \frac{\frac{1}{|p_i - \mu|} p_i}{\sigma} \tag{3.5}$$

where p_i denotes the value of amplitude of the peak i and μ and σ their average and variance values, respectively; are removed, as they are considered not of significant magnitude compared to the rest. This leaves a number of peaks equal or less than M in both signals, which has not necessarily be equal for h_1 and h_2.

Then, peaks in h_1 are related to the ones in h_2 by building a "distance" matrix where each value follows

$$coeff_{i,j} = \frac{1}{(1+\Delta s_{ij})(1+\Delta p_{ij})} \tag{3.6}$$

where Δs_{ij} and Δp_{ij} describe the absolute value of the difference in samples and amplitude, respectively, between peak i of h_1 and peak j of h_2. Recall that both values should be normalized to the range $[0,1]$ so the weighting is proportional. As an additional constraint, peaks spaced away more than 1000 samples are directly not related. Thus, high values of coefficient depicts potentially related peaks. Therefore, just by looking to the coincident maximum values in both columns and rows, indexes of the related peaks are obtained. Note that if more than one peak in h_1, for instance, is related to a single peak in h_2, only the one with the maximum value of coefficient is chosen and the rest are dropped. Hence, in the end, the same number of peaks are chosen in h_1 and h_2 with one-to-one relationship. This way, the algorithm is very robust to wrong alignments of the blocks that do not have a match in the other signal. Finally, maximum length blocks without overlapping (but never more than 600 samples, to avoid too much stretching), are formed around peaks. Recall that parameters values mentioned above, such as inter-peak spacing of minimum 100 samples, are values that worked fine for typical sampling frequencies and for our measurements. Nonetheless, they admit a refinement for each particular case.

B. Gravity Points and Warping vectors calculation

The second step constitutes one of the main advantages of this algorithm compared to other warping methods as DTW. Unlike DTW, which warps the signals to a "warped time-domain" and then needs to bring them back to the "unwarped time-domain" after interpolation, our new algorithms employs what we call gravity points to warp the signals to approximately the interpolation position. Therefore, interpolated signal does not have to be unwarped after interpolation nor the warping vectors need to be interpolated, which results in less operations.

In fact, gravity points represent the exact common sample index where we want the related peaks in h_1 and h_2 to coincide (Figure 3.4. They are represented in Figure 3.3 as the black squares. Indeed, they can be defined as a function of the position to interpolate, x_{int}, the positions x_1 and x_2 and the samples' indexes of the peaks. Thus, again using equation 2.11, gravity points can be obtained for each pair of blocks. As their name suggests, gravity points will produce an attraction force over the blocks, which will result in a non-linear stretching.

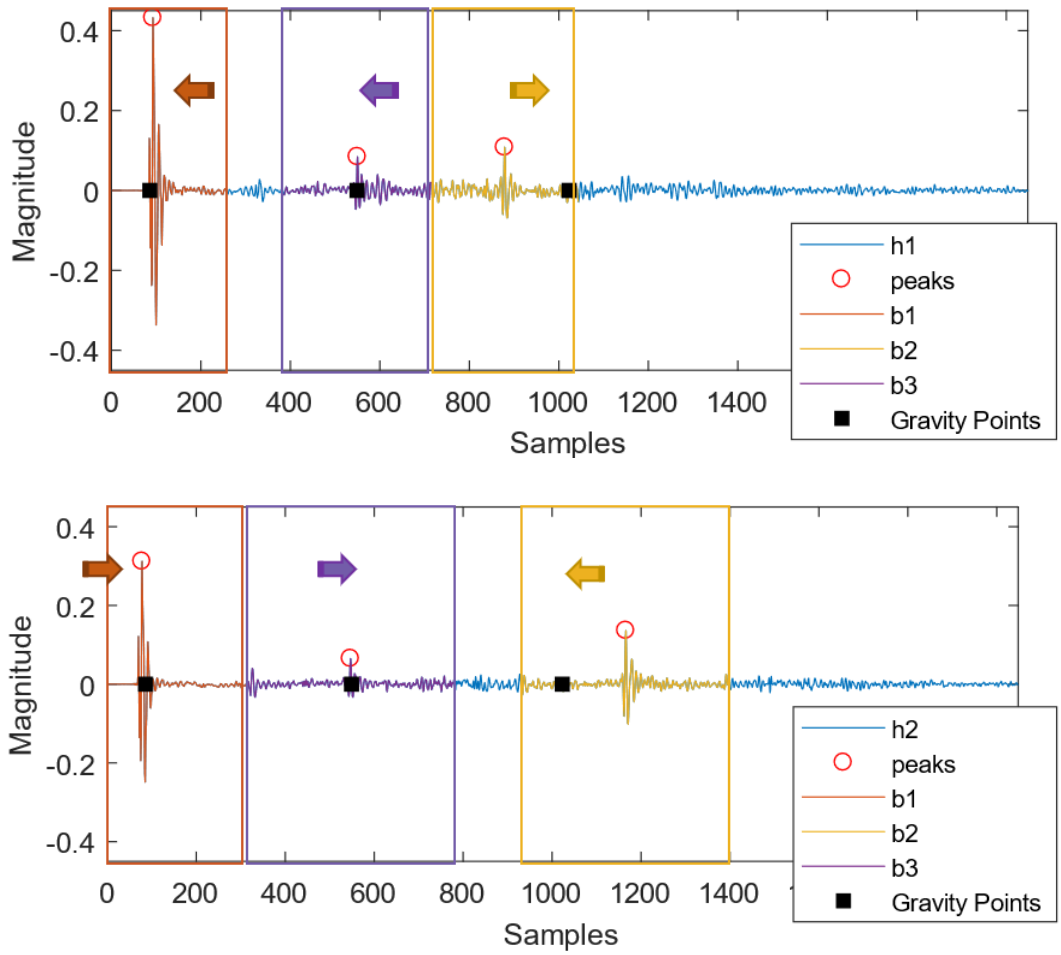

Figure 3.3: Related blocks and Gravity Points

This warping is based on the assumption that, within each block, there is a part where the impulse itself is concentrated, which must move as a single body (as a spring-mass system). This means that each sample must have the same displacement there. On the other hand, there must be another part which does not belong to the impulse which has not relevance for localization an thus can be considered as a spring which can be stretched, meaning that each sample is subjected to a different displacement.

Various functions could be used for modeling such displacement. We have opted for a sigmoid function, but cubic, hyperbolic tangent or other functions could also perform well provided that they are monotonically increasing functions. Sigmoid function is of particular interest because it comprises the range of values $[0,1]$ with a non-linear slope. It can be described and custom modeled as follows:

$$sigmoid = d_{max} \frac{1}{1 + e^{\pm \frac{(x - x_0)}{z}}} \tag{3.7}$$

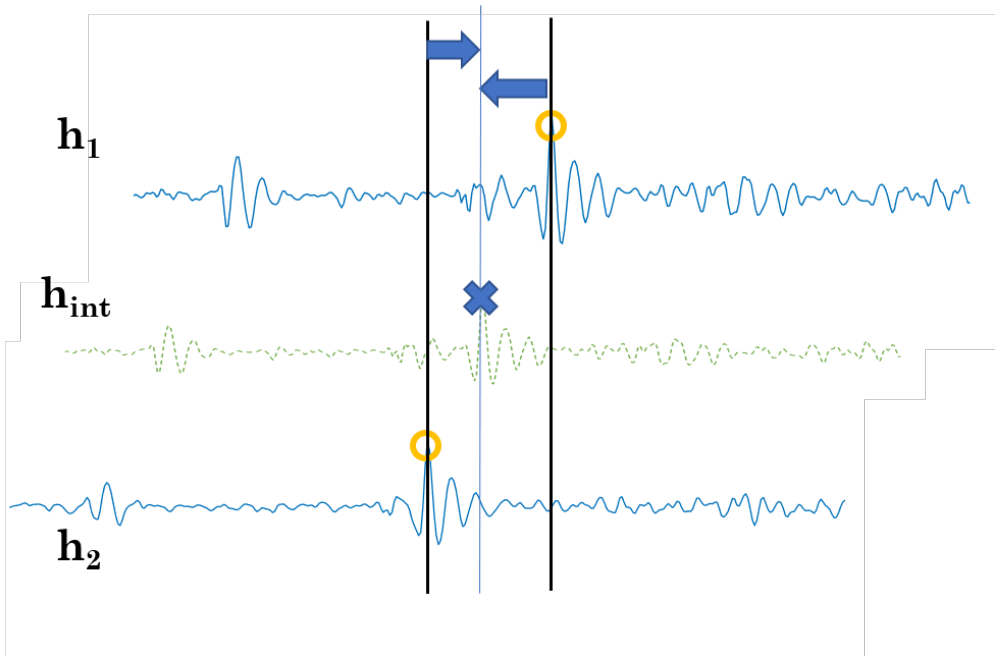

Figure 3.4: Gravity point concept

where d_{max} denotes the maximum displacement, in samples, to be applied to the block, x_0 is the sample index where to center the slope and z is a parameter used to control the slope rate (large values for smoother variations).

The maximum displacement for each block, indeed, is the distance from its peak to the gravity point. Thus, for each block in h_{e1} and h_{e2}, sigmoid function is modeled so that one extreme of the block remains motionless, maximum displacement is applied to the whole impulse and slope is applied to the part of the block where the energy is not concentrated (Fig. 3.5). This result into a particular warping vector for each block in both signals.

C. Warping and reconstruction

The final step consist into applying the aforementioned warping vectors to the blocks. This will be carried out in a "white canvas" so that extension of the blocks do not mess up the rest of the signal. Stretching of the signal will leave gaps between displaced samples belonging to the slope region of the sigmoid function. This gaps must be filled by interpolating linearly with the closest two non-void samples. Furthermore, warping is applied taking into account the limits of the signal, so that any sample that goes further is dropped.

Finally, the remaining warped signal is filled with the original samples of the signal at the indexes that have not been affected by the warping. Now, resulting signals can be interpolated linearly in time, likewise to the low frequency signal.

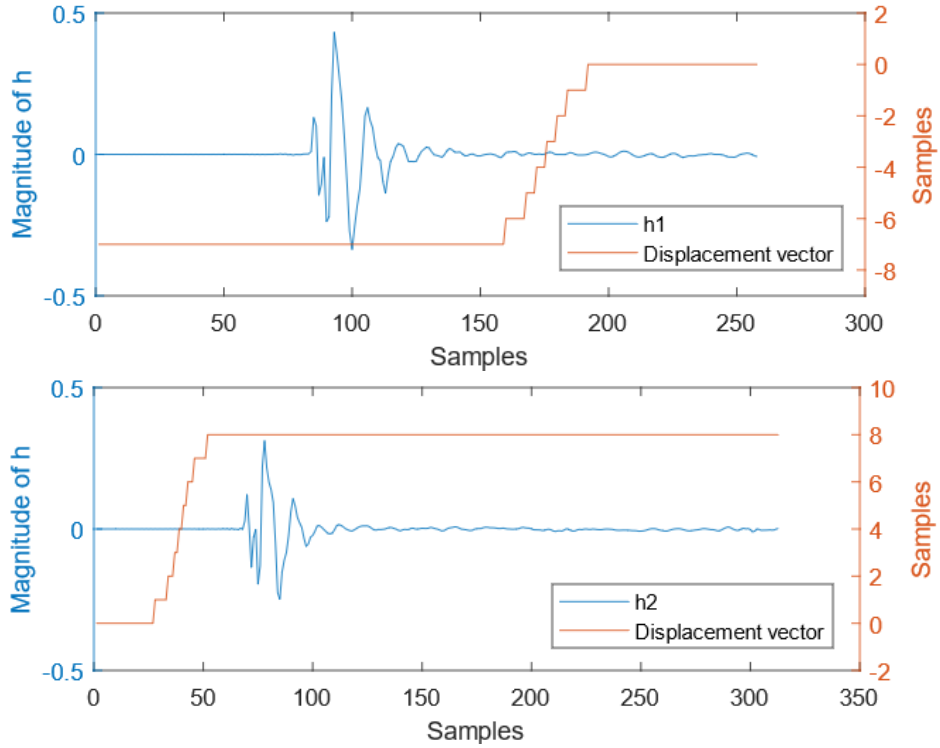

Figure 3.5: First blocks of h_{e1}^H (up) and h_{e2}^H (down) with its displacement vectors

3.1.3 Reverberation processing, delays and final mix

To recap, BRIR has been split in two parts: the direct impulse with early reflections, and the diffuse decay tail. The first part has been processed with our novel algorithm described above. However, to restore the interaural time difference of left and right ear BRIR, the delay must be reintroduced. Additionally, the reverberant tail must be joint again to the signal.

The reverberation tail is obtained by direct linear interpolation between the reverberation tails of h_1 and h_2. When interpolating the tails in time domain, distance between measuring positions of the BRIRs determine a transition frequency, below which, frequency components of the two signals to interpolate will be in phase, and, above which, will not. This results into a penalty for high frequencies that must be compensated in order to not alter the final timbre. A second-order shelving low pass filter has been applied after interpolation with an attenuation of 3dB centered at the transition frequency, obtaining good perceptual results. However, the transition frequency admits some refinement as a function of the room size.

Regarding the delay to be reintroduced, it is recalculated accordingly to the distance of each ear of the receiver to the interpolated audio source position by

$$delay_{int} = \frac{dist(x_{head}^{R/L}, x_{int}) f_s}{c} \tag{3.8}$$

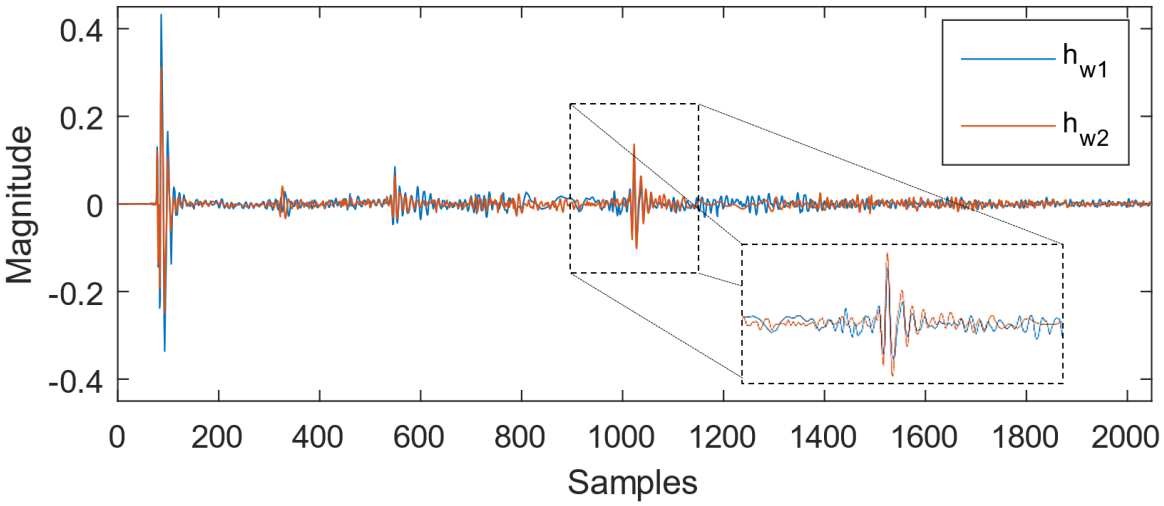

Figure 3.6: Resulting h_{w1} and h_{w2} aligned

where $dist(x_{head}^{R/L}, x_{int})$ describes the Euclidean distance between the right or left ear position and the interpolation position in meters, f_s denotes the sampling frequency in $samples/s$, and c is the speed of sound in the air in m/s.

Finally, to obtain the full interpolated BRIR, $delay_{int}$ zeros are padded before the cropped interpolated BRIR, and the interpolated reverberation tail is linked with it through a cross-fade, with weights defined by

$$w(i) = \sqrt{\frac{(i-1)}{(ovlp-1)}} \quad \text{for i = 1,2,...ovlp} \tag{3.9}$$

where $ovlp$ is the total number of samples that the tails overlaps the IR part. Hence, crossfade is performed as:

$$h_{int}(j) = w(j)h_{eint}(j) + (1 - w(j))h_{rint}(j) \quad \text{for j = N-ovlp,...N} \tag{3.10}$$

where N is the total number of samples of IR. Rest of the interpolated tail is added to the end as it is. Resulting interpolated signal h_{int} is compared against the reference h_1 and h_2 in Figure 3.7. Results are discussed in the next chapter.

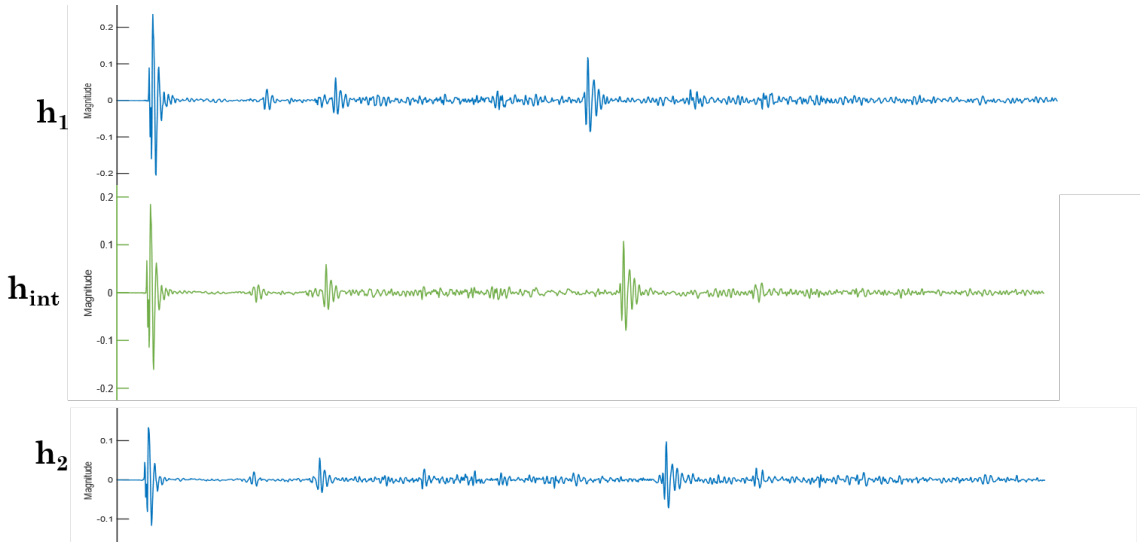

Figure 3.7: Resulting interpolated signal (green)

3.2 First Order Ambisonics B-Format interpolation algorithm

The motivation this time is to design a new FOA B-Format interpolation algorithm suitable for real-time. No expectations about quality could be defined as there were no other algorithms to take as reference. Nevertheless, the algorithm described for BRIRs will serve as the base upon which to build this new algorithm.

The difficulty of interpolating Ambisonics rely on the accurate estimation of the direction of arrival (DOA) of each impulse, which should also need to be interpolated and used later for reconstruct the B-Format signals. Besides, regarding computational cost, the similarities between W, X, Y and Z should be exploited in order to avoid to repeat processing between the four signals, thus reducing the total number of operations.

The whole block diagram of this algorithm resembles the one of the previous section. Again, signals are split between main impulse plus early reflections and reverberation tails. Later, these two blocks feed different processing blocks. The difference now is how early reflections are treated. In figure 3.8, it can be observed the new block diagram for dealing with the main impulse and the early reflections. In the following subsections, this processing flow for interpolating FOA B-Format signals will be explained. First, the search of the blocks with higher energy on the signals is explained. Second, it is described the calculus of the direction of arrival of each of those blocks, followed by the matching of those blocks between the two measurement positions. Next, interpolation itself is explained using all this information. Finally, the reconstruction of the B-Format signals from the results of the previous step is addressed.

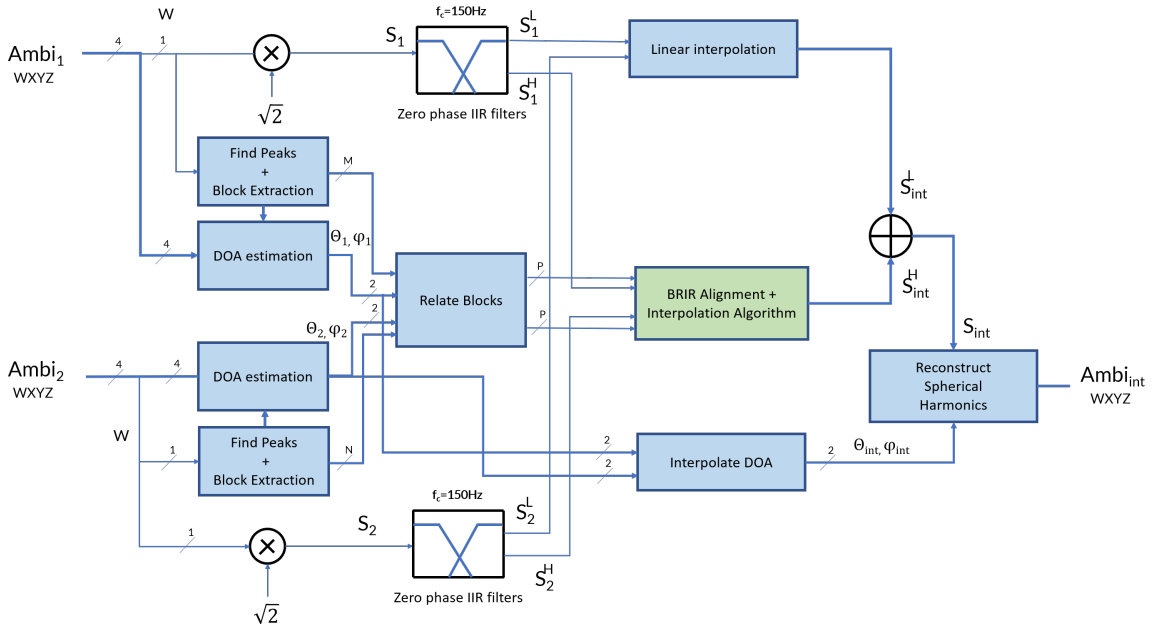

Figure 3.8: Block diagram for processing the main impulse and early reflections of B-Format signals. In green, it is highlighted the algorithm used to align and interpolate BRIRs.

3.2.1 Block search and extraction

First part of the algorithm is quite similar to the one of the BRIR interpolation algorithm: peak detection. For this algorithm, it is also assumed that important information for localization is condensed in just few blocks of energy on the whole signal. Hence, the same peak detection algorithm is applied here. To recap, K peaks are looked for in the signal with a minimum separation, and then a dynamic threshold is calculated to eliminate the peaks with lower energy, leaving M peaks. Recall that for the two measurements this number could not be the same.

The unique difference in this case is that the size of the blocks is limited even more, as it is convenient for a precise DOA estimation. The maximum block size for the measurements used has been set to the minimum separation distance between peaks. However, it admits refinements depending on the measured signals.

Notice that in this case, this process is only carried out on the W signal. The reason for this is that, according to the equations describing the FOA B-format signals (2.2, 2.3, 2.4 and 2.5), they all share the same common information, the signal S, which is basically W. Hence, the peaks appearing in W should also appear in X, Y and Z in the same location.

3.2.2 DOA estimation

Once the blocks have been identified for both measurements, next step is to obtain the direction of arrival of each impulse. The motivation behind this step, is that if we get the DOA for the impulses in W, we can discard the signals X, Y and Z for the rest of the processing, as they do not add any more valuable information, thus working with just one signal. Later, DOA will also be interpolated likewise the impulses, and the interpolated signals X, Y and Z will be reconstructed using the interpolated angles. Furthermore, the block matching step will be enhanced by taking into account the DOA of each impulse as well.

There are several ways to obtain the DOA. The method presented in this thesis is inspired by the analysis stage of the technique Spatial Impulse Response Rendering, presented by Merimaa and Pulki in [19]. This approach consists of an energy analysis of the sound field based on the concept of sound intensity. The instantaneous intensity is defined as the product of the sound pressure p(t) and the particle velocity vector u(t) as follows:

$$I(t) = p(t)u(t) \tag{3.11}$$

which in frequency domain is equivalent as

$$I(\omega) = P^*(\omega)U(\omega) \tag{3.12}$$

In a B-Format microphone system for FOA, the pressure can be regarded as the W signal

$$p(t) = W(t) \tag{3.13}$$

meanwhile the figure-8 signals X, Y and Z are proportional to the components of the particle velocity in the corresponding directions of the Cartesian Axes. Hence the particle velocity vector results in

$$u(t) = \frac{1}{\sqrt{2}Z_0} X(t)e_x + Y(t)e_y + Z(t)e_z \tag{3.14}$$

where e_x, e_y and e_z are the unit vectors of the three directions and Z_0 is the impedance of the air.

Putting all together, the intensity vector can be obtained for each sample/frequency bin, depending on the domain chosen

$$I(\omega) = \begin{bmatrix} I_x(\omega) \\ I_y(\omega) \\ I_z(\omega) \end{bmatrix} = \frac{\sqrt{2}}{Z_0} Re\left\{ W^*(\omega) \begin{bmatrix} U_x(\omega) \\ U_y(\omega) \\ U_z(\omega) \end{bmatrix} \right\} \tag{3.15}$$

where Re denotes the real part and W^* the conjugate of the complex value of W.

Azimuth and elevation angles are calculated straightforwardly from there:

$$\theta(\omega) = tan^{-1}\left[\frac{I_y(\omega)}{I_y(\omega)}\right] \tag{3.16}$$

$$\phi(\omega) = tan^{-1}\left[\frac{I_z(\omega)}{\sqrt{I_x^2(\omega) + I_y^2(\omega)}}\right] \tag{3.17}$$

Note that, these equations also apply in time domain, sample per sample.

However, with this, only the direction of arrival of a certain frequency bin or sample can be obtained. For a whole block, which contains several samples, each one with a particular DOA, some statistical analysis must be applied. Only considering the sample of the maximum peak could also be a valid approach but, after trying several methods the statistical one has proven to be the most reliable one.

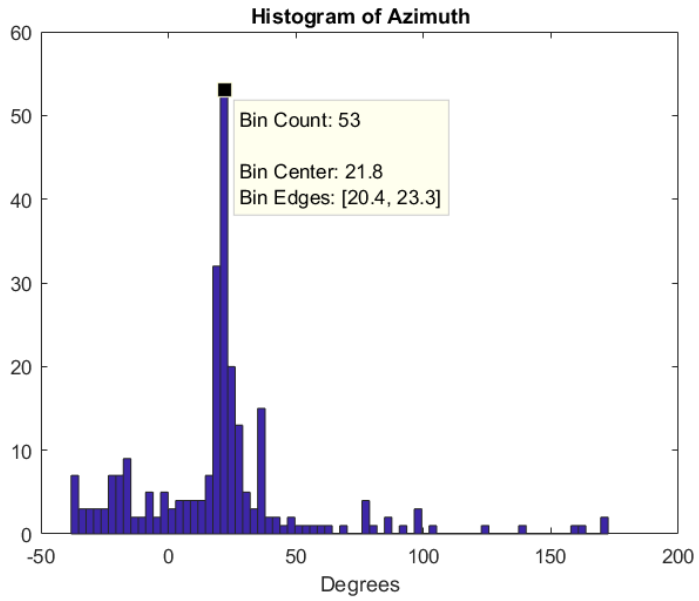

Figure 3.9: Histogram of calculated azimuths. Resolution of less than 3 degrees.

Hence, the implemented version, computes the spectra of each block and obtains the DOA per each frequency bin, with enough bins per block (value used was 512 bins for a single side spectrum). Then, it computes an histogram of the DOA (two in fact, one for azimuth and one for elevation) with 72 bars which, at worst, gives a resolution of 5 degrees (Figure 3.9. The DOA of arrival of that block will be the center of the greatest bar, i.e, the most repeated DOA value inside the block.

3.2.3 Block matching

Next step consists of matching the blocks. The approach followed here is quite the same as in the other algorithm, but this time the DOA is also considered. Peaks in W_1 are related to the ones in W_2 by building a "distance" matrix where each value follows

$$coeff_{i,j} = \frac{1}{(1+\Delta s_{ij})(1+\Delta p_{ij})(1+0.5\Delta\theta_{ij})(1+0.5\Delta\phi_{ij})} \tag{3.18}$$

where Δs_{ij} and Δp_{ij} describe the absolute value of the difference in samples and amplitude, respectively, between peak i of W_1 and peak j of W_2. $\Delta\theta_{ij}$ and $\Delta\phi_{ij}$ have a weight of 0.5 so as to count as one single feature. Recall that all values should be normalized to the range $[0, 1]$ so the weighting is proportional. Maximum coincident values between rows and columns are kept to relate the peaks, and consequently the blocks.

The outputs of this step will be an equal number of peaks with one-to-one relation between W_1 and W_2.

3.2.4 Interpolation

Continuing with the previous reasoning, the only signal that needs to be interpolated should be the one that is shared amongst the four signals, the S signal. Moreover, it can be obtained simply from W just by multiplying W by $\sqrt{2}$. Then, to obtain the interpolated X, Y and Z, the angles of DOA must be interpolated and applied back to the interpolated base signal S_{int} with the first order spherical harmonics equations. This method is way faster than interpolating each of the signals separately. Also it ensures a uniform processing amongst the four signals.

Sub-band processing

Likewise the BRIR interpolation algorithm, interpolation of the early reflections is split in two frequency bands, due to the very same reasons as before. A zero phase 3rd order butterworth filterbank is employed to separate the bands with 150 kHz as the cut-off frequency. Low frequency component is interpolated linearly meanwhile the high frequency component is interpolated after a previous alignment using the method of the gravity points and same warping vectors as in the BRIR algorithm. Resulting signals are added up again to form the interpolated base signal S_{int}.

DOA interpolation

To interpolate the DOA of each block in a proper way, a new algorithm based on the well known image source method (ISM), from room acoustic modeling, has been developed. The source image method is an algorithm that identifies the images of real sound sources by a mirror effect on each wall, which has multiple applications in the analysis of the reflected impulses. In Figure 3.10, image source of a first order reflection is illustrated. It is clear that the total path of the sound ($r_1 + r_2$) is equal from both sources, but for the image source it becomes a straight line. Higher order reflections can be modeled as well by keep mirroring the image source on the wall where the ray bounces.

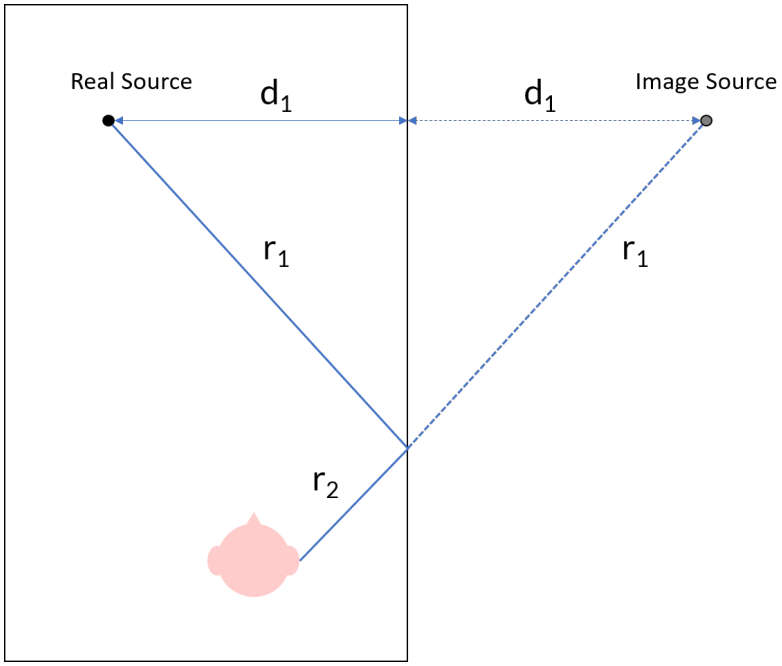

Figure 3.10: Image source method representation.

For DOA interpolation it is of particular interest as it greatly simplifies the geometrical calculations to obtain the interpolated angle. It does not matter which reflection order the ray is, using this method the path becomes a straight line between the image source and the listener. Hence, calculation of the interpolated azimuth and elevation is achieved after interpolating the position of the image source of the source at the real position where interpolation occurs, which is relatively easy to obtain (Figure 3.11).

First step is to compute the total length of the path that the impulse under study has traveled. This is calculated directly from the sample position at which the peak is:

$$r = \frac{s \cdot c}{fs} \tag{3.19}$$

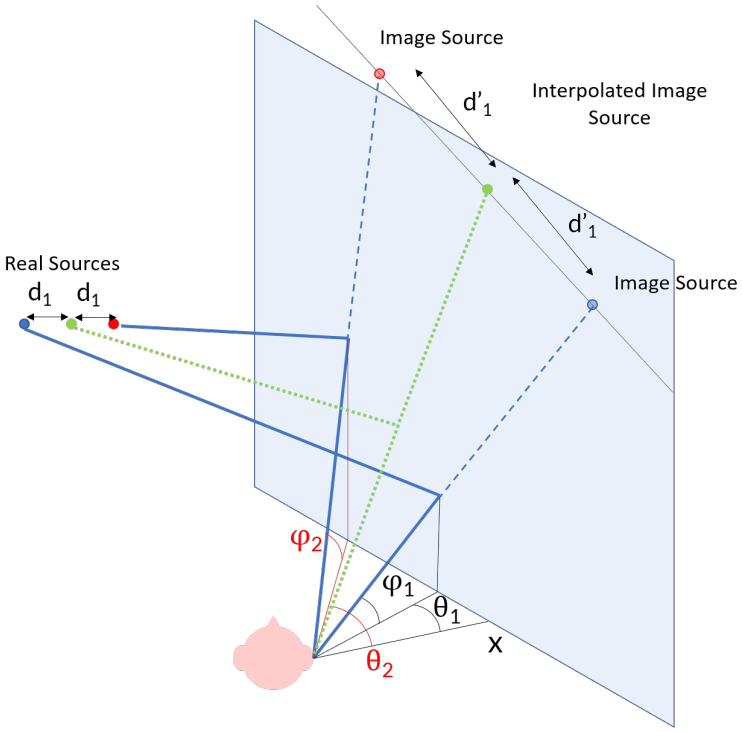

Figure 3.11: DOA interpolation visual representation. Image source method position is calculated to get the interpolated angles.

where fs is the sampling frequency, s the sample position and c the sound velocity in m/s.

Then, position of the two image sources of the sources at the measured positions are obtained by converting the spherical coordinates (θ_i, ϕ_i, r_i) to Cartesian coordinates (x_i, y_i, z_i). Note that θ_i and ϕ_i, for each block, are known from the DOA estimation stage. Recall that these coordinates are relative to the position of the listener, taking it as the origin of coordinates.

Once obtained the image sources, the ratio of the distance between the two image sources in the 3D space and the distance between the real source positions is calculated. Also, the distance between the real interpolation position and one of the real sources, d_1 in the figure, is calculated according to the Euclidean distance formula

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \tag{3.20}$$

With d_1 and the ratio between sources separation, the distance between one of the image sources and the interpolated image source can be obtained (d_1' in the figure):

$$d_1' = d_1 \cdot R \tag{3.21}$$

being R the ratio between separation of the real and image sources. Note that in the figure,

interpolation takes place at half way for simplicity, but it can be carry out closer to any of the sources.

Afterwards, to obtain the exact position of the image source, a system of equations must be resolved. Lets say that the image sources 1 and 2 are placed at the points $p_1 = (x_1, y_1, z_1)$ and $p_2 = (x_2, y_2, z_2)$, respectively. The interpolated image source must be placed in the straight line connecting those two points. This line has a direction vector $u = p_2 - p_1$ which can have two senses. Applying this direction vector in the two senses with a modulus equal to the calculated distance d'_1 from p_1 leaves two possible points where the interpolated image source could be:

$$p^1_{int} = p_1 + u\frac{d'_1}{|u|} \tag{3.22}$$

$$p^2_{int} = p_1 - u\frac{d'_1}{|u|} \tag{3.23}$$

where $|u|$ is the modulus of the vector u.

Of course, only one is of interest and is the one that is closest to the other image source, that is, the one that minimizes the distance with p_2. Once obtained the interpolated position of the image source, the interpolated azimuth and theta are obtained just by turning back the coordinates to the spherical system.

3.2.5 Reconstruction of the spherical harmonics

The last step is to reconstruct the W, X, Y, and Z signals from S_{int}, θ_{int} and ϕ_{int} by applying the equations 2.2, 2.3, 2.4 and 2.5, respectively. Note that this equations must be applied individually to each block of the interpolated signal since azimuth and elevation angles have been obtained in that fashion too.

One consideration here is that, when multiplying S_{int} with cosines and sines, which are functions whose multiplication also lies in the range [0,1], the result will be always equal or smaller than originally. This implies that it is likely that if the sound is coming from a direction which is close to a null in the polar pattern of that virtual microphone, impulses will become so small with respect to the rest of the signal. This is totally fine, but in order to avoid a great imbalance of the final mix, the remaining parts of the signals X, Y and Z are renormalized according to the penalty that the impulses suffer.

To do so, prior to apply the spherical harmonics equations, the total energy of the blocks is calculated according to

$$Eb = \sum_{i=1}^{N} \sum_{j=init}^{end} S_{int}(j) \tag{3.24}$$

where N denotes the total number of blocks and $init$ and end, the first and last sample of that block.

Later, after applying the equations, this calculus is performed again independently for X, Y and Z and the ratio between the energy after and before is used to renormalize the rest of the signal.

3.2.6 Wrap up

To summarize, early reflection processing, involves a analysis stage where DOA information is extracted from the four signals for each block of energy, and a synthesis stage where the four signals are reconstructed by using the interpolated data. This flow allows one to work with just the common signal S, speeding up the whole process.

Recall that delay and reverberation tails must be processed and reintroduced to form the final impulse responses. The processing of each of those part is identical to the one of the BRIR algorithm (Section 3.1.3) and reconstruction is performed with the same crossover between tail and early reflections as before. In this case, the tails are interpolated independently for each of the four signals whereas delay is shared amongst them. Results will be discussed in the next chapter.

Chapter 4

Experiments and Results

4.1 Measurements

Since the beginning of this thesis, all the data employed to design and check the performance of the new algorithms was real-world data. Therefore, the algorithms have been tested with real room BRIRs and FOA B-Format signals and, in the case of the BRIR algorithm, compared in terms of performance and quality with other interpolation methods.

To evaluate both algorithms, BRIRs and RIRs in FOA B-Format have been measured at multiple points in different rooms, following a grid similar to the one presented in Figure 1.1. In the case of the BRIRs, three rooms have been measured with different grids sizes. These are: in the large room, 12x8 m^2, in the medium-sized, 6x6 m^2, and in the small, 2.5x5 m^2, with resolution of 1 m, 0.5 m and 0.5 m, respectively. On the other side, for Ambisonics, only a small room has been measured with a grid of a size 2x5 m^2 with 0.4 m spacing. However, for interpolation, twice these resolution values have been considered in order to always have a real measurement point to be compared with the interpolated one.

The equipment for measurement BRIRs was composed by a B&K model 4100 acoustic mannequin, a flat response DSP processed Dynaudio Air6 loudspeaker, a Roland Octamic sound card with a computer, and an Optitrack Flex 3 infrared camera positioning system to get the current position of the loudspeaker when measuring. Logarithmic sweep signals, from 20 Hz to 20 kHz, were employed for obtaining the BRIRs through correlation [20]. This way, one ensures to excite all the audible frequencies in a room, resulting in a high quality impulse response. For Ambisonics, a Sound Field microphone was used to get the FOA B-Format signals.

An example of the set up used to measure an auditorium (large room) can be seen in

Figure 4.1. As said, the process of measuring is laborious and time consuming. For that large auditorium, the loudspeaker was placed manually one by one in over 64 positions. In all of them, it was put facing towards the mannequin to measure the impulse response. With this orientation, localization is maximized. However, in a more complex system, it could be also possible to measure in a position with several loudspeaker orientations and then interpolate between the orientations to provide the source with another degree of freedom, ie. rotation.

Figure 4.1: Measurement set up for a large auditorium. Infrared cameras were used to track the loudspeaker position

One last consideration about measuring is that, during the course of this thesis, an automated measurement system was devised and tested. This system intended to track the position of the loudspeaker with an infrared camera system and send and record the response automatically when the loudspeaker is the expected position. A custom Matlab script was in charge of synchronizing and orchestrating all the tasks between the camera system and the sound card. This way, the person measuring just need to move the loudspeaker continuously and forget about sending the impulse and saving the measure on the computer. The graphical user interface (GUI) of this script is shown in Figure 4.2. Nonetheless, this system has only proved to be useful and time saving in small rooms, as tracking became hard due to the limited range of the infrared rays in large rooms.

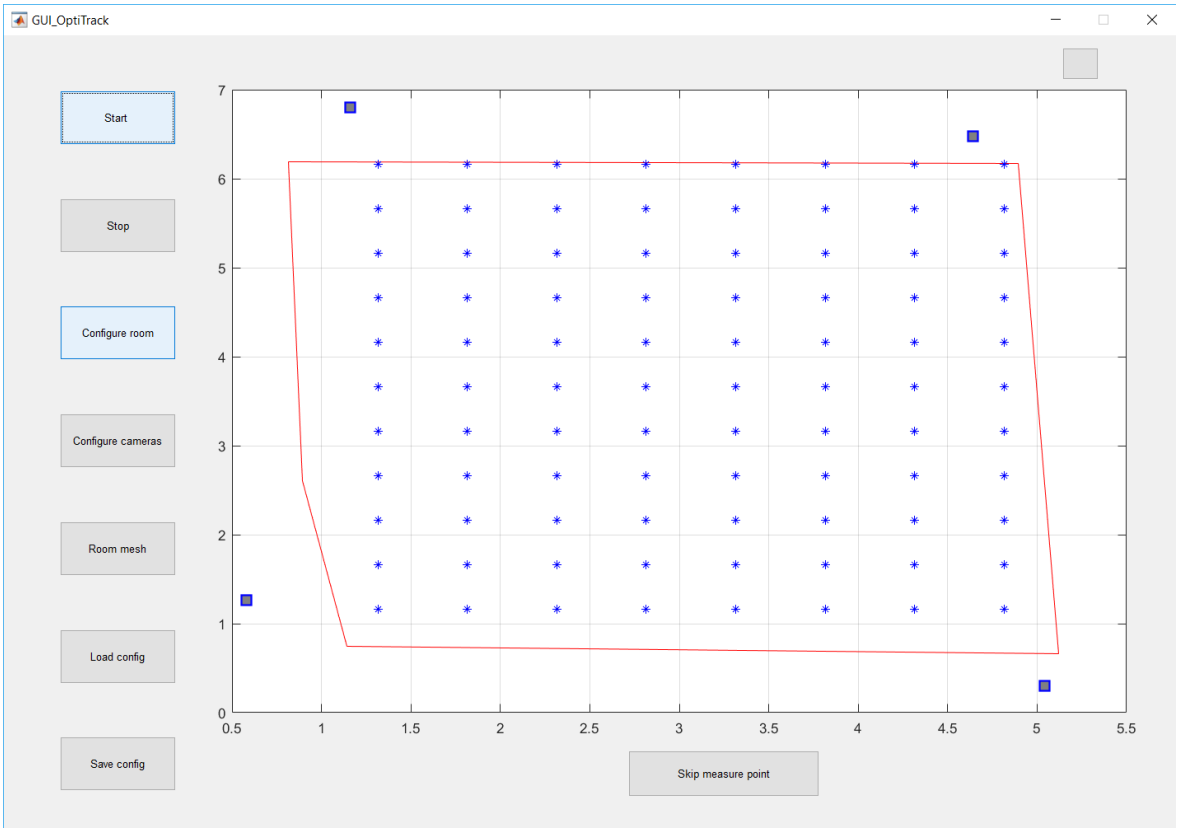

Figure 4.2: Matlab GUI to track the loudspeaker and sent and record the impulse response automatically. Squares represent the cameras and asterisks the grid or measurement positions to perform.

Finally, in order to compare the performance of the first algorithm, different interpolation methods have also been implemented, including a full implementation of the DTW interpolation algorithm presented in [7], the most straightforward time domain interpolation, and a frequency domain interpolation. Once implemented, they were tested using BRIRs of the three rooms and, at the same time, they served to evolve and refine the algorithm by introducing improvements.

On the Ambisonics side, the only way to evaluate the algorithm was subjectively with respect to the real measurement, as there were no other methods to compare with. Besides, it was also put in the spot for computational cost.

4.2 Discussion

In this section, both algorithms will be analyzed in terms of general performance, computational cost and subjective perceptions.

4.2.1 BRIR algorithm

General performance analysis

After running the different algorithms for BRIRs interpolation, the following observations could be collected. On the one hand, time domain interpolation stands out for being the fastest method. Notwithstanding, it presents replication of the impulses as the time peaks of the two BRIRs to be interpolated are not coincident in time. On the other hand, frequency domain interpolation has the advantage of not being dependent on the temporal alignment of the signals and it provides a good timbre quality result. However, it presents strong smearing in time if there exist accused early reflections, and it requires some more time for calculation. Dynamic Time Warping outperforms the rest in terms of quality, but not in computational efficiency.

Regarding the proposed algorithm, interpolated result of two BRIRs in time and frequency domain, compared against the measured BRIR at the interpolation point, is shown in Figures 4.3 and 4.4. The interpolation point was at the center of a quadrant so, result has contributions each BRIR at the four corners.

It can be observed that in time domain, main impulses are faithfully represented, whereas rest of the signal is quite similar but not equal. The first reflection impulse is slightly displaced in time, denoting that the movement of the early reflections might be better adjusted with a non-linear model. Flexibility of our algorithm makes this easily achievable by merely changing the Gravity Points equation (Eq. 2.11) by a non-linear one. It could be considered for future improvements.

Furthermore, Figure 4.3 exposes one of the limitations of interpolating RIRs: one cannot synthesize reflections not appearing in the RIRs used to interpolate. In the figure, a clear example can be seen around sample 520. This reflections will seldom occur due to the room geometry and the concrete presence of objects between source and receiver.

On the other side, Figure 4.4 reveals good adjustment of the magnitude of the spectrum to that of the measured BRIR, specially for frequencies above 2 kHz. In the figure, it is also compared the spectrum of the interpolation through DTW, showing that low frequency deviation, produced either by signal stretching or time domain interpolation, is properly solved in the proposed algorithm, thanks to the filtering stage.

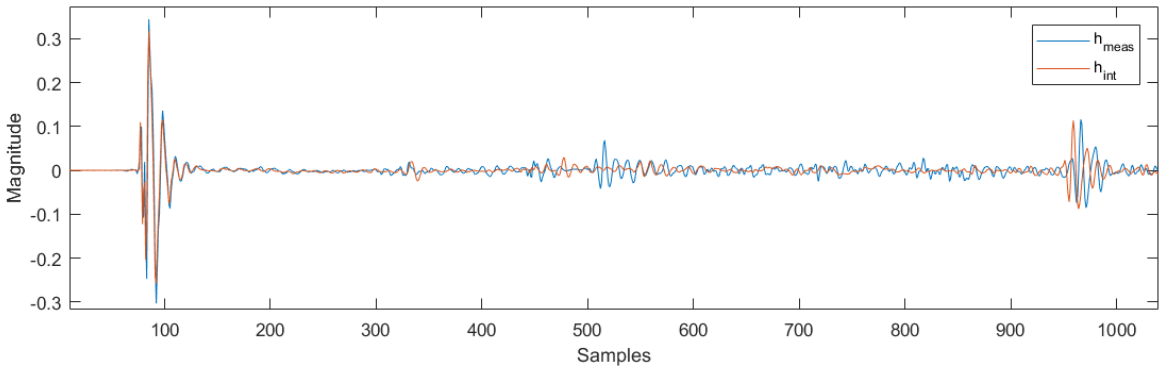

Figure 4.3: New algorithm interpolated BRIR versus measured BRIR (one channel).

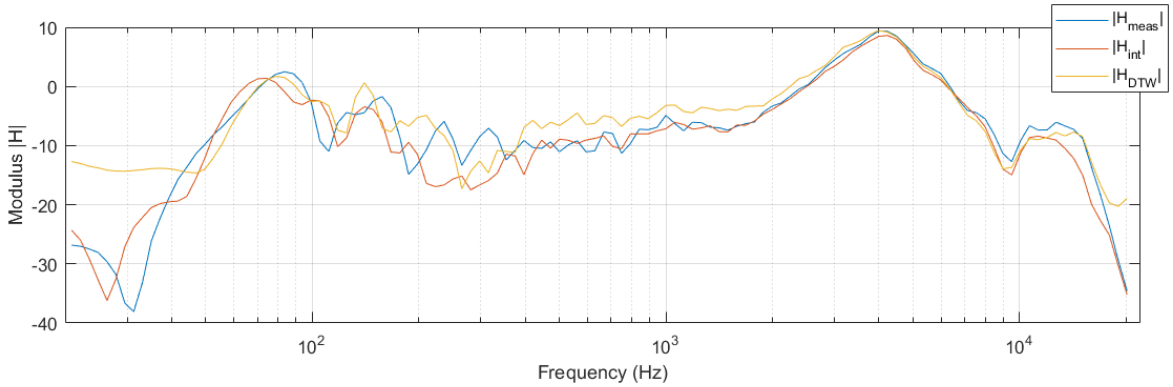

Figure 4.4: New algorithm interpolated BRIR versus measured BRIR vs DTW interpolated BRIR (one channel).

Preliminary subjective evaluation

To complement objective results, it is mandatory in this kind of audio algorithms to perform a subjective perceptual test. A preliminary MUSHRA test has been performed. In this test, subjects were asked to evaluate three samples of BRIR, interpolated with direct time interpolation, DTW and the proposed algorithm; in terms of source localization and timbre similarity with respect to a reference (the measured BRIR at the interpolation points). This was performed for six locations at just one of the measured rooms, the large one, in order to simplify the test. In total, 9 expert listeners related to audio engineering carried out the test. The sound sample consist of a sequence of guitar chords. Scores were rated from 1 to 5, being 1 not equal and 5 identical. Interpolated samples were shuffled with the reference and an anchor to check the significance of the answers. The test lasted around 30 minutes per person. The interface used for the test can be seen in Figure 4.5

Results of the test are shown in Figure 4.6 and summarized in Table 4.1. Highest scores were given to the new algorithm both in terms of location and timbre accuracy. Though source samples from Dynamic Time Warping interpolation were supposed to have better results, they were perceived somehow widened. This can be caused by the amplification of the low

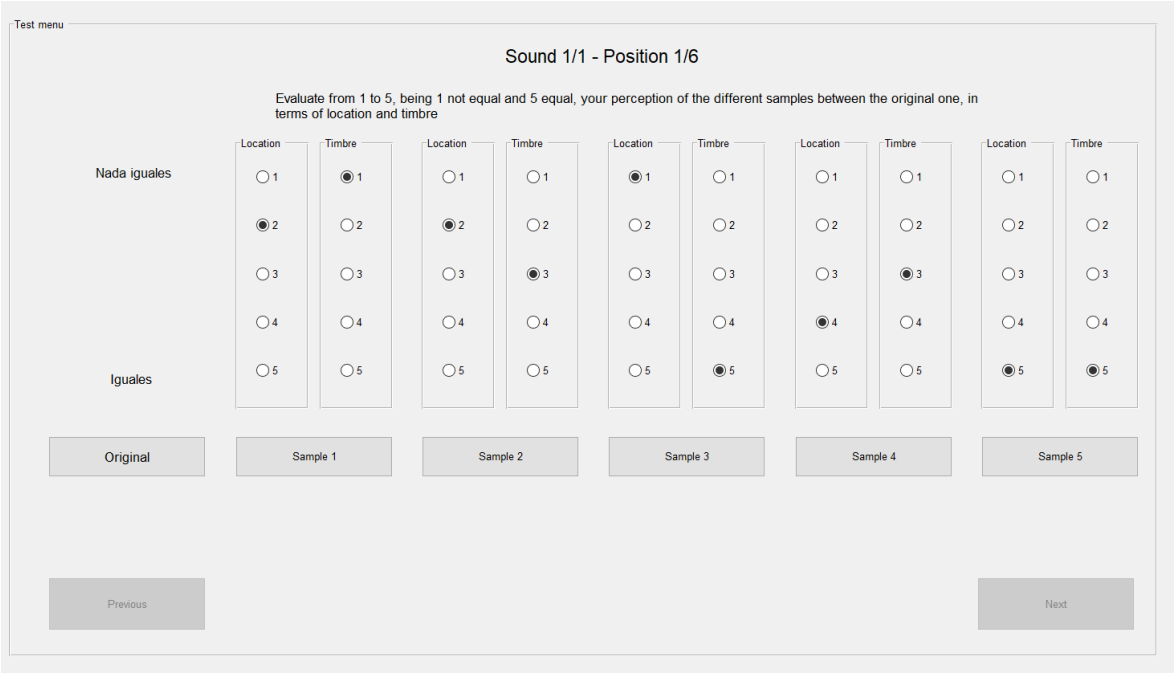

Figure 4.5: MUSHRA test interface in Matlab.

frequencies and, in turn, it affects spatialization.

Method	Location score (average)	Timbre score (average)
DTW	3.75	3.65
New	4.38	4.05
Time	3.47	3.84

Table 4.1: Averaged scores of preliminary subjective test for the BRIR algorithm

Computational cost

Table 4.2 depicts average timing performance running bilinear transformation with DTW, the new algorithm and direct linear interpolation in time domain, using MATLAB. However, they may decrease a lot using a compiled programming language as C, maintaining the percentage of reduction. Results reveal a noticeable time reduction with the new algorithm, making it suitable for real-time applications. Recall, that DTW has only been applied to the main impulse and early reflections as in [7].

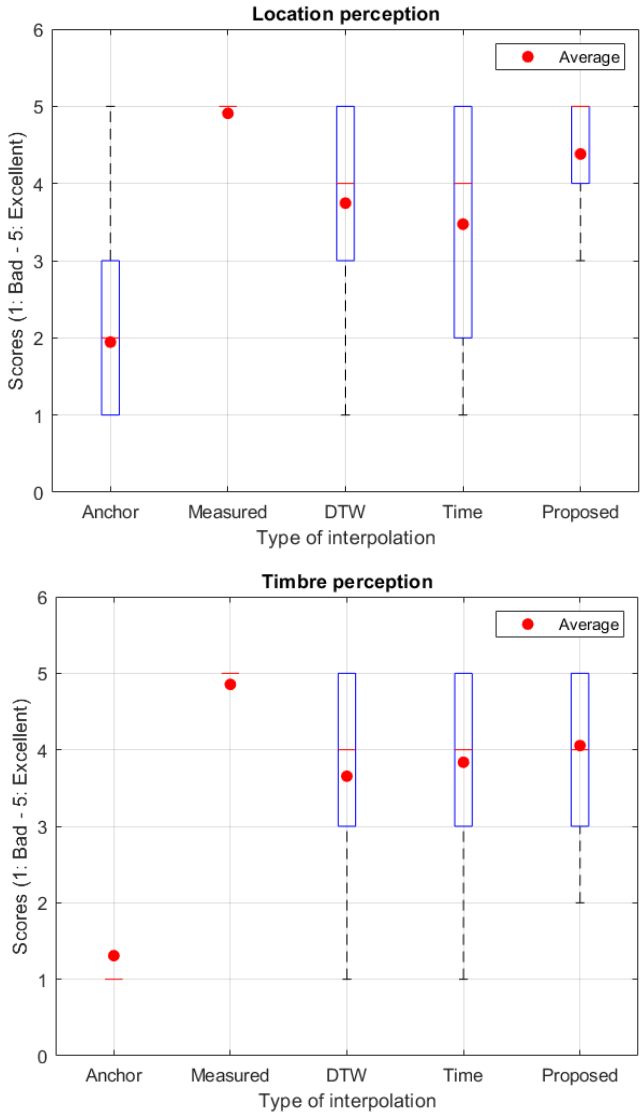

Figure 4.6: Subjective test scores for the BRIR algorithm

Furthermore, these algorithms have fed a real time interface in Matlab that allows the user to move a sound source in real time within the measured BRIRs grid. Interpolating with DTW resulted in lagged sound when the source was moved, compared to the smooth transitions provided by the proposed algorithm.

4.2.2 Ambisonics algorithm

General performance analysis

In this scenario, a direct comparison between the measured and interpolated B-Format signals reveals less information than in the BRIR case. This is because, as commented

Method	Time (ms)	% Reduction
DTW	156.0	-
Proposed	21.5	86.20
Time	0.6	99.63

Table 4.2: Timing performance comparison for BRIR bilinear interpolation using different algorithms

before, Ambisonic formats are independent of the reproduction format. Therefore, without considering the decoding stage, there is no clear evidence that a certain shape or feature in the signals could lead to an artifact in reproduction. Nevertheless, what is clear is that, if the same Ambisonics to Binaural decoder were to be used for the real and interpolated signals, the more they resemble each other, the more likely they are to be perceived equally. Figures 4.7 and 4.8 illustrate for a given position, the measured B-Format signals and the interpolated signals at that position, respectively.

Again, it can be observed that in time domain, main impulses are faithfully represented, whereas rest of the signal is quite similar but not equal. Reconstruction from W, has been quite accurate in X, Y and Z, in terms of direction (sign of the impulse), which denotes a correct DOA estimation. However, in those signals, there are small differences in the amplitude of the impulses and the shape of some of them, with respect to the measured ones. One explanation could be, that although theoretically, the impulses X, Y and Z could be reconstructed from W by applying the spherical harmonics equations, this only is valid if the impulses are sparse and well differentiated. It could happen that a same impulse has contributions from reflections of different order arriving a the same time, which modifies the shape of the greatest one. However, in the processing followed, only one DOA is estimated for a single impulse, hence ignoring the rest of contributions in reconstruction.

Preliminary subjective evaluation

Finally, likewise the BRIR algorithm, a preliminary subjective test has been carried out to evaluate the performance of the algorithm in terms of perception of the location and timber. The same MUSHRA test interface has been employed for simplicity but, this time, the testers were asked for a different thing as only the reference measurement can be evaluated against the interpolated. This test campaign has started very recently and will continue over the deadline of this memory. However, first results are optimistic until all the final data from the campaign is collected.

In addition, in this case there is a subtle difference with respect to the BRIR algorithm: signals must be decoded from B-Format to binaural. This step could seem trivial but it has a

huge impact on reproduction. For this reason, a decoder has been picked up and implemented using some well-known Ambisonics B-Format decoding libraries [16], [13]. The selected decoder has been the All Round Decoder from Matt Zotter. It decodes to a cubic loudspeaker layout (8 loudspeakers) using VBAP from a t-design layout. Afterwards the open-source MIT HRTF collection has been taken for convolving the resulting 8 signals with the corresponding HRTFs for those directions.

Hence, the procedure this time was to compare indirectly the interpolated sample with the reference one. For achieving this, several measured samples were picked and reproduced in triplets like if using 1D interpolation: first the left, if interpolation was carried out horizontally, or bottom, if vertically, reference signal was reproduced, then the interpolated or the measured one, and lastly the right or top reference signal. By reference signals it is meant the ones used for interpolation. Then, the testers are asked for evaluating if the central sample is perceived exactly in the middle of the reference ones and with good quality (timber) or not, again scoring form 1 (bad) to 5 (excellent).

This same procedure was repeated onto 6 different positions (i.e. 12 reproductions, as it must be carried out with the interpolated and the measured sample for each position). Samples were shuffled so, there is no biasing or direct comparison in the answers.

As it is too soon to have statistical meaningfully in the scores, the impressions so far will be discussed. Results revealed very good perception of the timber and fair localization. By comparing the interpolated sample directly with the measured localization seemed to displace a little to the left. This could be a particular effect of the measures in the room, so more rooms should be measured to discard any possible artifact. However, when reproducing several continuous interpolated samples around the space, which is more natural, this effect was reduced notably and were well located at the expected position. Thanks to the good preservation of the timber, sources were not perceived at all closer, further, widener or blurry, which in turn, helps to locate the sources. Localization is expected to be improved if a higher order of Ambisonics were employed. This is because, small mismatches in the estimated DOA provoke a diffuser localization in lowest order, where the virtual microphones polar patterns have low directivity. On the contrary, with higher orders, more directivity patterns are used and reconstruction will be perceived more clear even though it is not at the exact position.

Computational cost

Again, the algorithm has been tested using MATLAB. Several runs of the algorithm have been timed, giving an average timing of 71.93 ms for a 2D interpolation. Results are

promising as it only supposes a factor of 3 with respect to the BRIR algorithm, despite of dealing with 4 signals and doing some extra processing. This time, interpolation has not been tried on a real time interactive Matlab GUI, due to the impossibility to hear it smoothly in real time if binaural decoding were also to be applied in Matlab. Nevertheless, there is enough evidence to affirm that it would work using other language as C.

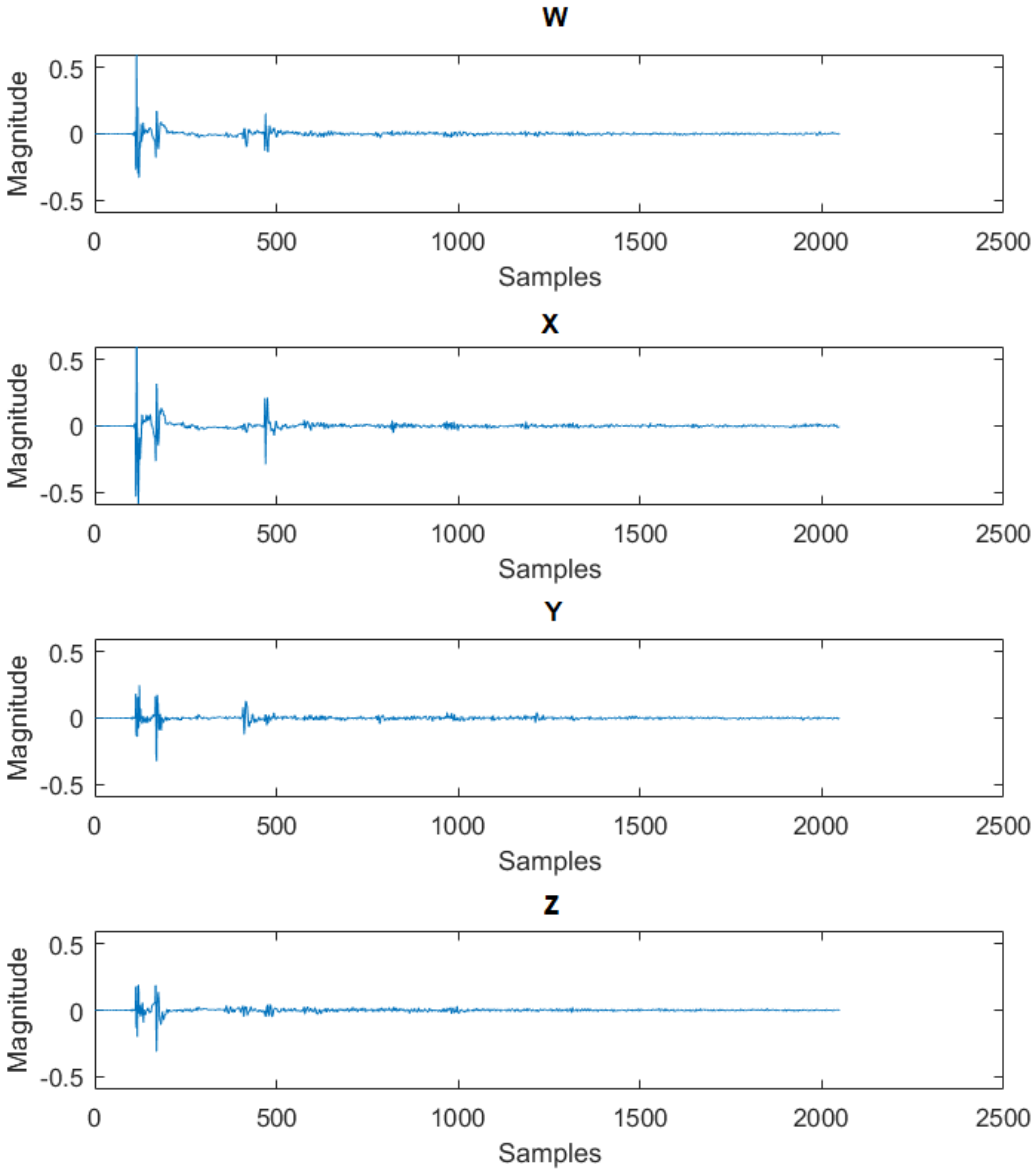

Figure 4.7: Measured FOA B-Format signals at the interpolation position

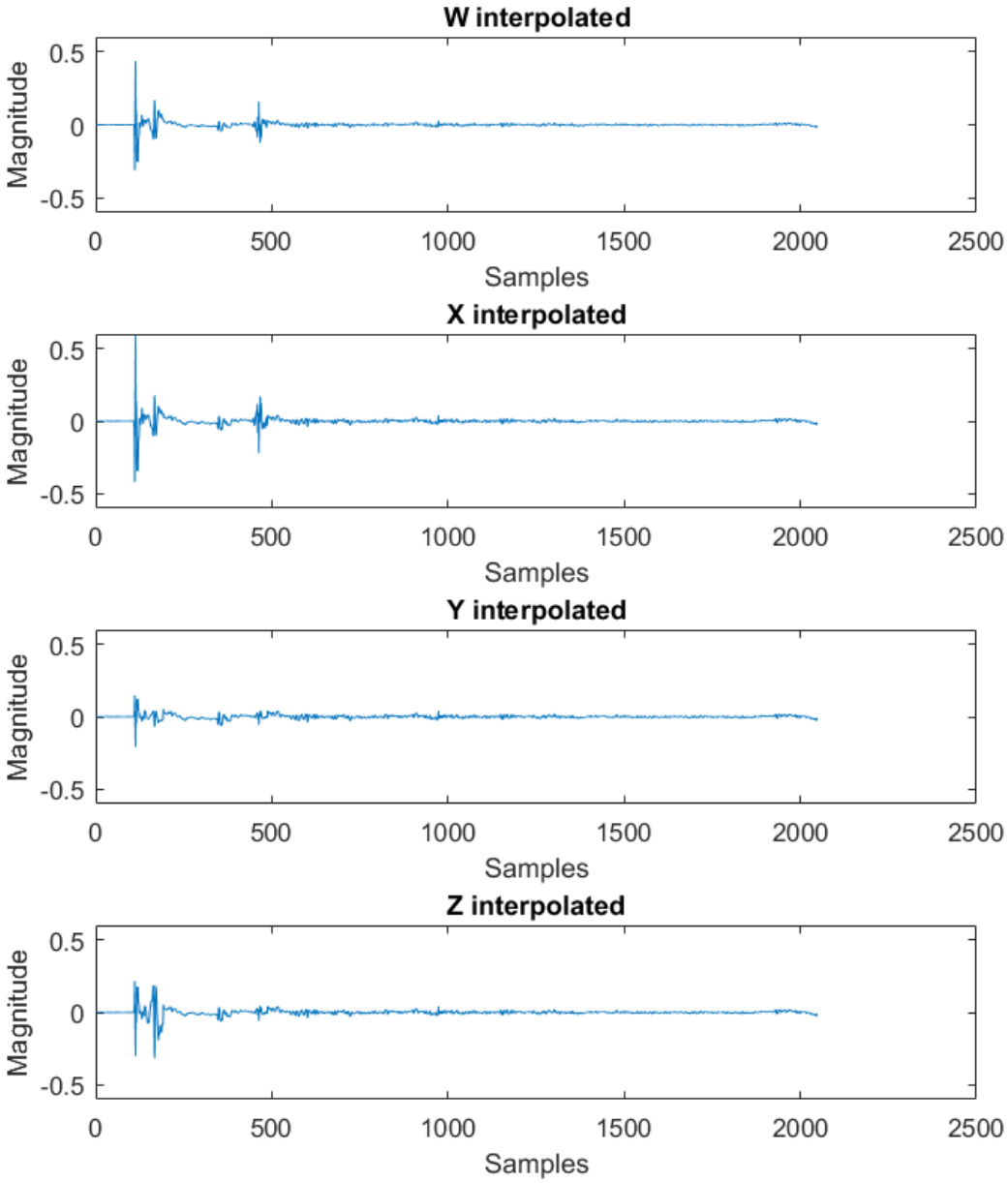

Figure 4.8: Interpolated FOA B-Format signals at the interpolation position

Chapter 5

Conclusions and Future Work

In this last chapter, we will sum up the achievements met according to the initial objectives. Afterwards, further lines of development will be proposed, followed by a personal reflection of the author and a special note on the publications.

5.1 Conclusions

At the beginning of this work, it was stated that the objectives were:

- To analyze the state-of-the-art techniques for spatial audio interpolation.
- To understand the underlying principles of the localization of sound sources for different recording and reproduction methods
- To learn new techniques which could be useful in the professional industry, regarding Binaural Room Impulse Responses and Ambisonics, and that could be easily deployed in a real system.
- To implement new interpolation algorithms, with Matlab, suitable for audio spatialization in real-time applications, evaluate and compare them with the existing techniques.

In this thesis, we have first stated the problem of interpolation in the context of audio spatialization and reviewed some of the state-of-the-art techniques, such as Dynamic Time Warping, pointing out their pros and their cons. This served as a basis to understand how interpolation worked and identify which parts of the signals were more relevant for interpolation: the main impulse and the early reflections.

With that knowledge, two novel algorithms for interpolation of Binaural Room Impulse Responses and First Order Ambisonics B-Format have been presented. These algorithms were devised with the idea to work in real-time, so as to meet with the needs of the industry and ease their deployment in real systems.

Given a measured series of BRIRs of Ambisonic records taken at several points in a room at any position, the algorithms are capable of finding the impulse responses at any point between them, by means of a 2D interpolation.

The first algorithm takes as inputs four Binaural Room Impulse Responses and the desired position where to interpolate and outputs the interpolated BRIR at that position. This algorithm is based on the assumption that only certain parts of the signal contains information about localization, thus focusing the processing on these. Moreover, this algorithm is provided with great flexibility in many critical points as peak matching, signal stretching and interpolation of the reflections. This has been achieved by means of a set of equations (Gravity Points, displacement function, and peak matching matrix) that can be adjusted to satisfy one's goals or to better adjust to a concrete measures. Results of a preliminary subjective test campaign denoted that this algorithm outperforms both in quality and in computational cost the state-of-the-art algorithm, the Dynamic Time Warping. Localization was accurate and clear.

The second algorithm works with First Order Ambisonic B-Format signals and also showed good performance and low computational cost, despite of not having any other algorithm as reference. This algorithm takes as inputs four B-Format signals (16 signals in total) and returns the interpolated signals at the interpolation position. Again, it works under the same assumption as the BRIR algorithm, and contains several parameters that are configurable in function of the room and measurements and that can be adjustable as a function of the requirements, providing the algorithm with high flexibility. Results showed that localization could be improved, as first order provides poor accuracy and made the sources seem widener, but the quality of the reproduction was not harmed.

Both systems have been implemented in Matlab, obtaining a gain of x7 in computational cost for the BRIR algorithm with respect to DTW. Despite of not being implemented in C language, it has been proven that it could be used in real-time applications such as virtual reality, videoconferencing and gaming. Ambisonics algorithm achieved very good timing considering that deals with more signals and performs extra processing steps such as DOA estimation and interpolation.

Being this said, it is fair to affirm that the objectives have been successfully accomplished.

5.2 Future work

As future work, there are several tasks that should be carried out in order to refine or improve the algorithms and/or add more features. Therefore it remains to:

- Make a comprehensive subjective test campaign that, in turn, allow to further refine the parameters and lay down the benefits of the algorithms.
- To test the algorithms with more data from other rooms to make sure that regardless the nature of the data, they work properly. Different rooms affect the shape of the response and it could make the peak matching fail, for instance.
- To extend the algorithms and validate to 3D. For this, measures would need to be taken in the Z axis, and perform 7 linear interpolations.
- Implement the algorithms in a compiled language such as C, to have a real deployable version of the algorithms. Although it has already been proven that they work in real time, this last step is needed to really make them distributable and usable in the industry.
- Generalize the interpolation problem to the case where the listener moves and the source is fixed. As early commented in Chapter 1, this would give the degree of freedom required to be able to reconstruct whichever sound scenario in 3D.

5.3 Reflections of the author

Lastly, I would like to bring up all the bulk work that could have not been reflected but have been required to achieve the results of this thesis. These include taking lots of measurements, countless hours programming tests to validate the algorithms so as to provide them of the rigor needed in official publications, debugging systems (really important this one, otherwise it would be impossible to find a fail within thousands of lines of code), adding features for supporting popular audio formats in the industry (such as SOFA), programming GUIs for evaluation of the algorithms, checking all the state-of-the-art software that could do something similar, reading and documentation of everything developed. As a result of all of these, a comprehensive amount of functions related to audio spatialization processing have been generated and documented, which can serve as a base for other people that could find them useful in their research. Moreover, they have been programmed both in object oriented programming (OOP) and functions so they are easy to use and provide a compact way of working with them. In fact, even mates in the office have started to use them for various purposes.

But, in fact, is all that work, including the one that did not get anywhere, to which I am really grateful as it served me to learn tons of new things and made me aware of the difficulty and dedication that research demands. Because, even the minimum thing which at prior seemed easy and fast to do, required a huge effort in the end. Because trying every option and keep failing stepped me towards the right solution with more confidence and criteria.

Looking backwards, I would have never imagined to be that immersed in the audio processing and almost catch up to the edge of innovation (because one never reaches it), and even less to contribute to it with publications. Also, I have been pretty lucky to be able to learn from very wise people who offered me their many years of experience condensed in few words. Personally, I really enjoyed this process of creation, which has been sort of artistic, as inspiration came when you less expected it and constancy was your unique ally when it did not show up. However, it is very encouraging to think that maybe one day, there will be people who find useful your contribution and who could build something meaningful on its basis.

Indeed, this is what the endless cycle of research is about: learning, experimenting, sometimes fail, other success, recap and repeat..

5.4 Publications

Finally, remark that the results of this work, allowed my tutor and I to make publications for both of the algorithms. The first one, was already presented as a congress paper for the Audio Engineering Society in Milan:

Victor Garcia-Gomez and Jose J.Lopez. ***Binaural Room Responses Interpolation for Multimedia Real-Time Applications*****. In proc. of Audio Engineering Society Convention, Milan (Italy), May 2018**

This paper covers the first algorithm explained in Chapter 3 along with its results. I had the opportunity to travel and assist to the Audio Engineering Society (AES) congress in Milan and present the paper myself the 24th of May, in an oral presentation of half an hour. It was quite good received by the audience, who seemed interested and asked for more details.

The second one, will be finished by mid of July and sent as a journal article to the AES Journal:

Victor Garcia-Gomez and Jose J.Lopez. ***First Order Ambisonics B-Format Impulse Responses Interpolation for 3D audio Real-Time Applications*****. Journal of the Audio Engineering Society, (to be submitted) July 2018**

In this article we cover the second algorithm explained, together with some extra results that will be gathered in the following weeks time. We expect it to be also welcomed as Ambisonics is of great interest at the moment for everybody in the audio spatialization industry (VR, videogames, audio and video producers,etc).

Bibliography

[1] Gavin Kearney, Claire Masterson, Stephen Adams, and Frank Boland. Approximation of binaural room impulse responses. *IET Irish Signals and Systems Conference (ISSC)*, 2009.

[2] H L Han. Measuring a Dummy Head in Search of Pinna Cues. *AES Journal*, 1994.

[3] Emanuel Aguilera, Jose J. Lopez, and Jeremy R. Cooperstock. Spatial audio for audioconferencing in mobile devices: Investigating the importance of virtual mobility and private communication and optimizations. *AES: Journal of the Audio Engineering Society*, 2016.

[4] Christian Nachbar, Franz Zotter, Etienne Deleflie, and Alois Sontacchi. Ambix - A Suggested Ambisonics Format. *International Symposium on Ambisonics and Spherical Acoustics*, 2011.

[5] Aaron J Heller, Eric M. Benjamin, and Richard Lee. A Toolkit for the Design of Ambisonic Decoders. *Linux Audio Conference*, 2012.

[6] Yoichi Haneda, Yutaka Kaneda, and Nobuhiko Kitawaki. Common-Acoustical-Pole and Residue Model and Its Application to Spatial Interpolation and Extrapolation of a Room Transfer Function. *IEEE Trans. Speech and Audio Proc.*, 1999.

[7] Gavin Kearney, Claire Masterson, Stephen Adams, and Frank Boland. Dynamic Time Warping for Acoustic Response Interpolation : Possibilities and Limitations. *17th European Signal Processing Conf. (EUSIPCO 2009)*, 2009.

[8] Claire Masterson, Gavin Kearney, and Frank Boland. Acoustic Impulse Response Interpolation for Multichannel Systems Using Dynamic Time Warping. In *AES 35th Int. Conf.*, 2009.

[9] Klaus Hartung, Jonas Braasch, and Susanne J Sterbing. Comparison of Different Methods for the Interpolation of Head-Related Transfer Functions. *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*, 3 1999.

[10] Victor Garcia-Gomez and Jose J. Lopez. Binaural Room Impulse Responses Interpolation for Multimedia Real-Time Applications. *In proc. of AES Convention Milan (Italy)*, 2018.

[11] Juha Merimaa. Modification of HRTF Filters to Reduce Timbral Effects in Binaural Synthesis, Part 2: Individual HRTFs. *Proceedings of 129th AES Convention*, 2010.

[12] Sanjeev Mehrotra, Wei Ge Chen, and Zhengyou Zhang. Interpolation of combined head and room impulse response for audio spatialization. *MMSP 2011 - IEEE International Workshop on Multimedia Signal Processing*, 2011.

[13] Franz Zotter, Matthias Frank, and Hannes Pomberger. Comparison of energy-preserving and all-round Ambisonic decoders. *Proceedings of the German Annual Conference on Acoustics (DAGA)*, 2013.

[14] Aaron Heller, Richard Lee, and Eric Benjamin. Is My Decoder Ambisonic? *Proceedings of The Audio Engineering Society Convention*, 2008.

[15] Svein Berge and Natasha Barrett. A new method for B-format to binaural transcoding. *40th AES International conference.*, 2010.

[16] Archontis Politis and David Poirier-Quinot. JSAmbisonics : A Web Audio library for interactive spatial sound processing on the web. *Proceedings of the Interactive Audio Systems Symposium*, 2016.

[17] Sylvain Busson, Rozenn Nicol, Vincent Choqueuse, and Vincent Lemaire. Non-linear interpolation of Head Related Transfer Function. 2006.

[18] Ming XU, Zidao WANG, and Ying GAO. Interpolation of Minimum-Phase HRIRs Using RBF Artificial Neural Network, 2017.

[19] Juha Merimaa and Ville Pulkki. Spatial impulse response rendering I: Analysis and synthesis. *AES: Journal of the Audio Engineering Society*, 2005.

[20] Piotr Majdak, Peter Balazs, and Bernhard Laback. Convention Paper 7019 Multiple Exponential Sweep Method for Fast Functions. 2007.