Universidad de Málaga

Escuela Técnica Superior de Ingeniería de Telecomunicación
Programa de Doctorado en Ingeniería de Telecomunicación

DOCTORAL THESIS

3D Binaural Spatialisation for Virtual Reality and Psychoacoustics

Author:

María Cuevas Rodríguez

Supervisors:

Arcadio Reyes Lecuona

Luis Molina Tanco

July 2022

AUTOR: María Cuevas Rodríguez

https://orcid.org/0000-0002-4698-5170

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña MARIA CUEVAS RODRIGUEZ
Estudiante del programa de doctorado INGENIERÍA DE TELECOMUNICACIÓN de la Universidad de Málaga, autor/a de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada: 3D BINAURAL SPATIALISATION FOR VIRTUAL REALITY AND PSYCHOACOUSTICS

Realizada bajo la tutorización de ARCADIO REYES LECUONA y dirección de ARCADIO REYES LECUONA Y LUIS MOLINA TANCO (si tuviera varios directores deberá hacer constar el nombre de todos)

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 6 de JUNIO de 2022

Fdo.: MARIA CUEVAS RODRIGUEZ Doctorando/a	Fdo.: ARCADIO REYES LECUONA Tutor/a
Fdo.: ARCADIO REYES LECUONA Y LUIS MOLINA TANCO (DIRECTORES DE TESIS)	

Edificio Pabellón de Gobierno. Campus El Ejido. 29071
Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10
E-mail: doctorado@uma.es

To my family.

A mi familia.

Acknowledgements

Firstly, I would like to express my deep and sincere gratitude to my supervisors, Arcadio Reyes Lecuona and Luis Molina Tanco, for the continuous support of my PhD study and related research, for their guidance and motivation. On many occasions we have worked side by side, in which they have shared their knowledge with me and helped provide great experiences.

A very special thanks goes out to Daniel Gonzalez; I couldn't imagine a better colleague for this doctoral thesis journey. For his motivation, encouragement, technical support, and for the many hours we were working together before deadlines. I would also like to thank the rest of my laboratory mates for the stimulating discussions and all the fun we have had in the last years.

I wish to thank the people I had the great opportunity to meet in the different European projects I have been involved in. Special thanks to Lorenzo Picinali, for his hospitality, his positive attitude, support, and his technical assistance.

I would like to strongly thank my family for supporting me through my entire life, and in particular, my parents, Andrés and María, they have always believed in me, encouraged me and supported me in all the decisions I have taken.

Finally, my deepest gratitude goes to my husband Alex, without whose love, continuing support, encouragement and assistance, I would not have finished this thesis. And to my little daughter. Alma, you are only 10 months old, but you have arrived to give me the strength I needed to finish this work and to brighten our days.

Index

Abstract

The incorporation of spatial audio to the simulation of immersive Virtual Reality (VR) environments is becoming essential. The aim of spatialised audio is to create in the listener the illusion of sound sources existing in three-dimensional space, increasing the realism of the VR environment. In the real world, the shape of the listener's head and pinna, together with reflections from the shoulders and the shadow produced by the head, act together as a filter that modifies the sound before it reaches the eardrum. This sound is interpreted by the brain to localise its position. The use of headphones in an immersive VR environment destroys this natural filter, so the sound is perceived as if it was inside the head. To overcome this limitation, we endeavour to simulate the natural filtering process by adding a series of cues to the original audio signal that can be interpreted by the brain for the spatial localisation of the sound source. Many of these cues are captured in a filter to characterise the listener, mathematically represented by the Head-Related Transfer Function (HRTF), which is unique for each person. If the simulation of the sound sources happens in an enclosed environment, the filter is the BRIR (Binaural Room Impulse Response) which characterises the listener and the environment.

This thesis describes a software library, the 3DTI Toolkit-BS. Its main task is to process an auditory signal based on its position within the VR environment and the characteristics of the listener and the environment. The 3DTI Toolkit-BS library offers a set of algorithms to simulate sources at different distances (including very close and very far distances), to customise the simulation for each listener (by means of HRTF interpolation and convolution) and to simulate sounds whose sources are in enclosed environments (by means of Ambisonics and convolution with BRIRs). It is an open source and multiplatform library developed in C++. The library is implemented with a flexible and modular structure, allowing new rendering methods to be integrated or the replacement of any module by others developed in the future.

Current immersive VR environments are dynamic and interactive. The user is constantly moving and interacting with time-varying elements in the environment. In

terms of audio, this means that the relative position between the sound sources and the listener are constantly changing. The 3DTI Toolkit-BS is designed to manage complex, dynamic acoustic scenes that change in real time. Smooth transitions for moving sources and/or listener were developed to avoid audible artefacts. The library has been evaluated with a battery of tests demonstrating its good dynamic behavior and performance.

The applications of the 3DTI Toolkit-BS library are not limited to Virtual Reality; the library aims to become a reference tool for the execution of psychoacoustics experiments, as it brings together in a single open-source tool several techniques and functionalities developed and evaluated of spatial audio research in the last 20 years. The 3DTI Toolkit-BS tool was tested in a psychoacoustics experiment, a study on the influence of non-individual HRTF on speech intelligibility. It is known that HRTF signals have an impact on speech intelligibility. However, how these cues affect each individual and, more specifically, the impact of HRTF choice in a Cocktail Party scenario (scenario where the listener tries to focus attention on a particular acoustic stimulus, filtering out and eliminating all other stimuli) has not been investigated in depth yet. In the experiment, the Speech Reception Threshold (SRT) was measured, showing significant global and individual differences between the SRTs measured using different HRTFs. These results confirmed that for these Cocktail Party situations, the choice of HRTF should be carefully considered for each individual.

Resumen

Los entornos de Realidad Virtual (RV) inmersivos son aquellos espacios donde la realidad física del usuario es reemplazada por un entorno artificial. En estos entornos, a pesar de que la simulación de la parte visual sigue siendo dominante, el audio está cobrando cada vez más importancia. La simulación de audio espacializado binaural pretende conseguir que, a través del uso de auriculares, una fuente sonora dentro de un entorno de RV suene como un sonido real y que el oyente pueda ubicarla dentro del espacio tridimensional. Esto mejora la sensación de presencia y realismo del usuario en entornos RV inmersivos.

En el mundo real, la forma de la cabeza y del pabellón auditivo del oyente, junto con las reflectiones en los hombros y la sombra que produce la cabeza, actúan como un filtro que modifica el sonido antes de alcanzar el tímpano. Este sonido es interpretado por el cerebro para localizar su posición. En el momento en que utilizamos auriculares en un entorno de RV inmersivo, perdemos esta ventaja y el sonido pasa a escucharse como si estuviera dentro de nuestra cabeza. Para evitar esta limitación, es necesario simular el proceso natural de filtrado. Este reto se consigue añadiéndole a la señal de audio una serie de indicios que puedan ser interpretados por nuestro cerebro para la localización espacial de la fuente de sonido. Para ello, el filtro que caracteriza al oyente se representa matemáticamente con una función de transferencia relacionada con la cabeza (HRTF), la cual es única para cada persona y caracteriza múltiples puntos del espacio tridimensional. Si el sonido se encuentra en un entorno cerrado hablaríamos de una transferencia que recoge características del oyente, pero también del entorno (BRIR).

Esta tesis doctoral presenta un conjunto de herramientas software que componen una librería llamada 3DTI Toolkit-BS, cuya principal tarea es procesar una señal auditiva, acorde a la posición de esta dentro del entorno virtual y de las características del oyente y el entorno. La librería 3DTI Toolkit-BS ofrece una serie de algoritmos que permiten simular fuentes a diferentes distancias (incluyendo distancias muy cercanas y muy lejanas), realizar la simulación de forma personalizada para cada oyente (mediante la interpolación y convolución del HRTF) y simular sonidos cuyas fuentes se encuentran en entornos cerrados (basada en una aproximación Ambisónica y convolución con BRIRs). El conjunto de herramientas que componen la librería se ha desarrollado en

C++, es de código abierto y multiplataforma. La librería ha sido implementada con una estructura flexible y modular, que permite el uso de cada componente de forma independiente, así como la integración de nuevos métodos de renderizado o la sustitución de unos módulos por otros.

Los entornos de RV inmersivos suelen ser entornos dinámicos, variantes en el tiempo, donde el usuario se encuentra en constante movimiento e interactúa con los elementos del espacio. Esto hace que la posición relativa entre las fuentes sonoras y el oyente cambien constantemente. Para ello, el 3DTI Toolkit-BS permite crear escenas acústicas complejas en tiempo real, teniendo en cuenta en todo momento la posición de las fuentes y del oyente dentro del espacio virtual. Se ha prestado especial atención al desarrollo de algoritmos que permitan transiciones suaves cuando una fuente o el oyente están en movimiento, evitando que se produzcan artefactos audibles, que enturbien la experiencia del usuario. La librería ha sido evaluada con una batería de pruebas donde se demuestra su buen comportamiento dinámico y rendimiento.

Las aplicaciones de la librería 3DTI Toolkit-BS no se limitan a la Relidad Virtual; la librería tiene como objetivo convertirse en una herramienta de referencia para la ejecución de experimentos de psicoacoústica, ya que agrupa en una sola herramienta de código abierto varias técnicas y funcionalidades desarrolladas y evaluadas en los últimos 20 años de investigación sobre el audio espacial. Para testear el uso del 3DTI Toolkit-BS en un experimento de psicoacoústica, se ha llevado a cabo un estudio sobre la influencia de la HRTF no individual en la inteligibilidad del habla. Se sabe que las señales de la HRTF tienen un impacto en la inteligibilidad del habla. Sin embargo, aún no se ha investigado en profundidad cómo afectan estas señales a cada individuo y, más concretamente, el impacto de la elección de la HRTF en un escenario de C*ocktail Party* (escenario donde el oyente trata de enfocar la atención en un estímulo acústico en particular, filtrando y eliminando el resto de los estímulos). En el experimento, se midió el umbral de recepción del habla (SRT), mostrando diferencias globales e individuales significativas entre los SRTs medidos utilizando diferentes HRTFs. Estos resultados confirmaron que para estas situaciones de *Cocktail Party*, la elección de la HRTF para cada individuo debe ser considerada cuidadosamente.

Acronyms

3DTI	3D Tune-In
ANOVA	Analysis of variance
BEM	Boundary Element Method
BILD	Binaural Intelligibility Level Difference
BMLD	Binaural Masking Level Difference
BRIR	Binaural Room Impulse Response
DSP	Digital Signal Processing
DT	Detection Threshold
EC	Equalization Cancelation
EoB	Energy out of Band
FEM	Fast Multipole Method
FFT	Fast Fourier Transform
GPU	Graphics Processing Unit
HATO	Head-Above-Torso
HMD	Head Mounted Display
HpEq	Headphone Equalization
HpTF	Headphone Transfer Function
HOA	Higher-order Ambisonics
HRIR	Head-Related Impulse Response
HRTF	Head-Related Transfer Function
IACC	Interaural Cross-Correlation
IHL	Inside-the-Head Localization
ILD	Interaural Level Difference
ITD	Interaural Time Difference
KEMAR	Knowles Electronics Manikin for Acoustic Research
MC	Masker Configurations
RIR	Room Impulse Response
SOFA	Spatially Oriented Format for Acoustics
SNR	Signal-to-Noise Ratio

SRM	Spatial Release from Masking
SPK	Speaker
STR	Speech Reception Threshold
Toolkit-BS	Toolkit Binaural Spatialiser
UPOLS	Uniformly Partition Overlap-Save
VAS	Virtual Auditory Space
VBAP	Vector-based amplitude panning
VE	Virtual Environments
VR	Virtual Reality
VST	Virtual Studio Technology
WFS	Wavefield Synthesis

Chapter 1

Introduction

This opening chapter aims to introduce the basic concepts of spatial hearing and how these concepts are used to simulate 3D binaural spatial audio (Sections 1.1 and 1.2). Then, Section 1.3 describes the context and the motivation of this PhD thesis. Finally, the research objectives and the structure of the document are outlined in Sections 1.4 and 1.5. respectively.

1.1 Spatial hearing

A fundamental function of the auditory system is the *perception of spatial sound*: the cognitive process that allows us to identify the location of any sound in space. Sound spatialisation plays an important role in daily life, including functions such as spatial awareness, object localization and avoidance, or the ability to focus on one speaker when multiple speakers are participating simultaneously in a conversation. Perceiving spatial sound is something natural and contributes significantly to the feeling of physical immersion.

The auditory system allows us to perceive sounds coming from our front, sides and back, at any elevation. This can be explained by reference to our anatomy, the physiology of the auditory system, the cognitive processes of the central nervous system and the characteristics of the environment. Having two ears (*binaural hearing*) allows us to *lateralize* sound very easily: if we perceive a sound reaching our right ear before reaching the left one, our brain interprets that this source is in our right side. Our head, neck, torso and the outer ear filter the sound before it reaches our eardrums, and they do it differently when a sound comes from the front, than when it comes from the back.

These modifications are captured by the auditory system and interpreted by the auditory cortex to estimate the direction of arrival of the source of the sound. The environment also modifies the sound arriving to the listener, and these modifications will also be interpreted by the auditory cortex to extract information about the environment and the location of the sources of sound within the environment.

To go deeper into the concepts of sound spatialisation it is necessary to understand how the human auditory system works, and which are the auditory cues that make our brain perceive the location of a sound source.

1.1.1 The auditory system

Understanding the anatomy of the auditory system helps understanding its function. This section presents its physical structure and introduces the physiological and cognitive processes that take place in it, which will be described in more detail in Chapter 2, Section 2.5.1 Auditory signal processing.

The auditory system can be divided into two parts: *mechanical* and *neural*. Figure 1 shows these parts and the *auditory pathway* through which the information in the sound reaches the primary auditory cortex in the brain.

The *mechanical part* is divided into the outer, middle and inner ear:

- The **outer ear** is the *pinna* (or auricle) and the *auditory canal*. The sound enters the ear, which modifies the spectrum of the sound in a way that is dependent on the direction of the sound. The modified signal is then transmitted by the auditory canal to the eardrum. The auditory canal also modifies the sound signal, as it resonates around frequencies relevant to speech.
- The **middle ear** consists of the *eardrum* (or tympanic membrane) and three small bones, the chained *ossicles*: malleus, incus and stapes. The eardrum is a thin membrane that moves due to the variations of the air pressure in the auditory canal. These movements are transmitted to the inner ear by the ossicles, which transform the variations of pressure into mechanical movement. This part of the ear ultimately amplifies the sound and transfers it from the air-filled external ear to the liquid-filled cochlea.
- The **inner ear** is composed of the *cochlea* and the *auditory nerve*. One of the most relevant parts of the cochlea is the *basilar membrane*, where the *organ of Corti* rests. The basilar membrane presents the first level of frequency analysis in the cochlea. This membrane is a single structure that varies in mass and stiffness along its length. Each point of the membrane has a different resonance frequency

to which it responds maximally. The resonance frequency range available on the basilar membrane determines the frequency response of the human ear (20 – 20000 Hz). The basilar membrane movements activate the sensory *hair cells* contained in the organ of Corti, which elicit neural activation patterns in response to the movements. In the hair cells the mechanical sound signal is finally converted into electrical nerve signals, which are transmitted through the auditory nerve to the *neural part*.

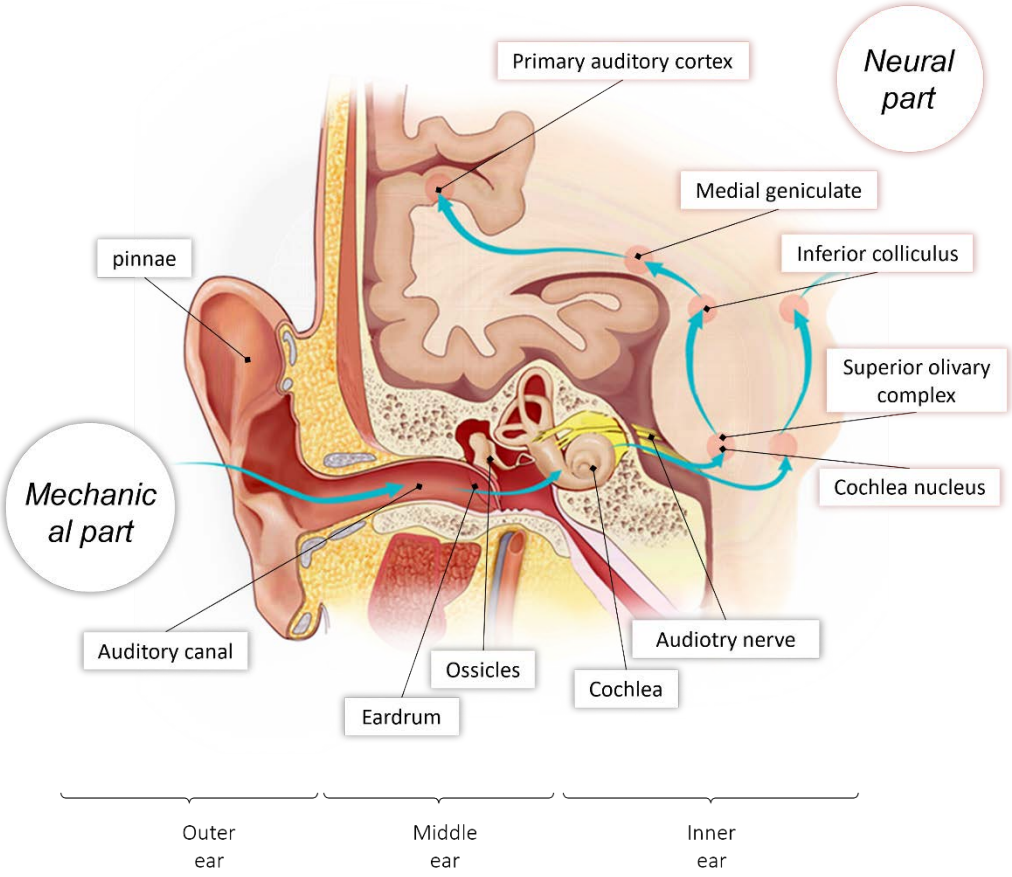


Figure 1. Anatomy of the human ear and auditory pathway[1].

The second part of the auditory system, the *neural part*, is the part of the nervous system responsible for audition. Neurons in the cochlea nerve carry information encoded in electrochemical signals from the inner ear to the central nervous system. The signal ascends in the auditory pathway towards the primary auditory cortex, getting decoded along the different stages in the pathway: The *cochlea nucleus* first decodes basic signal characteristics such as duration, intensity and frequency. The information travels to the

[1] Original image from Wikipedia (Public domain US government, https://commons.wikimedia.org/wiki/File:Hearing_mechanics_cropped_-_Acoustic_radiation.jpg), modified to show the names of the different elements in tags added later.

superior olivary complex, which integrates information from both ears, playing an essential role in the localisation of sound sources. This complex is believed to measure, among other things, the difference in level and time of arrival of the signals from both ears, which is considered as a major cue for estimating the lateralization of a sound source (localization cues will be described in detail in the next section). The signal continues its journey to the *inferior colliculus*, where the major ascending auditory pathways converge. This part is believed to be involved in the integration and routing of multi-modal sensory perception. Finally, the *medial geniculate* is part of the auditory thalamus, functioning as a relay station in the central auditory pathway, which receives the information from inferior colliculus and ultimately reaches the primary auditory cortex in the temporal lobe. The whole *auditory cortex* is where the conscious perception and the voluntary motricity response take place, recognizing, memorizing and processing the previously decoded signals (Pujol, 2020).

1.1.2 Sound localization cues

The human auditory sense can localize sound sources in the surrounding environment thanks to several *localization cues* embedded in the sound arriving at the two ears. This section presents a brief description of these localization cues. Later, in Section 1.2.2, more details about these cues are presented. Section 2.1 presents a technical overview of background research in this topic.

The localization cues are known as:

- *ITD (Interaural Time Difference),* which is the difference in the arrival time of the sound signal at the two ears.
- *ILD (Interaural Level Difference)*, which represents the differences in sound level at the two ears.
- *Monaural spectral cues*, which correspond to each ear separately and arise from the direction-dependent filtering of the listener head, torso, pinna and ear canal.

1.1.2.1 Binaural cues

The two *interaural differences*, ITD and ILD, are known as the *binaural cues* since they are calculated using signals arriving at both ears. ITDs (Figure 2) are caused by the difference in time between the arrival of sound at each of the two ears. When a source is on one side of the listener, it arrives first to the *ipsilateral ear* (the ear closer to the sound source) and then to the *contralateral ear* (the ear farther to the sound

source). This delay is noticed and interpreted by the brain as being caused by the sound coming from one side of the head. Something similar happens with the amplitude differences of signals at both ears, represented by the ILD. In this case, the signal arrives with more amplitude to the ipsilateral than to the contralateral ear. The attenuation is mainly caused by the shadowing effect of the head, as illustrated in Figure 2. ILDs are frequency-dependent, being notably larger for higher frequencies (when the wavelength is small compared with the size of the head) than for low frequencies. There is also a small frequency-dependency of ITDs, but this effect is still considered irrelevant from a perceptual point of view (Benichoux et al., 2016).

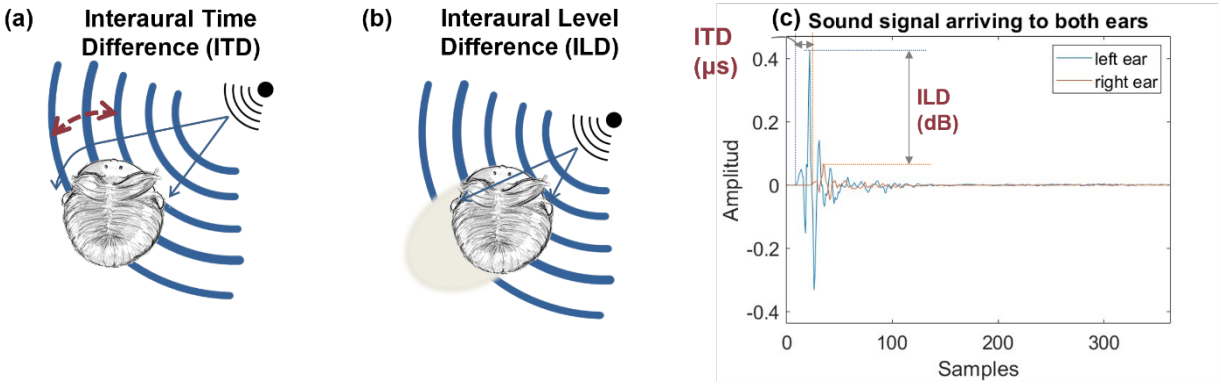


Figure 2. Graphical representation of the ITD (a) and ILD (b). Sub-figure (c) shows the signals arriving to the left and right ears with different in time of arrival and level.

These binaural cues allow the spatialisation mainly at the horizontal plane, where the accuracy of the human localizing sound is better that in the median plane. However, the accuracy in the perception of sound localization caused by the binaural cues is compromised by an effect called the *cone of confusion* effect (see Figure 3). These cones are regions centred on the *interaural axis*. Any sound coming from points on the surface of a given cone will result in approximately the same values for ITD and ILD. The brain is confused, as a source in the back hemisphere can be perceived to be in the front hemisphere, and vice versa.

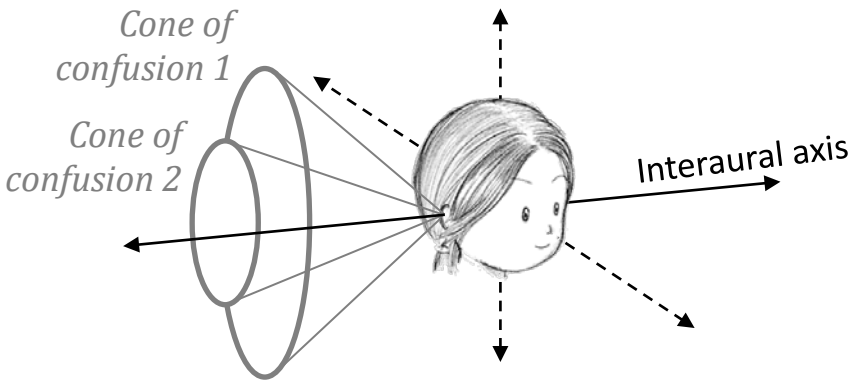


Figure 3. Two cones of confusion are represented.

1.1.2.2 Monaural cues

In addition to the interaural differences, humans also make use of *monaural cues*, i.e. not related with differences between the two ears. These cues arise from sound reflections on the listener itself. The wavelengths of audible sounds (2cm – 20m) are comparable to the dimensions of the human body (head and torso) and the outer ear (pinnae and auditory canal), which form a set of direction-dependent filters which result in modifications in the spectrum of the sound waves before they reach the eardrum.

Figure 4 shows a schematic representation of the filtering in the pinna. Due to the multiples cavities of this part of the auditory system, the sound is modified in different ways, depending on the location of the source. In Figure 4 it can be seen that for a frequency around 7 kHz, there is a *notch* when the sound comes from the front (top graph) that is not present when the sound comes from above (bottom graph). Although there is still much research to be done in this topic, it is well established that the notches that the pinna create in the signal spectrum provide the primary cues to perceive the height of a sound source. The frequency of these notches depends on the elevation of the source and varies greatly between individuals (Middlebrooks & Green, 1991).

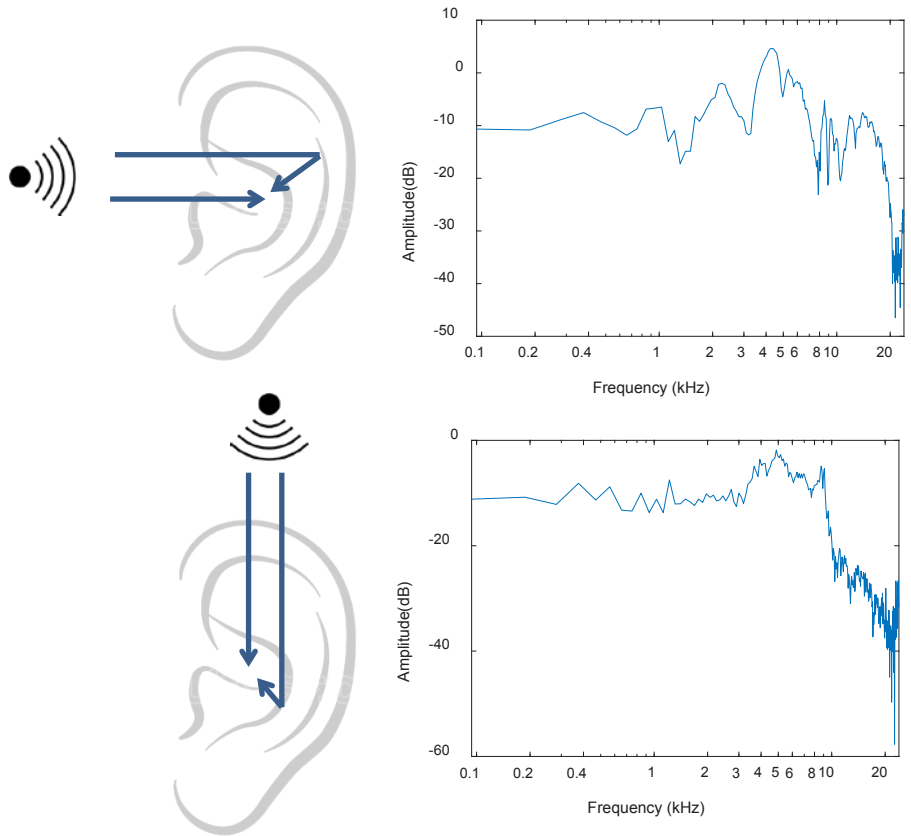


Figure 4. Schematical representation that shows measured frequency responses for two different directions of the sound source, from the front (first row) and from above (second row).

1.1.2.3 Auditory perception of distance

Until now, all presented cues refer to the direction of the source. However, there are a set of cues that make us *perceive the distance* of the source, such as loudness, reverberation and cognitive familiarity.

One of the main cues is the signal *loudness* (perceived level) and the fact that the direct sound level from the source to the listener decreases when the source distance increases. To use loudness as a cue, the brain considers the nature of the stimulus, for example, if it is a whisper or a scream. Therefore, it is considered that the combination of loudness and cognitive familiarity provides useful information about the distance of the sound source.

When the source is inside a room, the *reverberation* of the environment also becomes a cue that helps with distance perception. Inside a room, the sound is reflected many times on several surfaces, but the reverberant energy arriving to the listener does not change much with the distance between listener and source. In this way, the ratio of direct to reverberant sound changes as a function of distance between the sound source and the listener. As the listener moves away from a sound source, the level of the direct sound decreases, while the reverberation level remains invariant. This modification of the ratio is interpreted by the listener as a change in the distance to the source.

Binaural cues also play an important role in the perception of sound distance, when sources are located in the near proximity of the head. The level difference for nearby lateral sound sources (less than 1m from the head) is larger than when those sources are further away. These ILD variations contribute to the perception of a source as being very close to the head.

1.1.3 Beyond localization: the Cocktail Party phenomenon

Spatial hearing involves more concepts than just sound localization, especially when more than one sound source is involved. Using the localization cues introduced in the previous section, the listener can focus their attention on a specific sound, even when other sounds are being produced at the same time. This ability of the brain allows us to understand speech in a multi-talker situation, usually described as the *cocktail party effect* which first defined by (Cherry, 1953).

The auditory system allows listeners to focus their attention on a specific sound arriving from a specific direction (target sound) when one or more interfering sounds are also arriving to the listener (maskers). This ability is largely due to non-spatial attributes

such as sound signal intensity, pitch, timbre and rhythm. However, the auditory system can take the advantage of the spatial separation between the target and the maskers to detect the target sound more effectively. This phenomenon is called *spatial release from masking* or *spatial unmasking* and contributes to the ability of the brain to solve the cocktail party problem (Adelbert W Bronkhorst, 2000).

This psychoacoustic phenomenon is generally considered as the consequence of the cognitive processing of binaural sound information carried out at the neurological level of the auditory system (Haykin & Chen, 2005). However, many aspects are involved, and the understanding of how the human auditory system solves the cocktail party problem is still an open and very interesting line of research. As part of this PhD work, an experiment based on the spatial unmasking phenomenon has been carried out and will be described in detail in Chapter 5.

1.2 Simulation of spatial sound

An area where the spatial sound has found enormous and successful application is the field of Virtual Reality (VR) systems. Immersive VR systems have experimented a constant growth and popularization during the last decades, when most of the effort in research and development has been made on the visual modality. However, the real world is full of auditory stimuli and in our daily life we are constantly exposed to a three-dimensional experience of sound. It seems logical that spatial audio must be included in VR applications, for the sake of immersion and realism. Fortunately, the situation is changing, and interactive 3D audio is slowly becoming more and more present in VR, where the contribution of auditory stimulation in creating immersive experiences is becoming increasingly important.

A Virtual Auditory Space (VAS) is the name given to an artificial environment in which humans can perceive different virtual sounds as if they were real, located in specific points in space. When a sound is presented through headphones, the sound is directly injected in the ear canal. Therefore, if a mono-audio signal is reproduced, sounds are perceived by the listener as being inside their head. If a sound is presented using loudspeakers, it is perceived as coming from the physical position of the speaker. To achieve the *simulation of a spatial sound*, the sound signals must be processed before being delivered, to include the cues that lead to the perception of that location. The assumption is that, if these cues are simulated accurately, a listener immersed in a VAS will have the feeling that a virtual sound is coming from any place in the space, regardless the physical mode of delivery.

3D audio in Virtual Environments (VE) has become an important and strong line of research in recent years, where several approaches and innovative techniques that allow to include spatial audio in a VR system have emerged (which will be seen in detail in this chapter and the following). A sign of this can be seen in the fact that, while the number of publications on 3D audio in VE between years 2002 and 2011 was at 2502, in the last 10 years (between 2012 and 2021) these have increased to 5215[2]. 3D audio has also attracted attention of the major actors in the virtual reality industry, such as Google or Oculus (owned by Meta, formerly Facebook). In 2017, Google released an open source tool to include 3D audio in VE, called Resonance Audio (Google, n.d.). This tool is one of the most used renderers for both commercial and research applications in VR. Other tools, such as Oculus VR (2020), the Microsoft audio rendering engine (Microsoft, 2020), Steam Audio (Valve_Software, n.d.) or VRWorks Audio (Abhijit Patait, 2017), integrate 3D audio in VE but in this case they are closed-source and also some of them under a commercial license.

1.2.1 Overview of approaches to simulation

There are multiple approaches to simulate spatial audio, which can be classified according to the type of rendering algorithm, the delivery system they use and other criteria. In this section we will group the different approaches according to the properties under control during the audio rendering, which can be either perceptual or physical (Roginska & Geluso, 2017). The perceptual level methods take advantage of what it is known about the auditory perceptual system. They generate sound signals with features that, once processed by the perceptual system, provide the spatial feeling in the listener. Physical level systems try to generate sound signals that are physically similar to what is generated when the sound travels from the source to the listener.

1.2.1.1 Perceptual systems

In systems that use perceptual methods, the main goal is to *create perceived auditory events* that provide the listener with the feeling that the sound is coming from a specific location in space. These include stereo, surround and binaural systems.

[2] This data was obtained from Scopus platform (www.scopus.com), searching papers containing in the title, abstract or keywords the following words: (“audio” OR “sound” OR “auditory” OR “acoustic” OR “acoustics” OR “hearing”) AND (“Virtual Reality” OR “Augmented Reality” OR “Mixed Reality” OR “Extended Reality”).

Stereo systems

Stereo systems are designed to create the perception of sound coming from different directions on the horizontal plane. The sound is reproduced by using two audio channels and delivered by two loudspeakers or stereo headphones. In this way, the listener perceives the directional sounds arriving from a location between the two loudspeakers. One of the most used techniques is *stereo panning,* shown in Figure 5. The basic amplitude panning technique consists of two speakers (stereo) where the signal amplitudes of the left and right channels change to suggest a sound source (virtual source) that is localized on a two-dimensional sector defined by locations of the loudspeakers and the listener. This is the simplest approach for spatial sound, insufficient for front-back or out-of-plane localization.

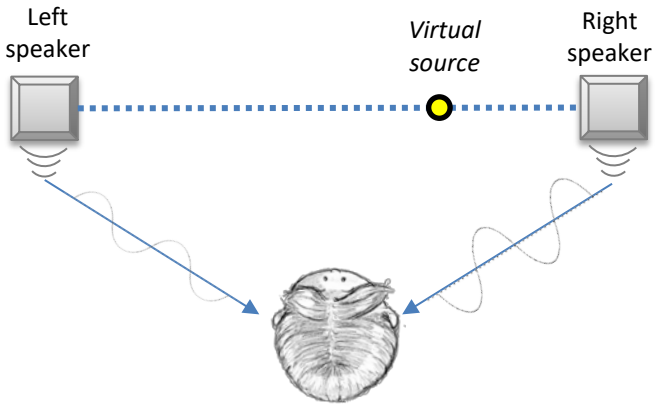


Figure 5. **Stereo panning**. Virtual sound can be simulated anywhere along the single line from left to the right by changing the level of the signal in both loudspeakers.

Surround systems

Surround sound is a term used to describe any configuration of loudspeaker reproduction system that includes more than two loudspeakers arranged around the listener (even in different planes) providing sound from multiple directions (Roginska & Geluso, 2017). A simple example can be seen in the *Vector-based amplitude panning* (VBAP) technique, which allows creating virtual sounds by panning among two or three dimensional distribution of an unlimited number of speakers, nearly equidistant from the listener (Pulkki, 1997). The most common surround sound system is the *ITU's 5.1 standard* (ITU-R, 1993), with 5 loudspeakers located at the center, front left, front right and two surrounds (left and right), as shown in Figure 6, and a subwoofer (the '.1') which corresponds with the low-frequency effects channel and for which position is not critical. This system has advantages compared to other surrounds systems: it does not require a large number of speakers. This makes this kind of systems useful for a number of real-world situations such as concerts, stage productions, installations in public places

and home theatres. However, similarly to the stereo system, this technique suffers a limitation known as the *sweet spot* area, which is based on the assumption that the position of the listener is known, fixed and restricted to a small area.

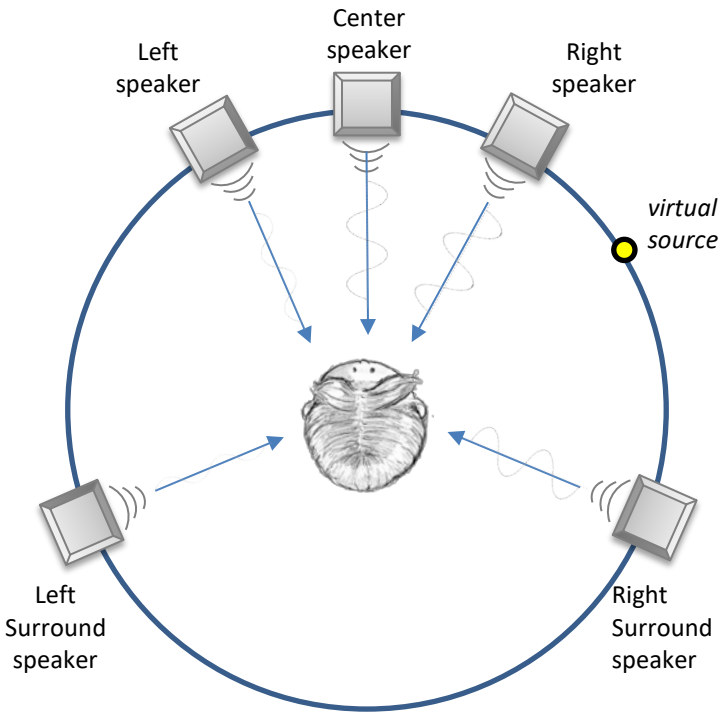


Figure 6. The 5.1 surround system. The distribution of the 5 speakers is: left and right speakers in front of the listener at ±30°, centre speaker in front of the listener at 0° and left surround and right surround speakers behind the listener at ±110°. The position of the '.1' speaker for low-frequency effects is not shown.

Binaural systems

Stereo and surround systems apply variations to the signal level. However, as seen in Section 1.1, our auditory system employs many other localization cues to perceive spatial sound. Using headphones, *binaural systems* deliver sound signals containing additional amplitude, frequency and phase modifications which simulate the changes that sounds experiment in the real world along the path to our eardrums. These effects create an illusion of spatial sound, a sound as real and natural as possible. The binaural, headphone-based approach is one of the main components of this thesis and will be described in detail in Section 1.2.

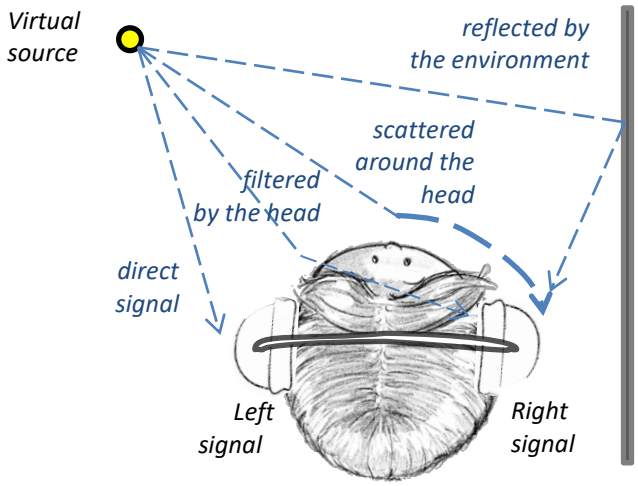


Figure 7. Headphone-based binaural system.

We speak of *transaural systems* when the binaural sound is delivered using loudspeakers (Roginska & Geluso, 2017). In this case, cross-talk cancelling filters are necessary to cancel the left-loudspeaker signal in the listener's right ear and vice versa (Cooper & Bauck, 1989). This technique requires the listener to stay in a small region, the previously mentioned *sweet spot* (Kyriakakis et al., 1999). In addition, the walls of the room where the system is working should be treated to avoid reflections that can modify the sound signal before it arrives to the listener's ears.

1.2.1.2 Sound field methods (physical systems)

Sound field methods attempt the direct capture and physical reconstruction of the sound waves (the sound field) that result from producing a sound in a specific real-world environment. The most widely used methods are *Ambisonics* and *wavefield synthesis*.

Ambisonics

Ambisonics is a technique based on spherical harmonic decomposition of sound, which allows recording, synthesis and playback of full-sphere surround sound (Gerzon, 1985). To synthesise spatial audio using Ambisonics, sound sources are encoded in an Ambisonics format and then decoded on a given set of loudspeakers, each one at a different location around the listener. The Ambisonic is generally expressed at a given order (e.g. first order Ambisonics, second order Ambisonics, etc), where higher orders correspond to an increased spatial accuracy. Figure 8 illustrates the spherical harmonics up to 3rd order.

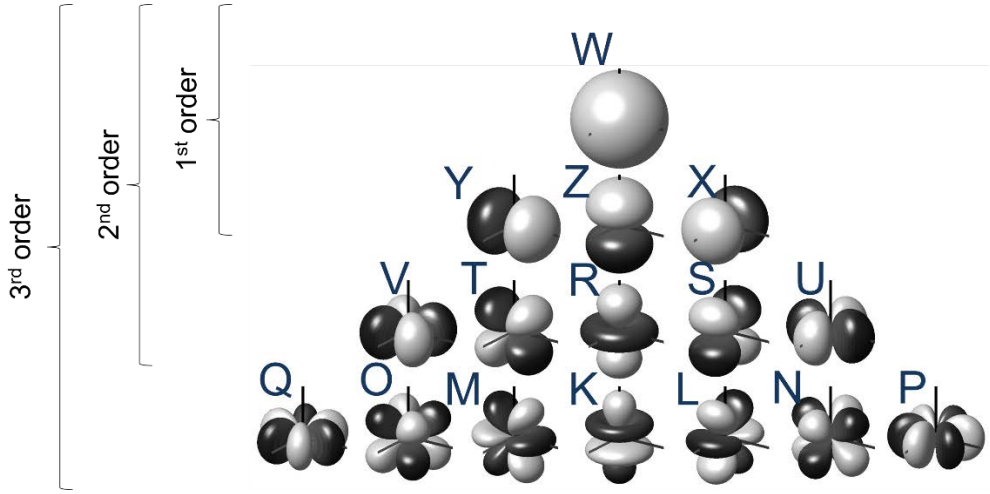


Figure 8. Visual representation of the Ambisonics components up to third order[3].

The simplest and most used Ambisonics format is called the B-format, which corresponds to 1st order Ambisonics channels W, X, Y and Z. The W channel is omnidirectional, Y represents the left-right directions, Z up-down and X front-back. A source signal S located at a specific direction (θ, ϕ), where θ is the polar azimuth and ϕ the polar elevation (see Figure 10), is distributed over the four channels using the following encoding functions:

$$W = S \cdot \frac{1}{\sqrt{2}}$$

$$X = S \cdot \cos\theta \cdot \cos\phi$$

$$Y = S \cdot \sin\theta \cos\phi$$

$$Z = S \cdot \sin\phi$$

The spatial resolution of 1st order Ambisonics is quite low with a small sweet spot. This resolution can be increased, and the sweet spot enlarged, incrementing the number of channels used to represent the sound. All formats with an order larger than one are collectively called *Higher-order Ambisonics (HOA)*. The codification becomes more complex as the HOA order increases, but the number of channels remains the same regardless of the number of spatialized sources.

[3] Image from Wikipedia, CC BY-SA 3.0, author: Dr Franz Zotter (https://en.wikipedia.org/wiki/Ambisonics#/media/File:Spherical_Harmonics_deg3.png). Modified to show the names of the Ambisonic channels.

Ambisonics assumes that the position of the listener is known, with fixed orientation and restricted to the sweet spot. However, Ambisonics allows for very efficient simulation rotations of the listeners head, since it can be implemented with a simple algebraic operation (see Zotter & Frank (2019) section 5.2.2).

The Ambisonics channels contain all the information of the sound field, but these channels cannot be sent directly to the reproduction system. A decodification must be carried out to convert the encoded signals to loudspeakers signals, which will depend on the configuration of the reproduction system, i.e. the number and position of the loudspeakers. Section 3.6 Reverberation path will show a specific case of Ambisonic decodification.

There are some techniques that combines the use of the Ambisonics and binaural sound synthesis. The most widely used technique is called the *Virtual-Ambisonics* (McKeag & McGrath, 1996). This technique consists in encoding the sound sources using the Ambisonic technique and then decoding them on a series of virtual loudspeakers in a specific position, then the virtual loudspeakers signals are rendered in the binaural domain. Virtual Ambisonics technique has been used in this PhD thesis and will be described in detail in Section 3.6. Recent studies have suggested an alternative formulation for Ambisonics and binaural playback that encodes the HRTF in the spherical harmonics domain in order to operate there with the Ambisonic channels directly (Engel et al., 2022).

Wavefield Synthesis

Wavefield Synthesis (WFS) relies on producing artificial wavefronts synthesized by a large number of loudspeakers. This technique is based on the Huygens principle (Jérôme Daniel et al., 2003), which proposed that a wave can be synthesized by adding the contributions of waves produced by a set of secondary sources positioned along the wave front. In this way, the reproduction system consists of a planar listening area using linear loudspeaker arrays, as shown in Figure 9. This kind of systems uses a high number of loudspeakers and solves, to some extent, the sweet spot problem presented by the Ambisonics, since it presents large listening areas (Spors et al., 2008). This technique is out of the scope of this PhD thesis and general overviews on WFS can be found in the literature (de Vries et al., 1994; Theile et al., 2003).

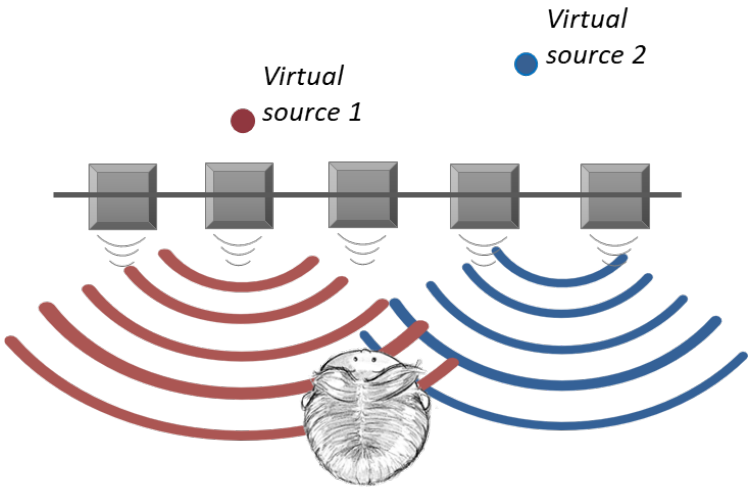


Figure 9. Diagram of a WFS system, where a lineal loudspeaker array synthetizes the spherical waves produced by two virtual sources at the wavefront.

1.2.2 Binaural systems

Binaural systems render auditory scenes by simulating all the localization cues believed to be used by humans to localize sound. In the following sections we review how this simulation is organized, by examining the role of each of the basic blocks of a binaural system: source (sound), medium (space or environment) and receiver (listener).

1.2.2.1 Coordinate system

Previously to revising how a binaural system works to simulate the position of a source, it is useful to define this position in a coordinate system which is adequate to how the brain performs this task: the *vertical-polar coordinate system*, which gives position in terms of radial *distance*, *azimuth* and *elevation* (see Figure 10a). The origin of the coordinate system is situated at the listener's head center, defined as the midpoint of the line segment that connects the two ears. The azimuth (θ) of the sound source is defined as the angle between the direction that the listener is facing and the horizontal projection of the line connecting listener and source, being positive towards the left, with $0° \leq \theta \leq 360°$. The elevation (ϕ) is the angle between the horizontal plane and the line between listener and source. It is positive in the north hemisphere and negative in the south hemisphere, with $-90° \leq \phi \leq 90°$. Finally, the distance is calculated with respect to origin, with $0 \leq d \leq \infty$.

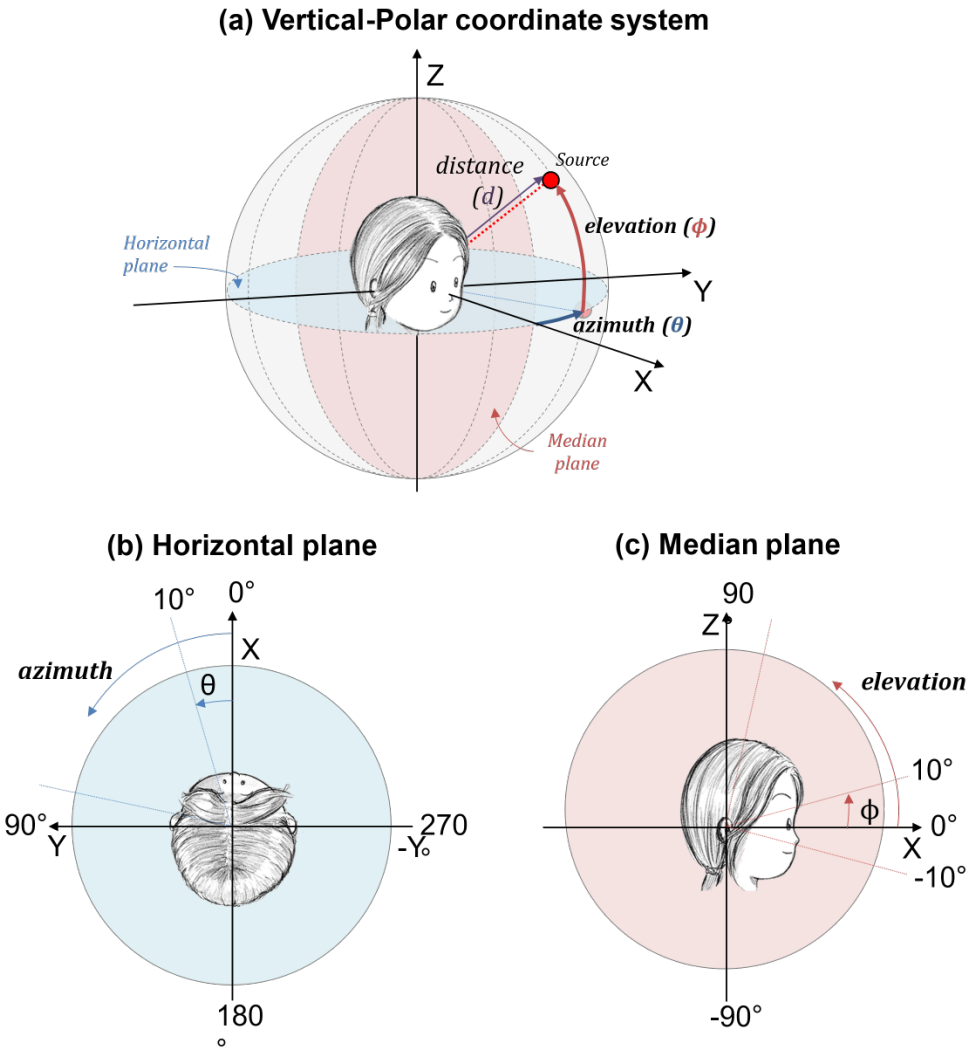


Figure 10. Vertical-polar Coordinate System. The position of a sound source can be defined in a spherical coordinate system centred at the listener's head. The azimuth is the angle on the horizontal plane that includes the listener and the source, is positive going counterclockwise, and is given in degrees. The elevation, also in degrees, is the angle on the vertical plane that includes the listener and the source and is positive going upwards. The radial distance is given in meters.

When Cartesian coordinates are used, the Z will be pointing up, Y will point towards the left ear, and X towards the front of the listener. We will often refer to specific planes, shown in Figure 10: the *horizontal plane* and the *median plane*. The horizontal plane is the XY plane. Sound sources on this plane have 0° elevation. The median plane is the XZ plane, and sound sources on this plane have 0° azimuth.

Additionally, two terms that will be frequently used are: ipsilateral (same side) ear and contralateral (opposite side) ear, to refer to a specific ear with respect to the sound source position. In this way, for the presented coordinate system, when the source direction is between (0°, 180°) the ipsilateral ear corresponds with the left ear and the

contralateral with the right ear. When the source direction is between (180°, 360°) the ipsilateral ear corresponds with the right ear and the contralateral with the left ear.

Sometimes a different coordinate system is adopted to describe the position of the source, which is known as *interaural coordinate system*. This coordinate system is shown in Figure 11, where the sound source position is described by the interaural azimuth (θ_I), the interaural elevation (ϕ_I) and the distance (d). The interaural azimuth is defined as the angle between the median plane and the directional vector of the sound source, with $-90° \leq \theta_I \leq 90°$. The interaural elevation is the angle between the projection of the directional vector to the median plane and the interaural axis (Y axis), with $0° \leq \phi_I \leq 360°$. The distance is defined as in the vertical-polar coordinate system, with respect to the origin ($0 \leq d \leq \infty$).

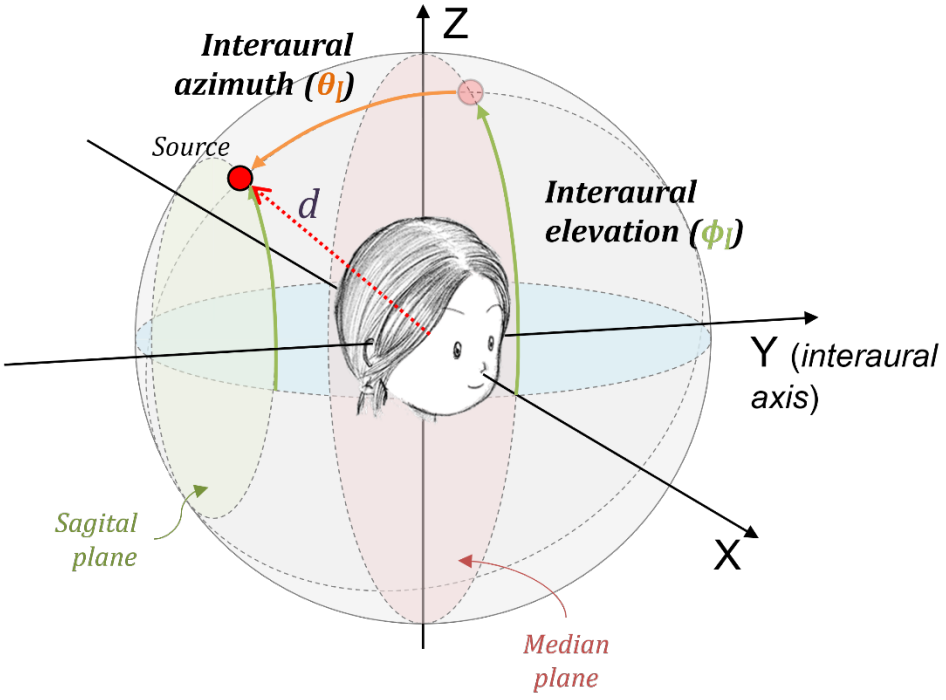


Figure 11. Interaural Coordinate System. The sound source is described by (θ_I, ϕ_I, d), as the interaural azimuth, interaural elevation and distance.

The vertical-polar coordinate system is adopted as the default in this thesis, while the interaural coordinate system will be occasionally used but it will be specified.

1.2.2.2 The Listener and the HRTF

Consider a free-field situation, where no reflections take place. The synthesis of VASs consists in rendering the binaural and monoaural cues presented in Section 1.1.2. These cues, which depend on the source location, are created by the transformations to the

sound in the paths to the listener's eardrums. We can model these transformations as a digital system characterized by its impulse response: the well known *Head-Related Impulse Response (HRIR)*, or alternatively its Fourier transform, the *Head-Related Transfer Function (HRTF)*.

An HRTF represents a transfer function from a point in space where the sound is located, to a point in the listener. More specifically, to two points: both ears. The variations in the frequency and amplitude of HRTFs are unique for each individual: like a fingerprint. Indeed, individual HRTFs vary significantly, and it is assumed that we have the most realistic perception of spatialized sound when we use our own HRTF in the simulation. This requires methods to measure an individual's HRTF.

The most accurate HRTF measuring systems are complex. They require the person for which we aim to measure their HRFT to stay still inside an anechoic chamber (a room with no echo) for quite a long time, with a set of microphones placed inside their ear canals. A set of loudspeakers is distributed around the listener, usually at a fixed distance. Then, a sine sweep is played from each loudspeaker, usually one at a time but some systems choose to overlap the sweeps between speakers, to get faster measurement processes (P. Majdak et al., 2007). Those signals are recorded with the in-ear microphones, capturing the modification of the acoustic waves caused by the listener's body, head and torso. The resulting HRTF is a set of discrete measurements at various locations around the listener, which provide a full characterisation of the auditory cues used by the specific listener.

The use of a person's specific HRTF (called *individual HRTF*) to simulate a VAS offers a better performance and an increase in the sense of presence (Väljamäe et al., 2004; Xu et al., 2007). However, measuring individual HRTFs for each specific listener is not practical. There are some publicly available HRTF databases measured in real people, or in manikins. HRTFs obtained using a different person than the end-listener are denoted as *non-individual HRTF*. HRTFs measured in a manikin are referred to as *generic HRTFs*. These concepts are reviewed in Chapter 2, Section 2.3.3.2 HRTF individualization.

For simplicity, in this thesis we use the term HRTF for the full transfer function and, by extension, the full set of measurements. We use HRIR to refer to each of the measurements or estimations of this function at a specific location, which together characterise the HRTF. Each location is specified in polar coordinates (see Figure 10). Since the distance is usually the same for all the points of the HRTF, we will refer to each HRIR by its direction, azimuth (θ) and elevation (ϕ). For each listener there is always a pair of HRTFs, one for each ear, called from now on, *left-HRTF* and *right-HRTF*. However, the prefix to indicate the specific ear will be often omitted when the

discussion is general to both ears. Once the listener is characterized by the HRTF a VAS can be synthesized by filtering sound signals with the HRIR estimated for the direction of the corresponding sound sources. The convolution between the sound signal and the HRIR (θ, ϕ) for both left and right ears, results in a binaural signal that is delivered by headphones. Let us illustrate the behaviour of HRIRs with the examples of Figure 12 and Figure 13, which are both for HRIRs measured at the left ear.

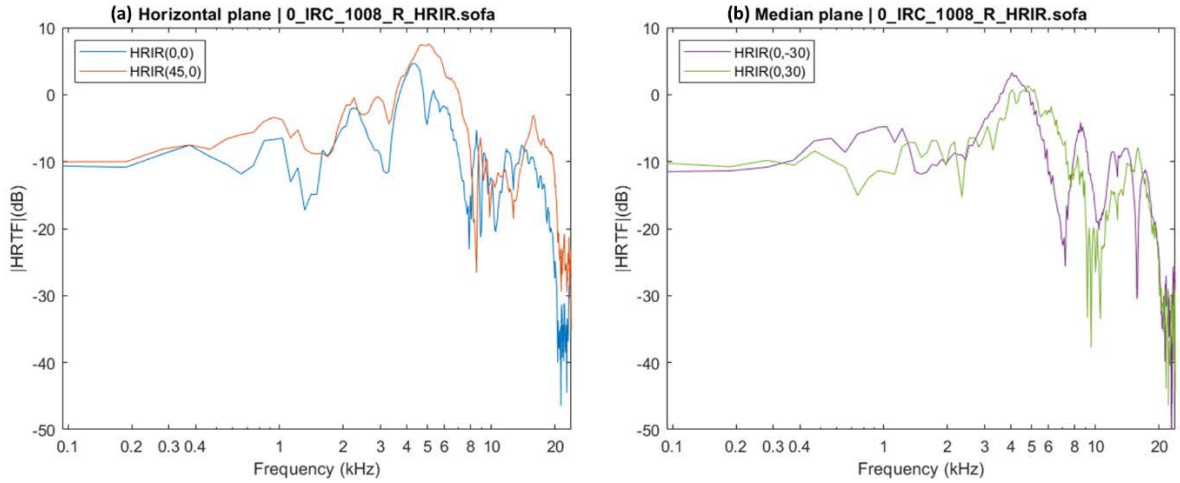


Figure 12. Magnitudes of an HRTF measured at the left ear (Left-HRTF) from the LISTEN database for two different directions on the horizontal plane (a) and two different elevations on the median plane (b).

Figure 12 shows the variation of HRIR magnitude in the frequency domain for specific directions in the horizontal and median plane. At frequencies below 0.4 kHz the HRIR magnitudes are nearly constant, because at these wavelengths the effect of the head is negligible: they are almost frequency- and direction-independent. But, as frequency increases, the HRIR magnitudes vary with frequency and direction in a complex way. These variations are due to the comparable size of the wavelengths with that of the head, torso and pinna of the listener, which now create a filtering which varies largely with the wavelength and the source location. For example, for frequency values larger than 5 kHz, it is well know that the peaks and notches illustrated in Figure 12b are generated by the listener pinna and that the first two spectral notches are cues used for localizing sources in the median plane (Takemoto et al., 2012).

Figure 13 shows the HRTF magnitude for the left ear for many different directions in the horizontal (Figure 13a) and median plane (Figure 13b). On the horizontal plane, when the sound is coming from the left (azimuth from 0° to 180°), the magnitude is larger than when the sound is coming from the right (azimuth from 180° to 360°), where the plotted surface shows a valley in blue colour. The largest magnitudes are around 90° (dark red colour) because sound coming from this direction gets the least possible filtering from the head. The differences between 0° and 180° of azimuth are due to the

front-back asymmetry of the head and the fact that the ear pinna points towards the front.

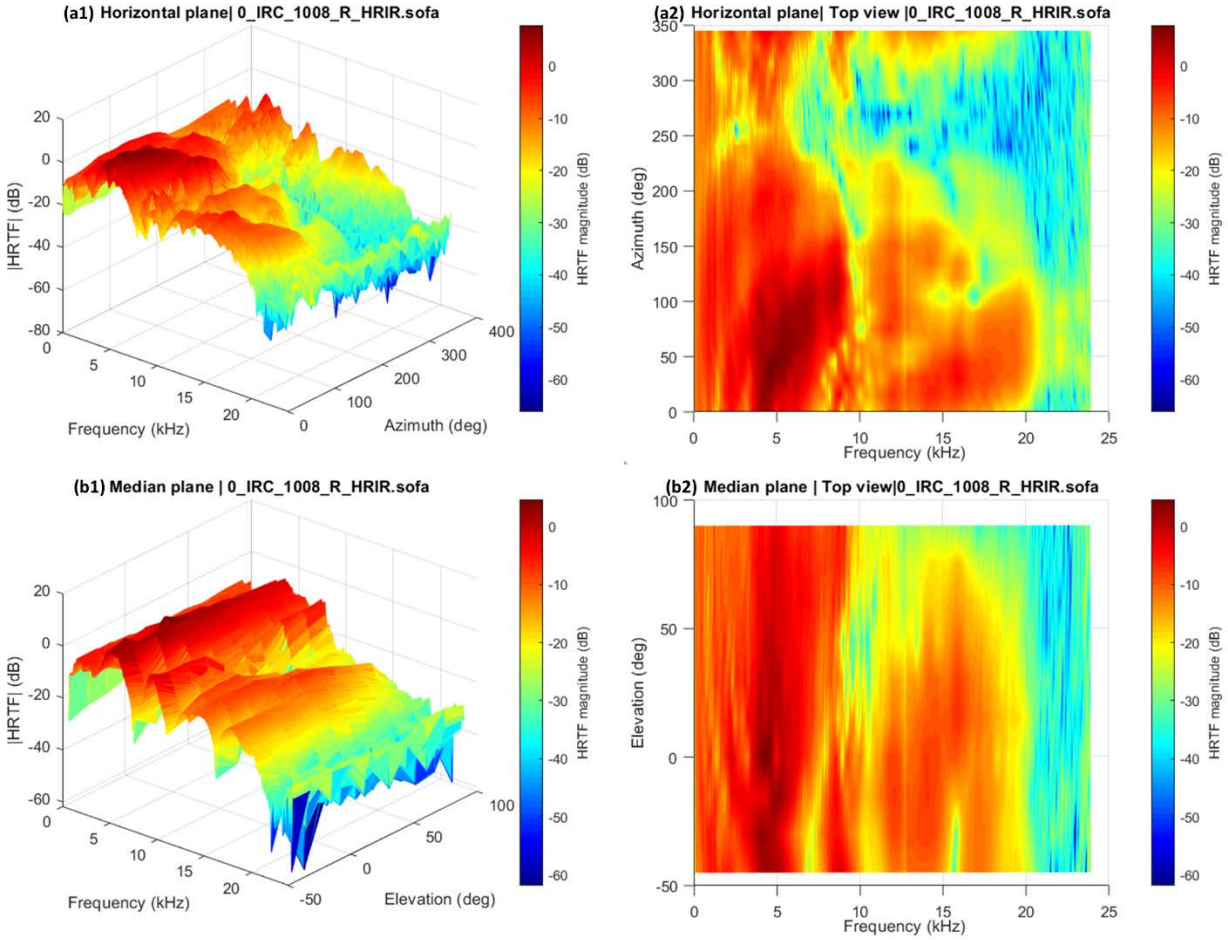


Figure 13. Left-HRTF magnitude representation in the frequency domain for both horizontal (a) and median plane (b), in a perspective (a1, b1) and top view (a2, b2).

Figure 13b shows HRTF magnitudes from many different elevations on the median plane, where a given ear gets approximately the same head shadow, regardless of elevation, so the variation of attenuation with respect to elevation is not as relevant as it was for azimuth. The most relevant elevation cues are the different spectral features caused by the pinna reflections and diffractions at frequencies above 5-6 kHz. These reflections create a set of elevation-dependent *notches* in the spectrum. These notches have been also demonstrated to be strongly individual-dependent. In the top view (Figure 13 b2) we can observe the HRFT magnitudes decreasing for frequencies above 6 kHz (yellow colour) and how these notches can be found at larger frequencies as the elevation increases. The primary features of HRTFs have been described in detail in (Xie, 2013).

1.2.2.3 The environment and the BRIR

The use of the HRTF to simulate spatial sound together with a distance attenuation simulation provides an anechoic model of spatial hearing. However, if we wish to simulate that the listener is inside an enclosed space (e.g., a room), the effects of the environment should be included. The transmission of the sound from the source to the listener in an enclosed space can be decomposed in *direct sound* and *reverberation* caused by reflections of the sound on the objects surrounding the listener, including the walls, the ceiling or the floor, represented schematically in Figure 14. Inside this generic room, the direct sound arrives first, since it travels from the source to the listener along a straight line. With the direct sound the listener can localise the source thanks to the natural filter produced by their anatomy (especially head and ears), which can be simulated by the HRTF. After the direct sound, the reflections reach the listener ears. Simulating reflections contribute to the source localisation and gives more realism to the simulation.

Reflections can be classified into *early reflections*, which arrive separately with different time delays, and *late reflections*, which arrive as a diffuse *reverberation* tail, as shown in Figure 14. Early reflections on nearby objects or room boundaries add colour to the sound source, and contribute to speech intelligibility, since they can be integrated with direct sound (Bradley et al., 2003). Late reflections or reverberation reach the listener after multiple reflections, creating a dense succession of echoes over time. The total energy of reflections decreases due to surface absorptions. The absorptiveness and diffusion of the reflective surfaces is frequency-dependent (Kaplanis et al., 2014; Xie, 2013). We estimate the size of the room mainly from the reverberation time and level, while the complexity of the room will shape the spatial distribution of the early reflections.

For a specific source location within an enclosed space, the temporal and the spatial information of the sound transmission from the source to the listener (both direct and reverberant sound) are encoded in the so-called *Binaural Room Impulse Response (BRIR)*. BRIRs, in the same way as the HRIRs, depend on the relative position of the listener and the source. In this way, rendering a 3D sound in an enclosed and reverberant space, for a binaural system, involves the convolution of the audio input signal with a pair of BRIRs (one for each ear). Within a real environment, the BRIR can be measured in a similar way as how we measure HRTFs, i.e. by reproducing a signal that contains all frequency components through a set of loudspeakers, recording the arriving sound with a set of in-ear microphones on the listener, all of it positioned within a room.

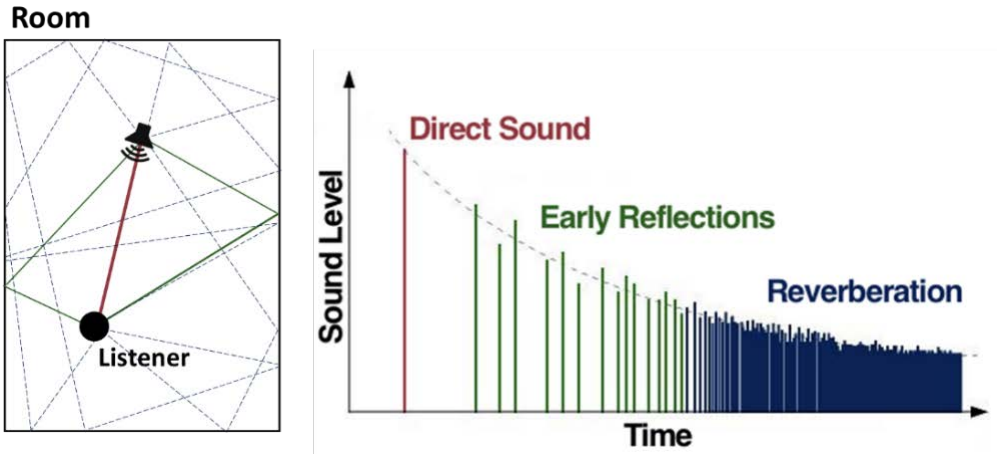


Figure 14. Schematic example of a room impulse response. Early reflections are represented by stronger coefficients than late reflections with more space between them[4].

A time domain representation of a BRIR for a recording in a library is shown in Figure 15.

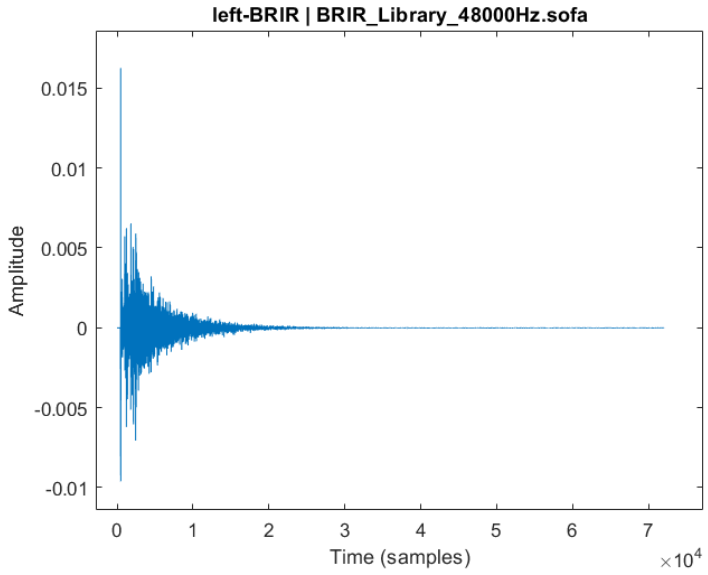


Figure 15. BRIR measured at the left ear of a KEMAR manikin in a moderate sized reverberant library, at position 0 degrees azimuth and elevation.

[4] Image of the right is from Wikipedia, licensed under the Creative Commons Attribution 3.0 author: Lee2008 (https://commons.wikimedia.org/wiki/File:Acoustic_room_impulse_response.jpeg).

1.2.2.4 Challenges in dynamic and real-time virtual auditory scenarios

The localization cues simulated in a binaural system depend on the relative position of the source with respect to the listener. In an immersive VAS, both the source and the listener can be in constant motion. Even in static situations, head movements are a natural component of spatial sound localization: we move our head and re-orient our body to make better decisions about the localization of an audio signal, especially to estimate elevation and front/back positions. These movements will cause the relative position between the listener and the source to change and therefore a modification of the auditory cues. Therefore, VAS systems should be able to dynamically detect the position and orientation of the listener with respect to the sources, and carry out the simulation of the spatial audio in real time. A basic structure of a binaural dynamic VAS system is shown in Figure 16. Depending on the application, the detailed structure can change but all have in common this general structure (Xie, 2013), which consists of three parts: the information input and definitions, real-time processing of the signal and the reproduction system worn by the listener.

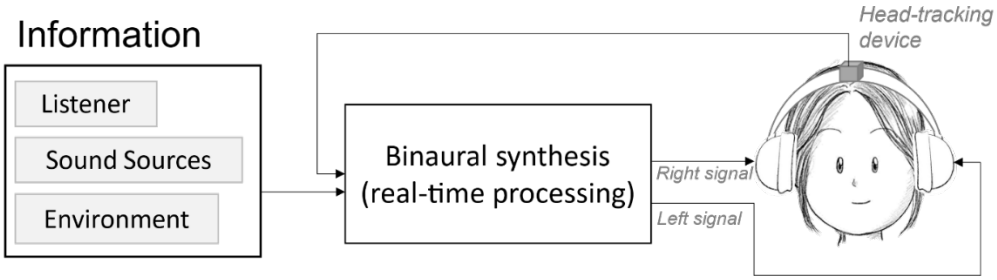


Figure 16. Basic structure of dynamic VAS systems

The information part provides data to the dynamic VAS. These data is related to: (1) the listener, as the individual information contained in the HRTF, (2) the sound source, as the stimuli, spatial position and orientation and sound level and (3) the environment, as the room or environment geometry, the BRIR and characteristics of the surfaces materials or the air absorption coefficients. The real-time binaural processing takes all previous mentioned information, together with the position and orientation of the listener head that comes from a head tracker device, to simulate the sound source. Finally, the binaural signals (left and right signals) are delivered to the listener through a pair of headphones. All these concepts are described in detail in Chapter 2, Section 2.3 Binaural rendering.

In a real-time digital audio system, processing becomes complicated with multiple sources moving at the same time, and changes in the listener position and orientation. This leads to a continuous update of the information and the signal processing, incurring

in a large computational cost. Furthermore, these systems have to process the audio using relatively large frame sizes (256 or 512 samples), so when different parameters and filters vary (due to, for example, changes in the relative position between the listener and the source), the change is usually reflected in the next frame. Even if the change was made smoothly, when the next frame is processed, the change is going to be abrupt producing audible artefacts, which can cause discomfort in the listener, as well as a loss of naturalness and presence.

In real-time audio rendering, the timing requirements are very important. The update rate and the system latency are crucial on the perceived quality of the sound. The higher the update rate, the closer the auditory perception to the real environment. The system latency refers to the time from which the listener or a source change their position to the time at which the corresponding change in the output of the binaural simulation is delivered to the listener. System latency should be reduced as much as possible, but it depends on many factors, such as head tracking response, data transmission, or time required for signal processing. Latency is determined by the hardware and the software structure of the system. Achieving certain requirements in update rate and latency, in a scenario composed of multiple moving sources in an enclosed space, demands a very high computational power. Depending on the method used in the simulation, as well as the implementation of the method, the computing power required can be higher or lower. Systems that make use of HRTF and BRIR are the most accurate in terms of simulating real sounds at its positions, but they are also the most computationally expensive. Chapter 2 discusses the existing methods, approaches, and tools for simulation of real-time spatial audio.

1.3 Context and motivation of this Thesis

This thesis has been developed within the DIANA research group at the University of Malaga, in the framework of the EU-founded project 3D Tune-In[5]. This project aimed at using 3D sound and gamification techniques to support people using hearing aid devices. Within the project, the DIANA research group was in charge of developing the "*3DTI Toolkit*", an open-source C++ library which integrates binaural spatialisation functionalities, together with other audio-related features such as a hearing loss and hearing aid simulation. Within the development of the 3DTI Toolkit, this PhD thesis has been focused on the design, development and evaluation of the *Binaural Spatialiser (3DTI Toolkit-BS)*. Within the 3D Tune-In project, several applications were developed using the 3DTI Toolkit, deployed on multiple platforms (e.g. mobile, desktop or

[5] http://www.3d-tune-in.eu/

browser), tailored to different target audiences (e.g. older users or children) and scenarios (e.g. music listening or noisy environment simulation).

The reason behind the need to develop a custom, open-source, multi-platform C++ library can be found in the challenging set of requirements on real-time performance and portability, as well as on the transparency of the audio processing chain. It is true that with the uprising of VR in the last years, a large number of binaural rendering tools has been released (a complete overview can be found in the Chapter 2, Section 2.4). Nevertheless, the main motivation for the development of a custom binaural spatialisation library from scratch was the need for several features which did not exist at the beginning of the development of this thesis. At the time of writing, and to the best of our knowledge, no other tools offer the following complete set of features:

- Full real-time 3D placement and movement of sources and listener, including near- and far-field simulation.
- Customization of HRTFs.
- Spatialised reverberation simulation.
- Smooth behaviour in dynamic situations.
- Multi-platform support (including web audio)

As described in the following chapters, the 3DTI Toolkit-BS integrates in one single open-source package several techniques and functionalities developed and evaluated in the last 20 years of spatial audio research. During the development stage, particular attention was put on the time-related aspects of the spatialisation, resulting in realistic and smooth simulation of moving sound sources, both in terms of direction and distance changes. The implementation of all these functionalities within an open-source tool provides full control on the spatialisation process, as well as the opportunity for future developments within the 3D audio and psychoacoustics communities.

The developed library was used to carry out a psychoacoustics experiment, presented in this thesis, where the 3DTI Toolkit-BS has been used to *study the influence of a non-individual HRTF on the speech intelligibility.* It is known that HRTF cues have an impact on speech intelligibility, however how these cues affect each individual and, more specifically, the impact of the HRTF choice on speech-in-noise performances in cocktail party scenarios has not yet been investigated in depth. The conduction of this study allowed both, to evaluate the performance of the 3DTI Toolkit-BS and to go deeper into the study of HRTFs and their relationship with speech intelligibility. In addition, with this experiment we want to test the use of the library as a tool to develop psychoacoustics experiments.

1.4 Research objectives

The aim of this thesis is the design and development of an efficient and flexible binaural spatialisation tool, to integrate 3D audio in the most effective way in a virtual reality environment. This means, the implementation of a set of algorithms to:

1. Simulate the propagation of the direct sound between the source and the listener in the most precise way, based on the individual characteristics of the listener and considering all the cues that lead to the perception of the localization of a virtual sound source.
2. Simulate the reverberation of the environment accurately, collecting the directional characteristics of the reverberant environment. This must be done efficiently, to allow a normal computer to perform the process.
3. Support static and moving sources to simulate dynamic scenarios.
4. Ensure smooth audio changes with non-audible artefacts in dynamic situations, when some characteristics of the scenario are modified.
5. Process the 3D audio in real-time on a "commercial PC" without specialized DSP (Digital Signal Processing) hardware and with no noticeable latency.

Additionally, to evaluate the binaural spatializer, test its use as a virtual psychoacoustics laboratory and to further study HRTF and its relationship to individual characteristics of the listener, a perceptual study has been carried out which main objectives are:

6. To study the impact of the HRTF on the speech intelligibility
7. To study the impact on individual listeners of different non-individual HRTFs on speech intelligibility within a VR Cocktail Party context.
8. To evaluate the use of the 3DTI Toolkit-BS in a virtual psychoacoustic experiment.

1.5 Outline of this Thesis

The thesis is organized as follows. Chapter 2 presents the state of the art of the 3D audio simulation in virtual environments. In order to put this work into context, the basic concepts behind binaural spatialisation are described and reviewed through a chronology of research milestones in the last decades. This chapter also surveys the currently available open- and closed-source binaural spatialisers and presents their most relevant features.

Chapter 3 describes and characterizes the 3DTI Toolkit-BS, presenting the technical details of this binaural spatializer renderer, outlining its software architecture, and describing the algorithms implemented in each of its components.

In Chapter 4 an evaluation of the 3DTI Toolkit-BS is carried out. The renderer performance, as well as some other relevant metrics such as non-linear distortion are assessed and presented. The 3DTI Toolkit-BS has been evaluated with special focus on how well it behaves with moving sources, estimating the distortion produced by our implementation in such situation. Algorithms to interpolate HRIRs and BRIRs are also assessed.

Chapter 5 presents and discusses an experiment where, making use of the 3DTI Toolkit-BS, the impact of a non-individual HRTF on the speech intelligibility has been studied.

Finally, Chapter 6 draws the conclusions and points out some future research.

Chapter 2

State of the art

This chapter introduces a brief description of the technical background required to follow the rest of the chapters, as well as a review of the state of the art and research lines in the field of binaural rendering. First, the most popular techniques and research milestones in the field of spatial audio are presented in Section 2.1, followed by a review of the role of the spatial audio in Virtual Environments in Section 2.2. Then, Section 2.3 presents a complete revision of the binaural rendering systems, including a detailed model of a state-of-the-art binaural spatialisation tool. In addition, this section describes the most used algorithms and techniques for the simulation of spatial audio. Section 2.4 shows the list and the description of the existing tools to create spatial audio in virtual auditory scenes. Finally, a description of the most relevant auditory models is included in Section 2.5.

2.1 Spatial audio techniques and research milestones

2.1.1 Binaural localization of 3D sounds

The first studies looking at *binaural sound perception* can be tracked back to 1907 by Lord Rayleigh, in the article "On our perception of sound direction" (Lord Rayleigh, 1907), where a series of experiments are described. Those studies investigated our capability to estimate the direction of arriving sounds and how we discriminate their relative position. Rayleigh distinguished between the capability of the listener to locate sounds coming from left-right and front-back locations, affirming that relative intensities and time differences between the two ears play an important role in the localization cues, calling it the “duplex theory”. This theory refers to what was later called ITD

(Interaural Time Difference) and ILD (Interaural Level Difference), differences in the signal time and level between the sound reaching both ears. Rayleigh also proposed that we localize low-frequency sounds from time differences (ITD cues) while higher frequency sounds are localized based on difference of intensities (ILD cues). He also suggested that these localization cues made right-left discriminations easier than front-back estimations and stated their frequency dependence. These experiments were the starting point for the following studies on acoustic localization cues.

The earliest paper that tried to measure the diffractions around the listener head seems to be the one published by Wiener & Ross (1946). They measured the variation of the sound pressure in the listener's ear, inserting a pair of small microphones in the subject's ear canals. After the experiment, they described the human ear as an effective acoustic "amplifier" and defined this amplification as the combined effect of diffraction around the head and pinna with the resonance in the auditory canal. The role of the pinna was later studied by Batteau (1967), describing the filtering effect caused by the interaction of the sound with the outer ear. However, it was not until 1980 that the term HRTF (Head-Related transfer Function) was used. The first paper that presented the term seems to be the one written by Morimoto & Ando (1980). In addition, one of the first extensive and rigours measurements of HRTF was carried out and published by Wightman & Kistler (1989a, 1989b), whose measurement system was later replicated by many in the field.

In 1983, Blauert published the first English version of one of the most important books in this area: "Spatial Hearing" (Jens Blauert, 1983). This book is considered by many the most important textbook on spatial hearing perception. It laid the foundation for what it is nowadays known as audio spatialization. Blauert presented the fundamentals of spatial hearing, based on experiments focused on auditory perception. The book described the advantages of binaural compared with monaural hearing, in terms performance localization of both single and multiple sources.

A decade later in 1994, D. Begault, in his book "3-D Sound for Virtual Reality and Multimedia", presented the context of a *Virtual Auditory Space (VAS)* and gave an overview of the spatial hearing problems. This and many others studies were performed by Wenzel and other colleagues affiliated to the NASA (National Aeronautics and Space Administration independent agency of the United States Federal Government). In the 90's-2000's NASA was one of the pioneers in spatial hearing research and real-time simulation (Foster et al., 1991; E. Wenzel et al., 2000). In this context, Begault and Wenzel did some foundational research on the binaural technique and its applications to air traffic control, presenting works looking at different factors influencing spatial perception of speech in VAS, via headphones, using HRTFs (Begault & Wenzel, 1990, 1993).

At the beginning of this decade, Blauert's studies in spatial hearing were also numerous. In (Jens Blauert, 1994) he presented the concept of *binaural technology*, defined as «a body of methods that involves the acoustic input signals to both ears of the listener for achieving practical purposes, for example, by recording, analysing, synthesizing, processing, presenting, and evaluating such signals». According to him, binaural technology involves all the techniques and devices that make use of three aspects: (1) physics, employed in the *reception of the sound signal*, which concerns the input signal before it reaches the inner-ear, (2) the subcortical auditory system, referring to the *perception*, i.e. the psychoacoustics of binaural hearing, the part of the auditory system that converts incoming sounds waves into neural spike trains and (3) the cortex, where the interpretation of the acoustic signals is carried out, which he calls the psychology of binaural hearing.

In 1997 Jens Blauert published an enlarged edition of his book written in 1983 "Spatial hearing: the psychophysics of human sound localization'' (Jens. Blauert, 1997). As the one presented in 1983, this book gives an analysis of the fundamentals of spatial hearing, in addition to a review of the most relevant studies on 3D sound localisation. This is one of the most referenced books and, at that time, it set the foundation for new challenges in spatial hearing research and VAS. In the same period, Møller and Hammershøi published a large number of papers on HRTFs. Some examples are "Fundamentals of binaural Technology" (Moller, 1992) and "Head-Related Transfer Functions of Human Subjects"(Moller et al., 1995), in which they studied binaural recordings and HRTF measurements, describing where to position the microphone in the ear canal during the measurements. In addition, they investigated the binaural playback system, looking also at audio transmission models inside and outside the ear canal and describing the sound transmission from a headphone. Regarding the modelling of HRTFs, other studies in the mid-90's should be noted, such as those by Duda (1993) and Lopez-Poveda & Meddis (1996), which studied the HRTFs and presented approximated physical models of the frequency transfer function.

Several other studies in those years investigated the possible causes for what is known as *IHL (Inside-the-Head Localization)*, which refers to poor sound source externalisation, that results from the listener perceiving the localized sound inside their head, without any perception of distance, especially when the sound is delivered by headphones. Many studies agreed that the effect of pinnae should not be ignored since they play an important role in externalization. However, this is only one of the main factors that help to avoid the IHL, as it described in (Mershon & Bowers, 1979). Bronkhorst & Houtgast (1999) demonstrated that the use of head-tracking to consider the relative position between the listener and the source in the simulation of the spatial sound improves the feeling of externalization. In addition, Hartmann & Wittenberg (1996) studied the characteristics of the signals that are responsible for the perception of externalization,

finding that externalization depends also on interaural differences. They showed that ITD affects externalization only for low-frequency components (below 1KHz). However ILD is important for externalization at all frequencies. They also highlighted the importance of headphone calibration to perceive a sound delivered as real. IHL is also very related with the perception of distance. Little et al. (1992) investigated how distance can be perceived thanks to the spectral content, i.e. we are able to discriminate that high frequencies are more easily attenuated over distance than low frequency sounds. A very recent review on sound externalization can be seen in (Best et al., 2020).

The importance of *dynamic localisation cues* was studied, among others, by Inanaga et al. (1995) and Perrett & Noble (1997), concluding that the localization accuracy and externalization can be improved with small *head movements.* To include this movements in a binaural sound simulation, the head position and orientation must be estimated, which requires some means to track the head movements. A study by Elizabeth M. Wenzel (1995) confirmed that, compared with static conditions, head movements provides listeners with extra information required to resolve localization issues. Begault et al. (2001), compared the impact in localization of accuracy of head tracking, synthesis of virtual rooms and individualised HRTFs. They found significant effect of the first two when looking at externalisation, while an optimal localization performance was achieved when the three factors were included.

In the field of spatial audio research, the work carried out by Pulkki and his team at Aalto University is also noteworthy. During the nineties they did an extensive work on the development of several *spatial audio reproduction techniques*, looking in depth at spatial hearing perception in complex audio-visual environments. They also created a computational model of auditory perception, which consisted in a model predictor to estimate the localization of different sources, based on objective measurements (Pulkki et al., 1999). Models of binaural perception go back to the 1980s, with early overviews in the subject such as (Jens Blauert, 1983; Colburn, H. S., & Durlach, 1978; Stern, 1988). Auditory models are reviewed later in Section 2.5.

All the findings presented above refer mainly to direct sound location in free field, where no reflections exist. *Spatial hearing in rooms* lays in the field of *room acoustics*, a subject treated extensively in publications such as (Beranek & Martin, 1996) or more recently the book (Kuttruff, 2016). Enclosed spaces create auditory effects due to reflections from walls, floors, ceilings, etc. The room acoustical parameters can be evaluated using the Room Impulse Response (RIR) (Schroeder, 1965), which describes the sound transmission between the source and the receiver in an enclosed space. Later, Allen & Berkley (1979) investigated how to simulate the reverberation of an impulse response between two points in a small rectangular room. When the RIRs are measured using microphones inside the listener ear canal, apart from the RIR it includes the HRTF

information. This function is called BRIR (Binaural Room Impulse Response). The foundations of what was known at the beginning of the 90's were laid by Lehnert & Blauert (1992). A very extensive review about "Fifty years of artificial reverberation" can be found in (Välimäki et al., 2012).

Reverberation is also important to estimate distances and to improve externalization, helping to reduce the previously mentioned IHL problem. Bronkhorst & Houtgast (1999) looked at the effects of room acoustic simulation on the perception of distance. They presented a couple of experiments that showed a simple model based on a modification of a direct-to-reverberant energy ratio to predict distance in an enclosed environment. In an enclosed room, when distance between the listener and the source increases, the level of reverberation remains similar. However, the level of the direct sound decreases considerably, allowing the use of the direct-to-reverberant ratio as distance cue (Békésy & Wever, 1960). In a more recent and very relevant study, Begault et al. (2001) showed that simulating reverberation improves externalization. However, this effect is stronger if the simulated reverberation matches the listener's expectations (Werner et al., 2016).

This section has described the most important research milestones from the beginning in 1907 with Lord Rayleigh to the end of the 90s, where research on spatial hearing started to become stronger and branched out into numerous fields of research and development, some of which will be reviewed later in more detail. For a history of binaural recordings, see (Paul, 2009) and for a more up-to-date review of spatial audio recording and reproduction see (W. Zhang et al., 2017)[6].

2.1.2 Other approaches to simulate spatial audio

Spatial audio first reached the public in theatres and movies, in the form of loudspeaker setups. In 1933, Stokowski and Fletcher produced a 3-channel transmission of the Philadelphia Orchestra, reproduced for an audience in Washington, D.C. (Torick, 1998). This can be considered as the first example of one-dimensional and highly truncated *Wave Field Synthesis* (WFS) reproduction system. The Walt Disney Movie *Fantasia* from 1940 is considered as the first movie that used surround sound conducted by Stokowski. Multichannel formats for the cinema became popular more than a decade later, at the end of the 1950s and during the 1960s (Rumsey, 2001). In the 1970s, new formats like quadrophony or Ambisonics were developed to bring surround sound

[6] For a non-research oriented review, see https://www.linkedin.com/pulse/history-binaural-audio-part-1-anthony-mattana and https://www.linkedin.com/pulse/history-binaural-audio-part-ii-resurgence-1940-2000-anthony-mattana. Retrieved January, 2022.

experience to homes. However, these formats failed to obtain commercial success. Later, these technologies finally migrated to homes with the Dolby Digital, and the popluar 5.1 standard (ITU-R, 1993). More details regarding the history of multichannel sound can be found in (Torick, 1998) and the book "History of 3D Sound", in the chapter by Roginska & Geluso (2017).

The basic idea of *Wavefield Synthesis* (WFS) was presented more than 70 years ago by Snow (1953) and later formulated by Berkhoul et al. (1993). The first works that implemented this technique were restricted to reproduce the sound in a planar listening area, using a linear array of loudspeakers. At the beginning of the 2000's, Hulsebos et al. (2002) presented a work where three different configurations were investigated: linear, cross-shaped, and circular arrays, with the last one having the best performance. Since then, many methods have been employed to improve the technique and allow practical implementations of WFS for two- and three-dimensional reproduction (Gauthier & Berry, 2006; Spors et al., 2008). In those years, with the increase in availability of computational power, this technology gained a commercial interest, supported by hardware and software solutions such as the ones proposed by Brix et al. (2003) and Pellegrini & Kuhn (2004), both authoring systems for WFS content production, mixing and mastering. An example of a commercial product can be seen in Holoplot[7], a speaker audio WFS based technology including both hardware and software that provides sonic experience in complex acoustic environments. Finally, some open source renders available to create spatial audio using WFS can be found in (Baalman, 2005; F. Völk et al., 2008).

The technique known as *Ambisonics* was developed in the 1970s by Gerzon (1973). However, it wasn't until the 1990s that these techniques began to gain interest, with the appearance of the *Higher-Order Ambisonics (HOA*) theory (Malham, 1999; Malham & Myatt, 1995). As introduced in Chapter 1, first order Ambisonics encodes the audio signal and models the spatial sound using four channels (B-format). The signals are subsequently decoded at the listener's reproduction system, using various techniques (Jérome Daniel et al., 1998). HOA offers more channels and therefore more digital components to the B-format, allowing more accuracy, enlarging the *sweet spot* and offering a better spatialization. In the 2000s, with the advances in technology and the miniaturization of audio devices, Ambisonics was mostly used for high resolution recordings (Moreau et al., 2006).

The renewed interest in Ambisonics comes, among other things, from the use of what is known as *Virtual Ambisonics*. This technique consists in spatialising the sound

[7] https://www.prosoundnetwork.com/business/innovations-holoplot-wave-field-synthesis-technology-immersive-audio-installation. Retrieved January, 2022.

sources, encoding the input signals using the Ambisonic technique, and then decoding them on a series of virtual loudspeakers, where each speaker signal will be finally rendered in the binaural domain through convolution with the HRTFs (Markus Noisternig et al., 2003). This technique will be explained in detail in Section 3.6.

Nowadays, Ambisonics is gaining a lot of interest in the area of virtual reality systems and games, thanks to the fact that it is a non-proprietary format and offers a flexible and scalable sound representation. It also has major limitations such as low resolution for low order Ambisonics (Frank et al., 2015), which leads to a poor source spatialisation. With high order Ambisonics we can get a finer resolution and then a better simulation of the source position but with a higher computational cost. That is why real-time applications use to work with low order Ambisonics. Companies such as Google use it in its VR audio technology, the Google Resonance Spatialization Tool, (Gorzel et al., 2019), as well as YouTube and Facebook in their 360-degree videos. Ambisonics is often used for recording purposes and many commercial microphones use this technology[8] (*Products | Mhacoustics.Com*, n.d.)

2.2 Spatial audio in virtual environments

2.2.1 Research milestones in VAS

The incorporation of 3D audio into virtual environments has always been closely linked to the development of the technology. While the science behind 3D audio was grounded through the whole 20th century, as seen in the previous section, the technology necessary to achieve a proper Virtual Auditory Space (VAS) did not emerge until the end of the same century (Carlile, 1996). The *Convolvotron* was one of the first commercial products able to create auditory virtual environments. The signal processor was marketed by Crystal River Engineering of Groveland CA and designed by the company's president, Scott H. Foster, as part of an auditory research program conducted by Elizabeth Wenzel of NASA Ames Research Center (Foster & Wenzel, 1992; E. M. Wenzel et al., 1988). This device created 3D sound by converting monoaural inputs into spatialized digital signals, through a Digital Signal Processing (DSP), using HRTFs filters and headphones. In the 90's Crystal River Engineering developed new devices to create a VAS in a more efficient way: the *Newtron*, which was able to perform frequency-domain filtering, or the *Snapshot*, a portable system for measuring HRTFs. In this period, Wenzel and Foster carried out several studies using these devices to perform the

[8] For example the one provider by mhacoustics (https://mhacoustics.com/products). Retrieved January, 2022.

signal processing of an auditory environment in real-time (Foster et al., 1991; Foster & Wenzel, 1992; Elizabeth M Wenzel, 1998). Another 3D audio dedicated hardware was the *Lake DSP Huron* which is considered as one of the first commercial reverberation products to use convolutions (Reilly & McGrath, 1995). A good review of this and other devices can be found in (Parker et al., 2008).

In the 1990s, the PC-based audio solutions started to emerge, where the signal processing was handled mainly by the computer CPUs or the sound cards. In 1994, Begault wrote the book, "3D sound for VR and multimedia", which described the system and the perceptual requirements of hardware and software necessary to create virtual spatialised sound (Begault, 1994). In 1996 Microsoft incorporated *DirectSound3D (DS3D)* into Microsoft's DirectX system, an attempt to standardize 3D audio in the Microsoft Windows OS. This system offered a standard API to create 3D audio which provided a direct communication with the soundcard. However, this API just allowed to control volume, pitch and a simple left/right pan. Many extensions were developed, as the *Environmental Audio Extensions (EAX)* by Creative Lab's, that allowed to add reverberation to the DS3D. Years later, in 1999 with the release of *DirectX 7*, the Microsoft API offered a spatialisation with two different HRTFs. Other implementations for audio spatialisation were the sounds modules of *Java 3D API* and the *MPEG-4*. More information about these APIs and others can be found in (Murphy & Neff, 2010).

From the middle of the 1990s, thanks to the technological advances and the increase in computer power, it is possible to carry out binaural signal processing in real-time which has led to an increase in these technologies (Blauert, 2013). In 1995, a real-time modular spatial-sound-processing software system, developed by IRCAM and called *SPAT* was presented (J. M. Jot, 1999). This was the first commercially available high-quality real-time software dedicated to binaural sound spatialization. IRCAM released several updates since then, integrating various additional features, such as HRTF selection, artificial reverberation, and sound source directivity. IRCAM celebrated its twenty-year anniversary with a position paper that traces the evolution of the software and talks about the past, present and future of binaural spatialisation (Carpentier et al., 2015). In the last decade, numerous VR applications have incorporated binaural reproduction methods and software (Xie, 2013). More details about relevant spatialisation tools will be included in Section 2.4 Existing tools to render binaural audio.

2.2.2 The importance of spatial audio in a Virtual Environment

Adding spatial audio to a Virtual Environment (VE) improves the way a subject interacts with the environment and increases the feeling of full immersion and the sense of presence (Bormann, 2005). Many works have studied the effects of spatialized sound on the subjective *presence* in VE. Hendrix & Barfield (1996) showed a significant increment on the sense of presence when adding spatialized sounds through a binaural system with HRTFs, comparing three scenarios for a navigation task with no sound, non-spatialized sound and with spatialized sound. In addition, according to (Väljamäe et al., 2004), using individual HRTFs showed a significant increase in presence ratings. Kapralos et al. (2004) studied the role of the auditory cues regarding self-motion (sensation of actual movement relative to a stable surrounding environment), obtaining that self-motion estimation was most accurate when both physical motion and auditory cues were performed simultaneously. Larsson et al. (2002) carried out a pair of experiments, regarding navigation and memory tasks, which suggest that high quality spatialisation of sounds may greatly improve the overall performance in the VE and give a better sense of presence. A more recent study by Kobayashi et al. (2015) shows an objective study using physiological and psychological measurements and which results suggested that there is a correlation between the psychological and the physiological responses in the spatialized sound condition, regarding the sense of presence which can be improved by the addition of auditory cues.

All previous works talk about presence as the illusion of "being there". However, when talking about the effect of the audio in the plausibility of a VE (Lindau & Weinzierl, 2012), i.e., the illusion that events of the virtual word are really happening, it is not clear that the spatial audio brings advantages. The previous mentioned work by Hendrix & Barfield (1996) found that spatialized sound positively influences presence, but not the perceived realism of the VE. In addition, Bergstrom et al. (2017) studied the plausibility of a musical performance in VR using four features: gaze (the musicians looking toward and following the participants), spatial sound, auralization (room reverberation) and environment (environmental sounds from outside the room). They found that the highest influence on the level of plausibility was given by the gaze and the environment, followed by the auralization and the spatialisation. However, it is important to consider how the spatial sound was created. In (Hendrix & Barfield, 1996) they used a generic HRTFs, with no reverberation and no head tracking. In (Bergstrom et al., 2017) the spatialisation was achieved through Virtual Ambisonics and non-individualized HRTFs. Many works have demonstrated that the use of individual HRTF provides better quality regarding the spatial simulation of a sound and improves the subject performance and interaction with the VE (Begault et al., 2001; Oberem et al.,

2020; Xu et al., 2007). However, none of them have discussed the relation between individual HRTF and plausibility. As far as we know, this is a topic that needs more research.

2.2.3 Real-time and dynamic VAS

When simulating a dynamic VAS, i.e. a virtual auditory space where either the listener or the sound sources are moving, real-time processing is required. The real-time processing of spatial audio in a VAS with multiple sources involves an intensive use of CPU, more so when room acoustics is included. Processes using convolutions with HRIRs and BRIRs are the most accurate ones but also the most computational power demanding, which makes it more difficult to meet the real-time requirements. In addition to the modelling of sound sources and the environment, a dynamic VAS must be able to constantly capture the position and orientation of the listener head which is usually detected with a head tracking and updated in real-time. According to (Sandvad, 1996), there are three relevant parameters to obtain a good quality of the auditory perception in a dynamic VAS: system *latency*, system *update rate* and *spatial resolution* of the HRTFs.

System latency refers to the interval time between the occurrence of an action, such as a head movement, to the auditory response. Latency is measured in milliseconds and is considered as an indicator of the quality of the VAS, in which the aim is to achieve the lowest possible latency. The *update rate* is directly related with the *audio frame size* used to process the signals and refers to the frequency with which the parameters of the system, such as the position of the sources, are updated. In addition, in a dynamic VAS system, the *spatial resolution* is important since the system would require a real time updating of the HRTF filters and the availability of those filters in many positions. In this way, the spatial resolution of the system can be increased using interpolation methods (Gamper, 2013). These parameters are discussed in more detail in Chapter 4, 3DTI Toolkit-BS Evaluation.

Sandvad (1996) investigated the perceptual impact of these three parameters, determining the necessary values for them in a dynamic auditory virtual environment with a localization experiment. The study reported that the *spatial resolution* seems to have very little influence on subject performance. However, reducing spatial resolution and the update grade, subjects seemed to be able to ignore the audible artefacts. In addition, Elizabeth M. Wenzel (2001) following the work of (Sandvad, 1996), studied the impact of the system *latency* on localization accuracy, finding that the localization accuracy, as a function of latency, was also moderately affected by the overall duration of the sound. They compared two stimuli of 3 and 8 seconds. The study results suggested

that listeners were able to ignore latency during localization of long duration stimuli, where localization was generally accurate, even with a latency as great as 500.3 ms. Yairi et al. (2007) estimated the *Detection Threshold (DT),* which indicates the maximum value of the latency of the system, for the listener to detect a delay. They performed a listening experiment with 9 subjects, where a sound source was rendered with five different latencies and subjects were asking to judge if they could detect the delay of the sound stimulus. Results suggested that, in order to avoid sound artefacts, the system total latency should be 45 ms at most. Yairi et al. conclude that, given the overall average of previous studies, the DT in a dynamic VAS can be estimated to be around 60 ms. Stitt et al. (2016) investigated the influence of head tracker latency on the perceived stability of virtual sounds, comparing a simple and a complex space and obtaining that the DT was 10 ms higher for the complex scene than for the simple one.

To improve the computational capacity of the system and optimize the previously described parameters, recent works propose some solutions as performing the DSP directly on the Graphics processing unit (GPU), taking the advantage of its superiority with respect to computational time and the highly parallel programmable capacity. Cowan & Kapralos (2008, 2009) presented a GPU-based convolution method that allowed for real-time rendering for an arbitrarily sized sound signal and a filter which is far more computationally efficient when compared to conventional, time- domain, software-based convolution. Belloch et al. (2013) described a headphone-based multisource spatial audio portable application, which carried out all the required processing on the GPU. This application could manage the HRTF-based rendering of multiple moving sources (up to 240) simultaneously without overloading the CPU thanks to a parallel computation on the GPU. In addition, GPUs can be used for enhancing room acoustics simulations. Lauri Savioja (2010) carried out a real-time set of room acoustic simulations of a modest-size geometry with a finite-difference time-domain model using the parallel capacity of the GPU.

Regarding devices that support interactive VAS, smartphones can be considered as a good candidate since they are widely used, can be used with headphones, incorporate compass and gyroscope. However, their computational power is very limited and therefore, real-time rendering using multiple sources and/or BRIR convolution cannot be performed. For this type of devices other approaches are used, such as the Virtual Ambisonic (used by platforms such as Facebook and YouTube). Other systems, such as the one presented by Katz et al. (2012), use a lightweight mobile PC in a backpack to implement a navigation system. This tool incorporates a 3D binaural spatialisation renderer with a distributed signal processing architecture that performs BRIR convolutions (Iwaya & Katz, 2018).

In recent years, the use of Head Mounted Display (HMD) has successfully re-emerged due to its application in video games. HMDs bring more realism to the games and this level of immersion has caused the increment on interest in 3D audio. Thanks to the fact that HMD incorporates head tracking it has become one of the most used devices that integrates scenarios with dynamic 3D audio and video simulation. Although, to do so, the HMD has to be able to process both 3D audio and video or be connected to a computer that offers enough computational power. For example, PlayStation VR supports spatial audio with his PSVR's Processor Unit, an external box that handles the processing of 3D audio. In addition, the HMD SDKs can incorporate spatialized audio, as the Oculus Rift SDK that allows spatial audio rendering through HRTFs.

In addition, and for scientific research and industrial applications, the use of CAVE-like displays to create an immersive environment is very extended (Cruz-Neira et al., 1993; Vorländer et al., 2010). These displays are room-mounted installations based on a combination of large projection screens surrounding the user, a loudspeakers setup to deliver the 3D audio and a set of tracking cameras to record the users' positions.

Figure 17. a) User wearing an Oculus Rift playing an 3D audio-based game. b) The immersive environment CAVE at RWTH Aachen University (Vorländer et al., 2010).

2.3 Binaural rendering

The aim of a binaural rendering tool is to make the listener, who is wearing a pair of headphones, have the perception that a sound is emanating from a specific location in the surrounding 3D space and within a given environment, as though it was a real sound. Spatial audio rendering using binaural technology is considered very close to natural listening (Langendijk & Bronkhorst, 2000; Martin et al., 2001). Research milestones of binaural audio, presented in Section 2.1, together with the Introduction chapter, might

help to understand the fundamentals of the binaural audio spatialisation. This section describes the basic principles, structure and implemented algorithms of a binaural rendering tool, which has been designed to simulate a dynamic, and consequently real-time VAS.

2.3.1 Components of a binaural rendering tool

The different components of a headphone-based binaural renderer tool can be classified into 4 groups which are presented in Figure 18.

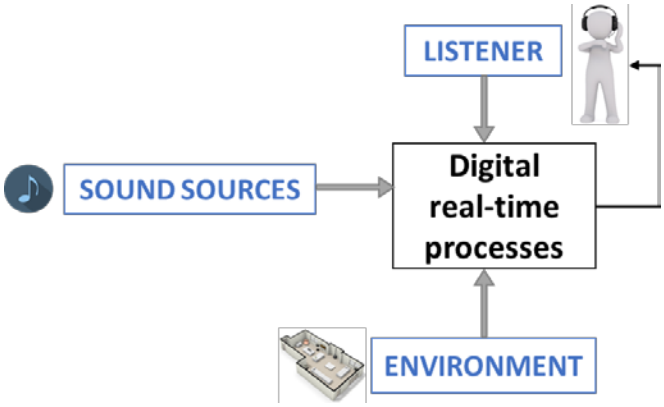


Figure 18. Simplified schematic model for a binaural renderer tool used to create an interactive VAS. The pipeline can be described as: the system spatialise in real-time a set of *sound sources* to be delivered to the listener, considering the *listener* characteristics and the *environment* where sources are placed.

The model shows the four major components in an interactive VAS. The three components highlighted in blue contain the *a priori* knowledge and data of the system to be implemented, i.e. definitions and processes to handle the sources, the receiver (listener) and the medium (environment):

- *Sound sources* include all the information regarding the virtual sources that are part of the VAS, as the source stimuli, the spatial location, the orientation, the directivity and the level of the source. These sound sources include also the ones generated by the listener while they are interacting with the VAS.
- *Environment* includes the room specifications, as the geometry or the absorption coefficients of the air and all surface materials.
- *Listener* includes all the data regarding the listener and the headphones, from HRTF's databases, headphone characteristics, anatomical characteristics, etc. to the processes to personalize the HRTF or to track the listener position.

Finally, the *digital real-time processes* incorporate all methods that are performed in real time. The main part is the DSP (Digital Signal Processing), where the sound signal (green line) is processed to carry out the simulation of the spatialisation of the sound. The DSP has as input multiple information from the whole system, both from processes that also take place in real time, such as HRTF and source position calculations as well as previously stored information such as the specifications of the room or the IRIR characterization. The spatialisation process performed in the DSP can be classified as *direct sound simulation*, *distance simulation* and *room simulation*. As mentioned, there are other algorithms that should be processes in real time and correspond with the ones that manage the information of the listener and sound sources, such as HRTF interpolation, calculation of the position of the source regarding the listener, etc.

The different components of each group are presented in Figure 19. This figure shows a complete scheme of a state-of-the-art of a binaural renderer tool. Many components of this diagram are also included in the scheme presented in (Serafin et al., 2018), others have been specified from other studies such as the one presented by Sunder et al. (2015) and (Xie, 2013).

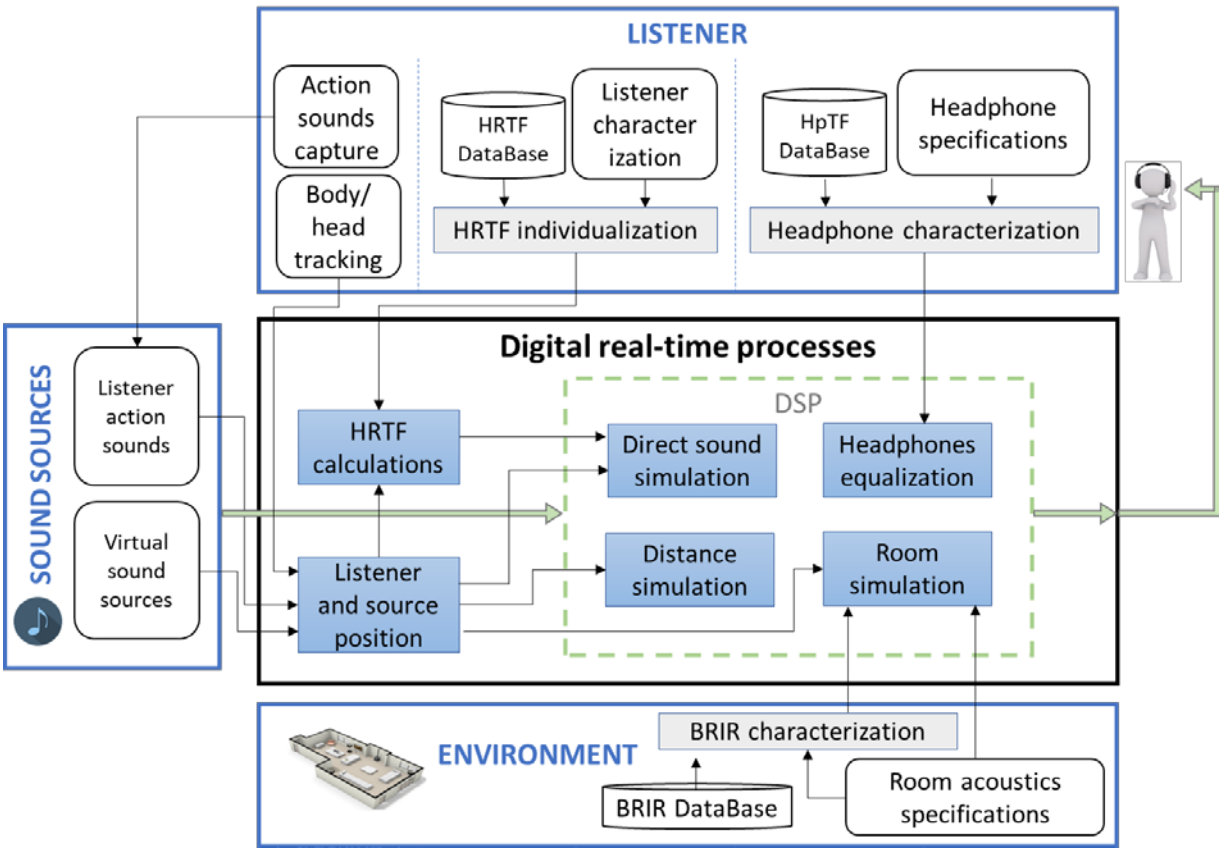


Figure 19. Diagram of a state-of-the-art of a binaural renderer tool

The components that form the diagram are of different types. Some of the components are data and information stores, indicated by the rounded corner boxes.

Others are offline processes that are carried out before the real-time rendering starts and are shown in the diagram as the grey shaded boxes. Finally, the real-time processes are represented by the blue shaded boxes. The data flow is represented with arrows. The following sections describe each component of the diagram in detail.

2.3.2 Sound sources components

These components contain all the sound sources that later will be rendered to create the VAS, virtual sound sources, which are generally pre-recorded digital audio, but can be synthesised sound as well, and sound produced by the listener's actions.

Usually, sources are represented as *omnidirectional point sources*, i.e., the sound is originated from a point in the scene and propagates equally in every direction. This representation is valid for many cases. However, some kind of sources, such as musical instruments, have radiation patterns that should be simulated as *directional point sources* and, therefore, their orientation is relevant. In addition, in the real word many sounds come from a large volume and can be considered as *volumetric sources*. Volumetric sources can be simulated as a combination of sounds from many directions and are used to simulate elements such as a river or wind blowing through the trees, where there is a broad soundscape (Schissler et al., 2016). Moreover, when a source is close to the listener, for example a guitar playing music, the sound is perceived as volumetric as this is emitted throughout the entire instrument and a point source can sound unnatural, as if the sound were coming only from the centre of the instrument.

Sound information can be managed using a paradigm called *object-based audio scenes* (Geier et al., 2010). Binaural systems usually follow this paradigm to transmit and store the spatial sound, where the objects are mainly the discrete sound sources that form the auditory space. These objects contain the sound source's stimuli and all the metadata describing location and other parameters relevant for its spatialisation. One of the main advantages is that this representation is more flexible in terms of the rendering methods used, i.e. the representation of the audio scene is always carried out in the same way, regardless of whether it is a binaural or a multi-channel system with several speakers. The problem with this approximation is that its feasibility is limited by the complexity of the sound scene, since rendering each source as a discrete object becomes impractical when the number of sources is very large.

Existing standards for broadcasting spatial audio address the need for suitable audio codification to store sound sources information in a more universal way. Jürgen Herre and colleagues proposed the standard for coding representation and rendering of spatial audio called MPEG-H 3D Audio (Herre et al., 2015). This codification supports the

object-based audio scenes and tries to facilitate the production, transmission and reproduction of audio material for immersive spaces.

2.3.3 Listener components

2.3.3.1 Listener actions and positions

In a dynamic VAS, the listener interacts with the environment in real-time, and, in order to know their *position and orientation,* a head/body tracking device should be used. Depending on the application, we may be interested in capturing the motion of the entire body of the listener or just the head. When full body tracking is required, optical trackers such as the well-known commercial systems Optitrack[9] or Vicon[10] can be used. This can be simplified to only track the head location and orientation and, this can be provided by an HMD, like Oculus[11] or HTC Vive[12], which are common devices in VR setups. A further simplification could rely on capturing just the orientation of the head. In this later case, inertial sensors attached to the head are a very cost-effective solution.

In addition, it should be considered that, the listener's movements, gestures and their interaction with the environment (*listener actions*) can cause new sounds, which must be incorporated into the set of sound sources to be rendered by the system.

2.3.3.2 HRTF individualization

A binaural renderer tool is based on the use of Head-Related Transfer Functions (HRTFs) and having a reliable HRTF data is decisive in this kind of systems. Despite this and the fact that numerous works have been carried out for the individualization and selection of the HRTF, it remains as a challenge. This section describes the most common “off-line” processes, i.e. processes that are not included in the real time processing of the system, to obtain the HRTF that will be used in the binaural rendering. More details about this topic can be found in the reference book about HRTF and VAS written by Bosun Xie (2013).

[9] https://optitrack.com/
[10] https://www.vicon.com/
[11] https://www.oculus.com/
[12] https://www.vive.com/

2.3.3.2.1 HRTF acoustical measurements

Individualization of the HRTF is a key for having a realistic virtual acoustic scenario for a specific listener (Xu et al., 2007). A typical HRTF acoustic measurement system is shown in Figure 20.

The listener is positioned in the centre of the structure, with a pair of in-ear microphones placed at the entrance of their ear canal (V.R. Algazi et al., 2001). The structure consists of one or several vertical arc-shaped array of loudspeakers, positioned at each direction of interest. During the measurement the subject must stay still, and a set of sweeps are played from the loudspeakers (Farina, 2000, 2007). The microphones record the played sound before it is modified by the listener anatomy. Logarithmic sweeps are the stimuli used by many systems. In this way, HRTFs are defined as (Xie, 2013):

$$H(\theta,\phi,f) = \frac{P(\theta,\phi,f)}{P_0(f)} \tag{2.1}$$

where $P(\theta,\phi,f)$ represents the sound pressure in the frequency domain captured by the microphones from the different directions at both ears and $P_o(f)$ the free field sound pressure in the frequency domain captured in the same position (at the centre of the head) but measured without head. Finally, the HRTF can be equalized in two different ways. One type of equalization consists in dividing the whole HRTF(θ,ϕ,f) by the one measured at a specific position HRTF(θ_0,ϕ_0,f), usually at the front, which is called free-field-equalized HRTF (Equation (2.2)); and the other one consists in dividing the HRTF(θ,ϕ,f) by a diffuse field average calculated as the root-mean-square value of HRTF magnitudes across all directions and called diffuse-field-equalized HRTF (Equation (2.3)) .

$$H_{free}(\theta,\phi,f) = \frac{H(\theta,\phi,f)}{H(\theta_0,\phi_0,f)} \tag{2.2}$$

$$H_{diffuse}(\theta,\phi,f) = \frac{H(\theta,\phi,f)}{\sqrt{\frac{1}{M}\sum_{i=0}^{M-1}|H(\theta_i,\phi_i,f)|^2}} \tag{2.3}$$

Although this is one of the most accurate measurement methods known to date, the results are highly dependent on the hardware and system configuration used (V.Ralph Algazi et al., 1999; Katz & Begault, 2007).

Measuring HRTFs in the previously described way, with high directional resolution for each individual is complex, very time-consuming and requires a specific and expensive equipment (Gardner & Martin, 1995). Many works have been carried out in order to measure the HRTF with a less-complex system. A very recent proposed system can be seen in (Jonas Reijniers et al., 2020), where a simplified method for an HRTF measurement in a home environment is presented. The system consists in a fixed loudspeaker, a head tracking system and a set of in-ear microphones. The listener carries out a set of unsupervised head movements in front of the speaker. The paper presents the differences between this dynamically obtained HRTF and the standard static one. In addition, Enzer and collegues (Enzer et al., 2013) presented a set of alternatives to HRTF measurements, considering near-fields, continuous HRTF measurements, etc. It is also worthy to mention the work of (Xu et al., 2007), where a review of research on HRTF individualization for VASs is presented.

Figure 20. Localisation measurement equipment from Rayleigh Laboratories and operated by ISVR Consultancy Services[13]. The photo on the left shows an HRTF measurement procedure of a real user and the one on the right of a manikin.

2.3.3.2.2 HRTF synthesis

Another interesting area of research within HRTF individualization is the HRTF synthesis. Despite many of the HRTFs used nowadays are the acoustically measured ones, several researchers in the past 20 years have attempted to efficiently and effectively synthesize HRTFs. A seminal work was done by Duda and Algazi on a spherical head models (V. Ralph Algazi et al., 2001; Duda & Martens, 1998), spherical model with torso, also called "snowman" model (R. Algazi et al., 2002; V. Ralph Algazi, Duda, et

[13] http://resource.isvr.soton.ac.uk/FDAG/VAP/html/facilities.html. Retrieved January, 2022

al., 2002) and adaptable ellipsoidal head models (Duda et al., 1999). Katz (2001b, 2001a) worked at HRTF synthesis using Boundary Element Method (BEM), obtaining HRTFs from 3D computational models, followed by the work of Kahana & Nelson (2007). The idea of these works is to acquire the 3D geometry of the listener, paying special attention to the pinna, by a 3D scanner or a 3D reconstruction technique from a photograph, and then carry out a numerical simulation of the propagation of acoustic waves around the listener to obtain the HRTF. A similar approach was followed by Fels & Vorländer (2009), which used BEM for investigating the respective contributions of head, pinna and torso on HRTF components. In addition, Kreuzer et al. (2009) presented a Fast Multipole Method (FEM) coupled with BEM to simulate HRTFs for a wide frequency range. Other approaches to calculate HRTF using wave equation simulations can be found in (Katz, 2001a). Spagnol et al. (2013) presented a model for real-time HRTF synthesis that, thanks to the studies carried out about the relationship between HRTFs and pinna reflection patterns, could customize the HRTF according to individual anthropometric data. A monograph with a collection of studies on HRTF individualisation, decomposition and modelling can be found in (Nicol, 2010). Finally, the review presented by Sunder et al. (2015) shows a comparison of various HRTF individualization techniques.

2.3.3.2.3 HRTF selection

Due to the difficulty of individualizing a HRTF for each listener of a binaural audio system, the use of what it is called a *generic HRTF* is very common. A generic HRTF is an HRTF measured in a manikin (Figure 20b). Two of the most used ones are the KEMAR (Knowles Electronics Manikin for Acoustic Research) dummy head and torso (Gardner & Martin, 1995) and the Neumann KU100 dummy head (Bernschütz, 2013). These manikins are based on worldwide average human head and torso dimensions, including a pinna which is acoustically and dimensionally similar to the human's pinna average dimension. Although many systems include this HRTF by default, the use of a generic HRTF in a binaural system results in a degradation of the listening experience in localization and realism (Xie, 2013).

A generally used alternative is to take an HRTF measured with another individual, from an available public HRTFs' database. From now on, we will call an HRTF measured from another individual as *non-individualized HRTFs*. A list with some public available databases is shown in Table 1.

The problem then becomes how to choose the HRTF from an existing database that best matches the listener's one. Several studies have been carried out in the past years to investigate this issue and, nowadays, there are multiple methods to select which, among the non-individual HRTFs, is the best fitting for a specific listener (*best-matching*

HRTF). This section describes a set of methods that can be classified into two categories: based on a*nthropometric data* of the listener and based on *psycho-acoustic experiments*.

Table 1. HRTF Databases public available

Databases	**Institution**	**Nº of Subjects**
LISTEN (Warusfel, 2003)	IRCAM	50
CIPIC (V.R. Algazi et al., 2001)	CIPIC Interface Laboratory	75
ARI (*ARI HRTF Database*, n.d.)	Austrian Academy of Sciences Acoustic Research Institute	132
RIEC (Watanabe et al., 2014)	Tohoku University	105
SADIE II (*SADIE | Spatial Audio For Domestic Interactive Entertainment*, n.d.)	Department of Electronic Engineering, University of York	20
FIU (Navarun Gupta et al., 2010)	FIU DSP	15
HUTUBS (Brinkmann et al., 2019)	Technical University of Berlin, Munich Research Centre and Sennheiser electronic GmbH & Co	96

The first group of methods, based on *anthropometric data*, are built on the idea that there is a set of anthropometric parameters which strongly influence the HRTF (M. Zhang et al., 2011). These methods select a HRTF by means of listeners' anatomical characteristics from a database that contains this kind of information. Zotkin et al. (2002, 2004, 2003) presented a novel way to obtain the HRTF from a database based on anatomical measurements. To select the best-matching HRTF, they presented an algorithm that calculates the difference between the listener anthropometric ear pinna parameters and those in the database, selecting the HRTF that minimizes the error. They used only seven morphological parameters measured on a picture of the listener's pinna, finding that the selected HRTF improved the localization performance. Lei & Xiangyang (2016) went one step further, using a correlation analysis between listener anthropometric features and HRTF to find the most appropriate HRTF. Finally, Yao et al. (2017) used a neural network (previously trained by listener's perception scores) to predict the best-matching HRTF based on input vectors of anthropometric measurements.

The second group, based on *psychoacoustic tests* for perceptual evaluations, is the most widely used. Within this group, we can find mechanisms to select an HRTF by rating the quality of different sounds rendered with different HRTF (Katz & Parseihian,

2012). In addition, these researchers performed another experiment where they used a binaural quality evaluation study to propose a global perceptual distance metric to describe HRTF and listener similarities, assessing the quality of sound trajectories of a set of HRTFs (Andreopoulou & Katz, 2016b). Another example can be found in (Iwaya, 2006), where the authors proposed a tournament-style listening test called DOMISO (Determination method of OptimuM Impulse-response by Sound Orientation). In that test, the listener had to listen to a set of sounds following various trajectories and select the one that better resembled the a given described trajectory. Finally, Seeber et al. (2003) proposed a two-step method where subjects had to select an HRTF that best matched a set of criteria. They evaluated the results through a localization experiment, where subjects were asked to report the location of a virtual sound source. Then, errors in localizations were measured as an estimation of performance, this being an objective indicator of how well the corresponding HRTF works.

The use of a non-individual HRTF involves a degradation of localization performance and realism. Many works discuss how listeners can localize virtual sounds using non-individual HRTFs (Adelbert W. Bronkhorst, 1995; Møller et al., 1996; Elizabeth M. Wenzel et al., 1993; Xu et al., 2007). They agreed that non-individual HRTFs increase elevation errors and front/back confusions. However, there are studies that suggest that HRTF adaptation improves the performance of those HRTFs. As Nicol (2010) stated, «the auditory system is able to modify its spatial decoding to learn the spatial mapping of another individual». Works, as the one presented in (Hofman et al., 1998), show that the human auditory system can adapt to localise sound sources using a non-individual HRTF. They compared HRTFs with different spectral elevation cues, obtained measuring the HRTF with pinna modifications by moulds. They discovered that, although the performance of localization of sound elevation was completely deteriorated immediately after the modification, an accurate localization performance was eventually recovered. Katz and colleagues carried out several works on this topic (Blum et al., 2004), confirming that quick adaptation to non-individualised HRTFs is possible through a controlled learning environment, with audio-kinesthetics interactions (and no visual cues) within an auditory scene. In a more recent work, Steadman et al. (2019) showed that, after a reduced number of training sessions, a significant effect of adaptation to a non-individual HRTF was appreciated, which was also retained across multiple days.

2.3.3.2.4 The Spatially Oriented Format for Acoustics (SOFA)

Once the HRTF has been obtained, it should be stored in order to use it later in a binaural spatialisation renderer. One of the most used formats is the Spatially Oriented Format for Acoustics (SOFA) (Majdak et al., 2013). SOFA is a data exchange format for reading, saving, and describing acoustic measurements, including HRTFs, in a general way, allowing to store also more complex data, e.g. BRIRs or multichannel

measurements such as those created with microphone arrays. In addition, the developers provide APIs for reading and writing the data in SOFA[14]. This format is in continuous development and it is already used by many HRTF databases, as CIPIC database (V.R. Algazi et al., 2001) or *ARI HRTF Database*. The AES standardization committee is working in a project which, building upon SOFA, will standardize an HRTF file format. There are other formats commonly used to store HRTFs, such as a plain text or Matlab (Mathworks, Inc.) file format.

2.3.3.3 Headphone characterization

Headphones used to deliver the binaural rendered signal to the listener are not acoustically transparent, since they produce spectral colorations and modify the timbral quality of the source, which can deteriorate the perceptual plausibility of the sound (Lindau et al., 2007). To solve this problem, an equalization filter can be used to compensate the effect of the Headphone Transfer Function (HpTF). This headphone equalization filter (HpEq) is obtained in the off-line process called *Headphone characterization* (Figure 19), designed to cancel those effect of the HpTF (Pralong & Carlile, 1996). The use of HpFT is highly dependent on the listener ear anatomy (as each headphone coupled differently depending on the ear), and on the headphone transducer response. Using individual HRTF and HpFT improves source localisations in VAS (Møller et al., 1995). For this same reason, it is reported that the non-use of HpEq can aggravate the front-back confusions and hinder elevation perception (Xu et al., 2007).

However, specifying a compensation filter for cancelling the HpTF is not easy and is highly dependent on the position of the headphones on the ear, as studied by Kulkarni & Colburn (2000). They also observed that a bad equalization may become worse than no equalization at all. Florian Völk (2014) studied the inter-individual differences and the intra-individual variability due to repeated positioning of the headphones. The study suggested that those fluctuations can influence the spatial hearing, especially for sounds in the frequency range above 6 kHz. In (Schärer & Lindau, 2009), an evaluation of different HpEq designs techniques was carried out, where several HpTF were measured on a manikin in multiple positions and objective and subjective evaluations were performed. The study revealed several aspects that should be taken into account as the individual headphone calibrations and also the adequate selection of the headphones.

[14] http://www.sofaconventions.org/. Retrieved January, 2022.

The headphone equalization filter (HpEq) is the input of the last component of the Digital real-time processes group, the *Headphone equalization*. This component performs the equalization of the binaural signals just before delivering it to the headphones, applying the HpEq to each channel (left and right).

2.3.4 Environment components

This component presented in Figure 19 contains all the information and data regarding the environment, such as the room geometry and the absorption characteristics of surfaces materials and air, or the BRIR in case we want to use it to simulate the environment.

As presented in the Introduction, to perform a binaural simulation of a source located inside a room, we can make use of the BRIRs (Binaural Room Impulse Response) (Jeub et al., 2009). Those functions encode the impulse response of the room and the listener. In contrast to HRIRs, BRIRs depend not only on the relative position of the source to the listener, but also on the position of the source and the listener in the environment. BRIR can be acoustically measured using a dummy head or a human subject within the room that will be simulated, as the database presented in (Jeub et al., 2009), where recordings took place in a studio booth, an office, a meeting room and a lecture room with a manikin head[15]. Another study that measured and used BRIRs was the one performed by Shinn-Cunningham et al. (2005). In this case, BRIRs were measured in a classroom for sources at distances under 1m. Recording BRIRs for any combination of listener and source position within a room can become impractical. A solution can be found in using synthetised BRIRs, based on room models (Borß & Martin, 2009). However, these functions tend to be very large, depending on the size of the room, and their storage will require a lot of capacity. For example, a BRIR for a sound source in a specific position inside a small room has 12000 samples at 44.1kHz, a medium size room 30000 samples at 44.1kHz, a large size room 55000 samples at 44.1kHz and a huge size room 177000 samples at 44.1kHz.

As mentioned, the BRIR for the specific position of the source relative to the listener is needed and, in the same way as the HRIR, interpolations are often carried out to get the BRIR for a specific direction. Garcia-Gomez & Lopez (2018) and Bruschi et al., (2020) presented a similar method for performing interpolation based on decomposition of BRIRs in time and frequency domain and algorithms for peak detection and early reflection matching, followed by an interpolation. Lindau et al. (2008) studied, with a

[15] Database link: http://www.ind.rwth-aachen.de/~bib/jeub09a. Retrieved January, 2022

listener test, the minimum spatial resolution of a BRIR dataset required for a dynamic binaural synthesis without audible artefacts, obtaining a grid of 2° x 2° for horizontal and vertical head movements for pink noise and 5° x 5° for a musical stimulus.

There are other types of methods used for rendering reflections in a binaural VAS which can be classified into two categories: *physics-based* algorithms and *delay networks* methods (Välimäki et al., 2012), and will be presented in Section 2.3.5.4 Room simulation. These methods required information about the environment, such as: size and geometry of the room, materials of the floor and walls, air absorption coefficients, etc. These data are stored by the *room acoustic specification* components and is used both by the component that is responsible for the characterisation of the BRIR and directly by the algorithms implemented in the Room simulation component.

2.3.5 Digital real-time processes

All the components described above store the necessary information to be able to perform the spatialisation of the sound, as well as carry out a series of processes that can be done off-line to increase efficiency and saving time during the execution of the system. However, there are other processes that must be executed in real time, especially those that are related to the position of the source and the listener, since we are working with an interactive system.

Within these processes that take place in real time, two groups can be distinguished, those that process information related to the source and the listener, such as *HRTF calculations* and *listener and source position,* and the ones that performs the *Digital Signal Processing* (DSP) that work with the audio stream and use all the information provided by the rest of the system processes. The DSP processes the signal using a stereo audio stream, discretizing both left and right channels. Typical sampling frequencies are 44100 Hz or 48000 Hz. In this way, for a real-time processing, the audio stream is sequenced in buffers with typical sizes of 256, 512 or 1024 samples. Using all the information regarding the source, the listener and the environment, this component carries out the rendering of spatialised sound sources, through the simulation of the direction of the direct sound, the reflected sound inside the room, the distance of the sound source and the equalization of the headphones used to deliver the sound to the listener. Processes that are performed in real time are described in the following sections.

2.3.5.1 HRTF calculations

As mentioned, HRTFs are measured for just a specific set of positions while the subject stays still. In a dynamic VAS, the source can be located at any place of the 3D space and the user is constantly moving and interacting with the scene. This means that the HRTF corresponding to a given position and pose of the listener might not be available in the measured set. To be able to simulate all these facts, the system can perform a set of run-time calculations which are presented in the next sections as HRTF interpolations, HRTF corrections for sources placed at the near field and HRTF corrections for different head-above-torso orientations.

2.3.5.1.1 HRTF interpolation

HRTF is generally obtained at discrete and finite directions, where the spatial resolution of measurements is usually 5 degrees but can be also lower. Since the localization accuracy for sources in the horizontal plane is approximately ±1.5 degrees for people with normal hearing (Jens Blauert, 2013), there is a need for HRIR reconstruction at unmeasured positions. Directional interpolation methods calculate the unknown HRIR using the nearby measured HRIRs values following different interpolation approaches (Xie, 2013). In addition, interpolation is necessary to simulate a continuous movement of sound sources and avoid audible jumps in the perceived sound location. Elizabeth M. Wenzel & Foster (1993) showed that 3D audio simulation for moving sources and listener was best achieved computing interpolations with minimum-phase HRTFs, to reduce dynamic comb-filtering effects.

HRIR interpolation methods are a widely investigated matter in binaural spatialisation and many approaches have been implemented and tested, from time-domain (Sodnik et al., 2005) to frequency-domain interpolation (Nishino et al., 1999), looking also at decompositions based on principal component analysis (Carlile et al., 2000) and spherical harmonics (Romigh et al., 2015). HRTF interpolation will be addressed in more detail in Section 3.5.2.

2.3.5.1.2 HRTF corrections for near field sources

As presented in the Introduction chapter, ILDs (the level differences in the signals at the two ears) appear at high frequencies due to the acoustic interference between the sound and the listener head, which is called the *head shadow* effect. In the far field, when distance is large (larger than 2 meter), this effect can be considered as only directional-dependent, i.e., the effect provides directional spatial information but is invariant with the source distance. However, when a source is in the near field, changes in the distance of the source cause changes in the ILD across all frequencies (Duda & Martens, 1998).

Various studies led by Brungart studied this issue, investigating the spatial perception and changes within HRTFs of nearby sources (D. S. Brungart & Rabinowitz, 1999; D.S. Brungart, 1999; D.S. Brungart et al., 1999). They suggested that binaural cues play an important role in auditory distance perception in the listener's near field. They measured HRIRs in distances shorter than 1 meter with a KEMAR manikin and carried out a set of experiments. Their results indicated that ILD increases substantially for lateral sources as the source approaches the head, however, ITD remains almost invariant to the distance. One of the experiments, focused on the stimulus effect, also suggested that ILD for low frequencies are the dominant auditory distance cue in the near fields. Finally, the elevation-dependent characteristics of the HRTFs (which can be considered as monoaural cues) were not strongly dependent to distance.

Shinn-Cunningham (2000) also defined how ILD behaves with distance. They observed that significant changes happen for lateral sources, where they found that the ILD varies over 20 dB as distance varies from 15 cm to 1 m to the head. When the source is more than 1 m away, changes in distance cause no significant changes in ILD, being ILD changes at these distances caused only by the direction-dependent head shadow effect. These works were followed by studies from Lentz and colleagues looking at near-field HRTF synthesis, and defining the limits of noticeable differences between near-field and far-field HRTFs, being around 1.5m for lateral sources and around 0.4m for front sources (Lentz et al., 2006).

Due to the aforementioned characteristics, HRTFs in the near field have some particularities that must be taken into account. Some studies and databases offer HRTF measurements specifically for the near field. Nishino et al. (2014) measured the near-field on an artificial head at distances 0.15, 0.20, 0.25 and 0.3 m, but just for the horizontal plane. Hosoe et al. (2005) measured the HRTF at 9 different distances from 0.2 to 1 m using a dodecahedral loudspeaker system. In (Yu et al., 2018) a database of near-field HRTF measurements in real subjects can be found. However, near-field HRTF measurements require a heavy workload, since HRTFs should be measured not only at many different directions but also at different distances. In addition, special equipment is needed to reproduce the sound as a conventional loudspeaker cannot be used (Xie, 2013). That is why most existing HRTF databases are measured at a single distance between the listener and the source, lacking data for near-field simulations. That is the case of those presented in Table 1. In this way, some works propose HRTF corrections to simulate sources in the near field using HRTFs measured in the far field as a basis. Romblom & Cook (2008) presented a work about how to compensate measured HRTFs for near-field simulations, based on the utilization of geometric head models from (Duda & Martens, 1998). These works and how to modify HRTF to simulate sources in the near-field will be described in detail in Chapter Section 3.5.4.

2.3.5.1.3 Head-above-torso orientation (HATO)

HRTF is usually measured for fixed head-above-torso (HATO) angles since the listener must stay still during the measurement process. The impact of the head-above-torso orientation has been investigated in different studies. Guldenschuh et al. (2008) presented a detailed research on this topic and provided 10 different measurements to assess the influence of the HATO on the horizontal plane. They extracted the influence of the HATO from the HRIR, denoting it as the torso related impulse response (TRIR). They realized that only a few configurations of sound source and head-torso orientations cause strong TRIRs, that correspond to reflections that lie slightly behind the direction where the shoulder is pointing to. They modelled the TRIR using those "strongest TRIR" and then interpolating. They suggested that adding the TRIRs to the HRIRs produces differences which are more noticeable in dynamical situations. In addition, a more recent study by Brinkmann et al. (2015) showed that taking into account HATO produces audible differences and listener achieves better binaural spatializations. They evaluated the influence of HATO on HRTFs, suggesting that there is a comb-filter effect caused by shoulder reflections, which was found to be most prominent if sound source, shoulder, and ear were aligned. They also affirmed that deviations in ITDs and ILDs could be neglected since they were below the threshold of audibility. Finally, V. Ralph Algazi, Avendano, et al. (2002) presented a work where they found that HATO related cues have an impact on the perception of elevation for sources outside the median plane for low-frequencies.

2.3.5.2 Direct sound simulation

In a binaural rendering tool, the direct path is simulated by filtering the virtual sounds signal with the HRIR of the specific position of the source. This filtering can be performed either in the time domain or in the frequency domain, both demanding a large number of operations, being the frequency domain operation the most efficient one. A very common method is to employ Fast Fourier Transform (FFT) algorithms, where the convolution of two signals of length N is carried out in the frequency domain by a multiplication of their FFTs and the application of the inverse FFT to the result. This makes a big difference, particularly when N is large, since the cost of the convolution in the time-domain is proportional to $O(N^2)$[16], and using the FFT-based convolution it is reduced to $O(N \log_2 N)$.

[16] We indicate the order of the function with the mathematical notation "Big O" (https://en.wikipedia.org/wiki/Big_O_notation, retrieved May 2022)

For simulating direct sound in a real-time environment is also important to take into account the *propagation delay* between the sound source and the listener. With a delay line we can simulate the behaviour of sound during its propagation through the air. This delay line is algo related with the *Doppler effect* since it can be simulated using a variable delay line. The doppler motion is commonly known as the effect of the sound pitch changing when a speeding object passes the listener. When the distance changes, so does the length of the delay, which also leads to a compression or expansion in time of the signal and, therefore, a change in the pitch of the sound. This is the basis of the Doppler effect, which can be perceived when the sound source or the listener are approaching or going away from each other at certain speed.

2.3.5.3 Distance simulation

Distance perception is an important aspect of spatial hearing, nevertheless it always receives less attention than the directional component. There are some distance cues that should be considered to render binaural audio. According to Shinn-Cunningham (2000), potential distance cues include overall level, reverberation and the ILD. These are the most important cues, although other cues have been identified, as spectral changes with distances (Coleman, 1968; Little et al., 1992). A more recent review of distance cues can be found in (Kolarik et al., 2016).

The *overall level* cue responds to the fact that the sound level decrease when the distance increase, according to what it is called the *inverse square law,* where the sound pressure generated by a point source in the free field is inversely proportional to the distance between the source and the listener. That is, a doubling of the distance from a source reduces the sound pressure level in 6 dB. This method is considered as an effective cue for distance perception and can be applied at every distance.

In addition, the judgment of the distance of a sound is also influenced by the *stimulus familiarity* (Kolarik et al., 2016). When a sound is perceived, listeners compare the sound level and the spectral content of the sound with what they know about the specific sound stimulus that they are hearing. For example, if a car horn is received with a low level, the listener will think that it is far away, since horns usually sound loud. Familiarity and the level of the signal are used in combination to perceive distances. For instance, a whispered speech with high level is associated with very nearby sources, while shouted speech has relatively more high-frequency energy than conversational speech and judged as far away source if the level of the signal is low. Several studies have demonstrated that distance estimations of familiar stimuli can be more accurate (Zahorik, 2002).

Regarding reverberation, the level of the signal reaching the listener ear decreases at a different rate than the direct sound. The propagation of the direct path expands spherically, where the power is distributed over the surface of the sphere, decreasing with the square of the distance (hence 6 dB), as mentioned before. However, this does not happen to the reflected path in the environment, which varies only slightly with distance. The reverberant sound is a collection of complex reflections, which depend on the size of the room and the acoustic properties of the reflection surfaces of the room (Zahorik, 2002). In this way, the *ratio between direct and the reverberant sound* becomes an important cue for the distance perception (Begault, 1994). How reverberation and direct-reverb ratio change with the distance is described also in (Adelbert W. Bronkhorst & Houtgast, 1999; Little et al., 1992; Shinn-Cunningham, 2000; Shinn-Cunningham et al., 2005).

Finally, when the source is located at less than 2 m, the interaction with the head must be considered. These alterations can be reflected in the HRTF, more specifically in the ILD, used to simulate the sound source, in the way that it was described in the previous section. However, the HRTF does not usually include measurements for these source distances and therefore, to carry out simulation of sources located in the near field (distances lower than 2m), a series of corrections on the HRTF should be performed (described previously in Section 2.3.5.1.2). In addition, sound sources placed at very far distances (larger than 15 meters) suffer a high-frequency attenuation caused by the air absorption, which acts as a low-pass filter modifying the spectrum of the sound.

A summary of the research status on auditory distance perception is presented in (Zahorik et al., 2005). In Section 3.3 these distance simulation cues are described deeply.

2.3.5.4 Room simulation

How enclosed spaces affect spatial hearing perception has been explained in Chapter 1, Introduction. The most relevant research milestones in this area were presented in Section 2.1.1 Binaural localization of 3D sounds. This section will present different techniques to render binaural audio within a reverberant room. An extensive description can be found in Chapter 11, “Binaural Room Modelling”, of (Xie, 2013), and one of the most relevant books in this area is “Room Acoustics” from (Kuttruff, 2016). In addition, an important review on this topic can be found in (Välimäki et al., 2012), titled “Fifty Years of Artificial Reverberation”.

Methods to perform room simulation can be classified into two groups: convolution-based methods and algorithmic reverb. Regarding convolution-based methods, analogous to the HRTF, for enclosed environments, the convolution in real-time of the input signal

with the BRIR allows to give a listener the impression of hearing the sound source within the room which is characterised by the BRIR. The duration of a BRIR depends on the size and shape of the virtual room and can often have a duration of several seconds. Therefore, using BRIR to render a virtual space requires additional computational power and a large memory space to store a high number of BRIRs to simulate multiple directions. Some methods have been designed to reduce this problem, which will be described in Section 3.4.

Algorithmic reverb simulations include several methods that can be classified in two categories: *physics-based* algorithms and *delay networks* methods (Välimäki et al., 2012). *Physically-based algorithms* try to simulate the physics behind sound propagation and reflection, aiming to reproduce the acoustics of the simulated room. According to physical principles, techniques for room acoustic modelling can be divided into two groups: *ray-based methods* and *wave-based methods*. *Ray-based* methods stand on the idea that sounds propagate in rays, following the rays' rules of reflections. These methods are recommended for high frequencies and smooth boundary surfaces. *Wave-based* methods aim to numerically solve the wave equation of the sound inside the room, using methods such as boundary element methods or finite-difference time-domain. These methods are more accurate but also demand more computational power and are feasible only for low-frequencies and small room modelling. The wave-based methods typically provide impulse responses directly while ray-based methods either produce time-energy responses or information of reflection paths that need to be converted to an impulse response. For more details and works related with these methods see (Välimäki et al., 2012; Xie, 2013).

The second category includes the approaches based on *networks of delay lines and digital filters*. In these approaches the input signal is delayed, filtered and fed back along several paths according to some precalculated room acoustic attributes or parameters. They can be classified as *perception-based methods* as they trust in the perception, rather than physical models (Xie, 2013). They use artificial delay methods and reverberation algorithms such as comb filters and all-pass filters, also known as Schroeder reverberation (Schroeder & Logan, 1961), Digital Waveguide Networks (Smith, 1985) and Feedback Delay Network reverberations (Rocchesso & Smith, 1997). These methods are less complex regarding computational power and can be used when a precise simulation is not necessary. Additional examples and methods of this category can be seen in (Välimäki et al., 2012, 2016).

To reduce the computational power required by simulating the previously described methods, many applications carry out an approximation where the virtual room is acoustically treated as a *shoebox-type room* (Schimmel et al., 2009). In this way, the room is approximated as a box and with no objects inside. A "shoebox" room is modelled

as a 3D space bounded by six rectangular faces (walls). Each wall can be independently defined regarding the material to simulate reflections on them, i.e. absorption and scattering coefficients can be configurable per wall.

Finally, and as mentioned in the Introduction, the sound reflections in an enclosed space can be divided in two groups, early reflections and late reflections. These two groups can be processed separately. Considering the characteristics of the room, the required level of accuracy and the amount of resources available, different previously described approaches can be applied to each group (L. Savioja et al., 1999). For early reflections ray-based methods are very used, while late reflections use to be simulating using recursive filters as feedback delay networks.

2.4 Existing tools to render binaural audio

There are many tools available for binaural spatialisation. Some of them have been implemented by important research groups in the field of spatial hearing and others by big companies in the virtual reality industry, such as Google or Meta (formerly Facebook). In this section, the most popular and representative open-source and commercial tools are described. All these selected tools are distributed as a software library implementation to create 3D audio for dynamic VAS by creating spatialisation effects of a sound source in real-time, and producing signals to be delivered through headphones (some of them include also algorithms to reproduce the audio on loudspeakers). There are as many implementation approaches, feature sets and license schemes as there are tools. The following tables are an attempt to make a comparison between the different tools, focusing on the functionalities and features of a binaural renderer presented in section 2.3 Binaural rendering.

Table 2 shows the list of the tools sorted by the year of the first release. This table shows in the first column the name and the reference where the documentation of the tool can be found, usually the article for the open-source tools and the website or manual for the closed ones. The second column indicates the author, company or community that has developed the tool. Then, the third column provides the website[17] with more information regarding the tool. The fourth and fifth columns provide the license and the availability of the code (the language of the library, the available plugins, the repository or web page where the tool can be download or purchase). Finally, the last column includes some additional information of interest.

[17] Retrieved July, 2021

Table 2. Real-time binaural spatialisation toolkits[18]

Toolkit	First release	Author/ Community/ Company	Website	License	Availability	Additional information
SPAT (Carpentier, 2018; Carpentier et al., 2015)	1995	IRCAM	https://forum.ircam.fr/projects/detail/spat/	closed-source Commercial	Max/MSP, Repository: https://git.forum.ircam.fr/beller/spat *Latest release: v5.1.7 (Jun 2020)*	First time presented in 1995 by "Le Spatialisateur". Improved along the years, it is still in use. In 2017 IRCAM, together with Flux, created SPAT Revolution (https://www.flux.audio/project/spat-revolution/)
Slab 3D (Miller & Wenzel, 2002)	2001	NASA	http://slab3d.sourceforge.net/	open-source NASA Agreement	C++ Library. Code Download: http://slab3d.sourceforge.net/downloads.html *Latest release: v6.8.3 (Aug 2018)*	Originally developed at the Spatial Auditory Displays Lab, NASA Ames Research Center. User Manual v6.8.3: http://slab3d.sourceforge.net/slabtools_user_manual.pdf
OpenAL Soft (Wu & Yu, 2016)	2007	Author: "kcat" https://kcat.strangesoft.net/	https://openal-soft.org	open-source LGPL	C/C++ library. Repository: https://github.com/kcat/openal-soft *Latest release: v1.21 (Jan 2020)*	Software implementation of the OpenAL 3D audio API (http://openal.org/). It was forked from the open-sourced version available originally from the obsolete *openal.org* SVN repository.

[18] This table was last edited on 30 July 2021

Rapture 3D (Blue Ripple Sound, 2016)	2008	Blue Ripple Sound Ltd.	https://www.blueripplesound.com/products/rapture3d-universal-sdk	closed-source Commercial	Native C/C++ integration, Unity, OpenAL driver. *SDK Latest release: v3.2.3 (Jul 2019)*	Blue Ripple Sound provides multiple products to create 3D audio, including a *Binaural Surround Plugins* that convert channel-based mixes into binaural 3D audio and *Rapture 3D Universal SDK* that supports HRTF and head tracking.
Csound for Binaural Processing (Lazzarini & Carty, 2008)	2008	Csound	http://www.csoundjournal.com/issue9/newHRTFOpcodes.html	open-source LGPL	Opcode/C. Repository: https://github.com/csound/csound/tree/develop/Opcodes *Latest commit on Mar 2018*	Csound was originally developed by Barry Vercoe in 1985 at MIT Media Lab. Nowadays it is a community of volunteers that contribute with the system. The Audio3D module is included in the opcode module (as *hrtfopcodes.c*)
Soundscape Renderer (Geier & Spors, 2012)	2008	Deutsche Telekom Lab and Universität Rostock	http://spatialaudio.net/ssr/	open-source GNU	Code repository: https://github.com/SoundScapeRenderer/ssr *Latest release: v0.5.0(Nov 2018)*	This tool provides different rendering algorithms: WFS, HOA and binaural techniques.
RealSpace ™ 3D Audio (RealSpace 3D, 2015)	2013	VisiSonics Corporation	https://realspace3daudio.com/	closed-source Educational and Commercial	Unity, Wwise, Unreal. Download: https://realspace3daudio.com/purchase-unity-plugin/ *Latest release: 2017.1.2.6361 (wwise authoring plugin) and 2019.2+ (wwise-unreal)*	Based on 10 years of research in the University of Maryland. In 2014 it was licensed by Oculus VR company (https://www.umventures.org/news/visisonics-realspace-3d-audio-software-licensed-oculus-virtual-reality)
Steam Audio (SteamAudio, 2014)	2014	Valve	https://valvesoftware.github.io/steam-audio/	closed-source	Unity, Unreal, C API, FMOD, Wwise. Repository:	The tool name changed from Phonon to Steam Audio when it was bought by Valve.

				Free (under a license agreement)	https://github.com/ValveSoftware/steam-audio *Latest released: 2.0-beta.18 (Apr 2020)*	See also: https://steamcommunity.com/games/596420/announcements/detail/521693426582988261
Oculus Audio Spatilizer (Oculus VR, 2020)	2015	Oculus VR (acquired by Facebook in 2014)	https://developer.oculus.com/audio/	closed-source CC 4.0	Native C/C++, Unity, Unreal, FMOD, Wwise, Digital Audio Workstation (DAW) *Latest release of Oculus Spatializer Native: v18.0 (Jul 2020)*	SDK to integrate 3D audio in a VR environment. See the available packages here: https://developer.oculus.com/downloads/audio/
Facebook 360 Spatial Workstation (Audio, 2020)	2015	Facebook	https://facebook360.fb.com/ https://facebookincubator.github.io/facebook-360-spatial-workstation/Documentation/SDK/Audio360_SDK_GettingStarted.html	closed-source Royalty-free copyright	FB360 Spatial Workstation and Rendering SDK (Unity package, Cross-platform C++ library and Android Java API). Download: https://facebook360.fb.com/spatial-workstation/#s2 *Latest release of Spatial Workstation: v3.3.3 (May 2020) and Rendering SDK: v1.7.12 (Dec 2019)*	This tool was initially developed by Two Big Ears and later bought by Facebook. Its engine, 3DCeption (now called audio360) dates from 2014. The *Spatial workstation* workflow can be seen in https://facebookincubator.github.io/facebook-360-spatial-workstation/KB/SpatialWorkstationWorkflow.html#spatial-workstation-workflow

Resonance audio (Gorzel et al., 2019)	2016	Google	https://resonance-audio.github.io/resonance-audio/	open-source Apache 2.0	Unity, Unreal, FMOD, Wwise, Web, DAW, Android, iOS, C++ and MATALB. Repository: https://github.com/resonance-audio *Latest release: see Release section of each repository in the previous link.*	A developer overview of the tool can be seen in https://resonance-audio.github.io/resonance-audio/develop/overview
MS HRTF Spatializer and Microsoft Spatializer (Sound, 2020)	2016	Microsoft	https://docs.microsoft.com/es-es/windows/mixed-reality/spatial-sound-in-unity	closed-source MIT	Unity. Microsoft uses it for different products. Repository: https://github.com/microsoft/spatialaudio-unity *Latest release: v1.0.18 (Jun 2020)*	The MS HRTF Spatialiser is a spatial audio plugin created by Microsoft as part of their Mixed Reality Toolkit (https://docs.microsoft.com/en-us/windows/mixed-reality/spatial-sound). In 2019 Microsoft launched The Microsoft Spatializer, to include 3D audio on the HoloLens 2.
Virtual Acoustics (Vorländer et al., 2010) (Aspöck et al., 2019)	2016	Institute of Technical Acoustics at RWTH Aachen University	http://www.virtualacoustics.org/overview.html	open-source GNU and Apache License v2.0	C++ libray, Matlab and Unity. Download: http://www.virtualacoustics.org/download.html *Latest release: VA_full.v2020a.win32-x64.vc12*	VA is a real-time auralization framework developed for scientific research that also provides modules and interfaces for experiments and demonstrations.
IEM Plug-in Suite	2017	Institute of Electronic Music and Acoustics from	https://plugins.iem.at/	Open-source	The plug-ins are created with the JUCE framework	The IEM Plug-in Suite is an audio plug-in suite including Ambisonic plug-ins up to 7th order and

(Schörkhuber et al., 2018)		University of Music and Performing Arts, Graz, (Austria)		GPLv3 License	Repository: https://git.iem.at/audioplugins/IEMPluginSuite/-/tags/v1.13.0 *Latest Version v1.13.0 (November 2021)*	
Anaglyph (Poirier-Quinot & Katz, 2018)	2018	Project supervised by Dr. Brian Katz.	http://anaglyph.dalembert.upmc.fr/	closed-source CC BY-NC-ND 3.0	DAW (VST audio plugin). Download: http://anaglyph.dalembert.upmc.fr/index.html#download *Latest release: v0.9.4 (May 2020)*	This tool integrates the results of over a decade of spatial hearing research. It was developed in the context of numerous academic research projects with the CNRS and Sorbonne University.
SOFAlizer **(Jenny et al., 2018)**	2018	Project supervised by Dr. Piotr Majdak	https://github.com/sofacoustics	open-source European Union Public License 1.1	Unity, VLC and pureData. Code Repository: https://github.com/sofacoustics *Latest Unity release: v1.2 (Mar 2019)*	A spatialisation engine allowing to use HRTFs stored in SOFA format and implement the convolution.
SPARTA Binauraliser (Mccormack & Politis, 2019)	2021	Acoustics Lab, Aalto University, Finland,	https://leomccormack.github.io/sparta-site/	open-source MIT License	VST audio Plug-ins Repository: https://github.com/leomccormack/sparta-site	SPARTA (Spatial Audio Real-Time Applications) is a collection of VST audio Plug-ins for producing spatial sound scenes.

Table 3 summarizes the different features of each spatialisation tool. The next sections describe each of these features, highlighting some special characteristics of each tool. Regarding the table, the second column indicates the main technique used to simulate the direct (anechoic) path. The third column shows additional techniques implemented by the tool to also simulate the direct path, but usually with less computational power cost. Next columns indicate if the tool allows the HRTF file import, the HRTF and/or ITD customization, the HRTF and/or ITD interpolation, the near field simulation and distance simulation. Finally, the prior-to-last column indicates the technique used to carry out the room simulation, and the last column shows information regarding the source settings allowed by the tool. Regarding the closed-source tools, in some cases it was not possible to gather detailed information about which approach/algorithm/technique they use, as they are not reported in the available documentation. In some cases, it was possible to infer this information from comments in the API code and screenshots of the GUI. When no information could be found about a specific feature, the abbreviation NR (not reported) is used in the corresponding cell.

Table 3. Real-time binaural spatialisation toolkits features[19]

Toolkit	Direct sound	Additional techniques for direct sound	Import HRTF	HRTF or ITD customization	HRTF or ITD interpolation	Near field	Distance	Reverb sound	Source configuration
SPAT	HRTF-based	HOA, stereo techniques, VBAP in 2D and 3D and more (see Carpentier et al., 2015)	SOFA format	No	HRTF, see (J.-M. Jot et al., 1995)	Yes	Level attenuation and air absorption	The reverb simulation consists of three temporal segments: few early reflections, a set of dense late reflections and late reverb tail. Different implementation options: - with delay lines, different for each segment, - convolution-based (parametric convolution) - hybrid: convolution reverb for the early reflections and a FDN for the reverberation tail	Source presence, warmth, brilliance, and radiation (aperture and orientation of the radiation pattern).
Slab 3D	HRTF-based	No	Own format	Stored separately	HRIR and ITD, spline interp.	No	Spherical spreading loss and FIR filter for far dist.	- Rectangular room with materials. - Ray-based model. Up to 6 early reflections. - Late reverberation not supported	Sound pressure level, waveform and source radius

[19] This table was last edited on 30 July 2021

OpenAL Soft	HRTF-based	stereo, 4-channel, 5.1, 6.1, 7.1 and B-Format	Own format and SOFA	Stored separately	HRIR and ITD, bilinear interp.	Near-field control filters	Configurable attenuation curve	Available through a 'standard reverb' and 'EAX extension', with configurable early and late reflections and FDN techniques with configurable parameters (gain, delay, etc.), see example in https://github.com/kcat/openal-soft/blob/master/examples/alreverb.c)	Point sources (with directional sound)
Rapture 3D	Virtual Ambisonics (HOA)	Stereo panning	SOFA format	NR	NA	NR	Level attenuation (configurable rolloff factor). Configurable speed of sound	Directivity controls and occlusion simulation. O3A reverb includes separate simulation for early reflections, convolution-based reverb and shoe box model reverb simulation (https://www.blueripplesound.com/products/o3a-reverb)	Point sources
Csound Binaural Processing	HRTF-based	Woodworth based SHM and Virtual Ambisonics	No	No	HRTF and delay linearly interp.	No	No	Dynamic FDN based diffuse-field reverberator with early reflections in a parametric room.	Point source
Soundscape Renderer	HRTF-based	WFS, VBAP, Ambisonics and Amplitude panning	WAV format	No	No	Near-field-HOA	Level attenuation, configurable attenuation slope	BRIR-based	Point source
RealSpace™ 3D Audio	HRTF-based	Fast spatialisation mode	Yes (format NR)	HRTF	NR	NR	Logarithmic, linear and custom rolloff att.	Room is divided into multiple shoeboxes, reflection coefficient can be configured per wall.	Point source. Audio clip pitch can be modified

							Min and max dist.		(speeds up or slow down)
Steam Audio	HRTF -based	Ambisonics and precomputed reverb	Own format and SOFA	No	Yes (nearest and bilinear)	NR	Level attenuation and freq. dependent air absorption	BRIR-based, using scene geometry. Includes raycast occlusion modelling of direct sound by solid objects, models partial occlusion for non-point sources.	Point and volumetric
Oculus Audio Spatilizer	HRTF -based	Virtual Ambisonics	No	NR	NR	Yes	Level attenuation and far dist.	Shoebox model, including early reflections and late reverberation	Point and volumetric
Facebook 360 Spatial Workstation	Virtual Ambisonics	No	No	No	NA	No	Linear and logarithmic attenuation curve configurable	Generates the first few orders of reflections inside a simple room model. Parametric room raytracing for important spatialisation reflections with pre-delay to enhance it. Allow use of external reverberation plugins.	Point and volumetric (width, spread). Directionality (source yaw, pitch, intensity and coverage angle)
Resonance audio	Virtual Ambisonics	Configurable Ambisonics order (HOA)	No	No	NA	Yes	Level attenuation	Ray-based method with support for arbitrary geometries, flexible assignment of surface heterogeneous materials. - Dynamic early reflections and late reverberation - Occlusion simulations (by treating high and low frequency components differently)	Point and Volumetric

MS HRTF Spat. and MS Spat.	HRTF -based	No	No	HRTF	NR	No	Attenuation curve and max distance configurable	SFX reverb (https://docs.microsoft.com/es-es/learn/modules/spatial-audio-tutorials-mrtk/10-use-reverb-to-add-distance-to-spatial-audio)	Point (omnidirectional and directional)
Virtual Acoustics	HRTF -based, HATO and Hp EQ	Virtual Ambisonics. HOA and VBAP for a loudsepakers setup.	OpenDAFF format	HRTF	NR	Yes	Level attenuation	Based on ray tracing methods. Other modes: early reflections, diffuse decay, medium absorption, temporal variation, scattering, diffraction, doppler effect, transmission of sound energy through solid structures and sound absorption by material (see more details in the web site of Table 2). Also includes BRIR-based but using RAVEN (see Ref of Table 2), which is not free.	Point (omnidirectional and directional)
IEM Plug-in Suite	Ambisonics using the MagLS approach (Schörkhuber et al., 2018)	No	No	No	No	No	No	Room encoder plug-in. It allows you to put a source and a listener into a virtual shoebox-shaped room and render over 200 wall reflections.	No
Anaglyph	HRTF -based	No	SOFA	ITD	HRIR, BRIR,	Yes	Level attenuation	BRIR-based (Virtual Ambisonics)	Point

					ITD and ILD. Linear interp.		(configurable exponent) Max and min dist.		
SOFAlizer	HRTF -based	No	SOFA	No	No	No	No	No	No
SPARTA Binauraliser	HRTF -based	Ambisonics (but in a different plug-in)	SOFA	Stored separately	HRIR and ITD, triangular-spherical interpolation	No	No	No	No

- **Direct path DSP**

The presented tools can be classified in two groups regarding the technique they use to simulate direct sound: HRTF-based and Ambisonics. In the *HRTF-based techniques* the spatialisation is performed by convolving the source signal with HRIR. In the second group sound sources are encoded into a set of *Ambisonic* channels, which are subsequently decoded, in most cases, into a set of virtual loudspeakers. Those virtual loudspeakers are spatialised as static virtual sources by convolving their respective signals with the corresponding HRIR. Notice that different Ambisonic orders can be used, allowing for a variable level of spatial resolution. Typically, renderers can be configured to use up to 3rd order Ambisonic (16 channels). Using higher order Ambisonic results in higher spatial resolution, at a higher computational cost.

In addition, other methods are included as a *low-quality alternative* but more computationally efficient. Csound tool for example offers a simpler and a low-resolution way of spatialisation, by using a spherical head model. In this case, the spatialisation is synthetically simulated by applying delays to simulate ITDs, and filters to simulate ILDs. These filters are designed according to mathematical models of sound propagation around a rigid spherical head (e.g. (Duda & Martens, 1998)). RealSpace 3D Audio includes a fast spatialisation mode that consist on just pointing directional sources, distance attenuation with logarithmic roll-off and no environment simulation. Other solutions for increasing performance at the cost of much lower spatialisation quality are: implementing distance culling, as far sources are not rendered (OpenAL Soft, RealSpace 3D Audio); projecting all sources into an Ambisonic sound field and some of them even configurable order (Resonance Audio, Oculus spatializer, OpenAL Soft, Sound Scape Renderer, Steam Audio and Virtual Acoustics); or using simple stereo panning or VBAP (SPAT, Rapture 3D, Sound Scape Renderer and Virtual Acoustics). Steam Audio also provides for reverb simulation an alternate listener-centric reverb which is precomputed over a grid of listener positions. The IEM plug-in implements an Ambisonic approximation to render binaural audio but it does not use virtual loudspeakers, but converts the Ambisonic signals directly to binaural headphone signals, with help of pre-processed HRTFs (Schörkhuber et al., 2018).

- **HRTF import**

The tools that allow to load HRTF files in SOFA format (Majdak et al., 2013) are: SPAT, OpenAL Soft, Steam Audio, Anaglyph and Sofalizer. Soundscape Renderer imports multichannel WAV files, with two channels for each direction. Some tools have converted some HRTF databases to their own custom format; Slab has translated the LISTEN and CIPIC, SoundScape Renderer has translated FABIAN and KEMAR, Rapture 3D includes a subset of LISTEN and OpenAl Soft offers a default HRTF

generated from KEMAR HRTF of MIT Media Laboratory. Using custom file formats implies that the users cannot (or hardly can) do their own translations of HRTFs and need to rely on the designers of each tool for this task. Finally, Virtual Acoustics includes ITA generic HRTF by default and supports the OpenDAFF format import.

- **HRTF and ITD customization**

Many existing tools do not allow any mechanism for HRTF customization, providing instead one fixed HRTF (CSound, Resonance Audio, Oculus and Virtual Acoustics) or a choice between a few presets (RealSpace 3D Audio, Rapture 3D). Resonance Audio includes a custom HRTF that was derived from a KU100 manikin HRTF from the SADIE database. They also offer the script used to create that custom HRTF. RealSpace 3D Audio works on HRTF personalization based on anthropometric measurements such as head and torso radius and neck height. Microsoft HRTF uses a fixed standard HRTF (averaged from anthropometric measures) but allows, using the interpupillary distance from the HoloLens, to adjust the HRTFs for the listener head size. Finally, Virtual Acoustics allow the customization of the HRTF using the listener head width, height and depth and supports HATOs.

In most tools, the delays for ITD are implicit in the HRIR data, however, some of them split the rendering of an HRTF into ITD component and minimum-phase HRIR, such as in Slab 3D and Open AL Soft. Slab 3D allows ITD customization using a spherical-head model (SHM) provided by (Woodworth et al., 1954). Anaglyph allow the ITD customization using listener head circumference dimensions.

- **HRTF and ITD interpolation**

To minimize discontinuities and artefacts when HRIR data is not available for a specific position, interpolation among different HRIRs included is necessary. Different approaches to perform the interpolation are employed. Tools such as Soundscape Renderer and Steam Audio select the nearest HRIR without interpolating. Steam Audio, together with Slab 3D, SPAT, Csound and Anaglyph, also include interpolation at runtime using neighbouring HRIRs. OpenAL Soft carries out a pre-process of the HRTF, where the HRTF grid is resampled using an offline process using a bilinear interpolation. Slab also includes this approach but also implements real-time interpolation at runtime, using a biharmonic spline interpolation method.

Some approaches extract the ITD from the HRTF and perform the ITD interpolation separately to avoid artifacts that can produce interpolating HRTFs that include ITDs (see Section 3.5.3 for more details). Examples are Slab 3D, Csound and Anaglyph. In addition, Anaglyph performs interpolation of ILD biquad filters. SPARTA Binauraliser extracts the ITD for each of the default or imported HRIRs, via the cross-correlation

between left and right HRIRs. Then, this tool interpolates the HRIRs at run-time by applying triangular-spherical interpolation on the HRTF measurement grid.

- **Near-field correction**

Regarding simulation of near-field sources, Oculus Spatializer and Resonance Audio model the effect of acoustic diffraction around the head. SoundScape Renderer provides a resource-expensive experimental solution using High Order Ambisonic (HOA). Open AL Soft allows to enable what they call "near-field control filters" that compensate for low-frequency effects caused by the curvature of nearby sound waves. Anaglyph implements a ILD modifier that applies a frequency dependant adjustment and parallax HRIR correction. SPAT also uses filters for correcting the ILD, cross-ear selection of HRTF filters, and geometrical corrections of monaural gains and delays.

- **Distance simulation**

Most existing available tools simulate distance through level attenuation due to sound propagation through air, which follows the inverse square law (attenuation of 6dB with every doubling of the distance). Although this is often the default setting due to its physical correctness, some tools provide customization of the distance attenuation curve (RealSpace 3D Audio, Open AL Soft, Facebook 360, MS spatializer, Rapture 3D and Anaglyph).

The effect of air absorption at high frequencies for large distances (e.g. more than 15 meters) has been addressed by some tools, such as SPAT, Slab 3D, Oculus spatializer and Steam Audio. Others just provide distance culling to save processing resources for far sources (Real Space 3D Audio, Slab 3D and Oculus spatializer and MS spatializer).

- **Reverberation**

Most available tools can simulate reverberation, employing three main approaches: BRIR-based, *physics-based* algorithms and *delay networks* methods.

The Soundscape Renderer, Steam Audio, Virtual Acoustics and Anaglyph are the ones that implement a BRIR-based method, where the impulse responses of the environment to be simulated are convolved with the audio signal. Usually, these impulse responses are binaurally registered using a dummy head microphone, with the sources positioned in different locations. This allows for a certain level of spatialisation of the reverberation sound. The main problem of this approach is the computational cost, as these impulse responses can be very long. An inconvenience of these tools is that they do not allow to import the BRIR. Instead, they provide a set of BRIRs to simulate the reverb of the room.

In the tools that implement a Synthetic reverberation, the response of the room can be simulated synthetically using different approaches. Resonance audio use ray tracing methods which can handle rooms with arbitrary geometry. In addition, this tool also implements Spectral Magnitude Decay (SMD) method in the frequency domain. This approach works for late reverberation as well, and are able to simulate simplified geometries, as a shoebox rooms. Facebook 360 also implements ray-methods and allows the use of external reverberation plugins. Virtual Acoustics, apart from BRIR-based methods also implements simplified models to create a physics-based auditory impression, such as ray tracing and image source algorithms and a combination of both approaches (Vorländer et al., 2010).

Csound's reverberation is mainly based on Feedback Delay Networks (FDN) in the time domain, using the shoe-box approximation with configurable number of reflections. OpenAL Soft also uses FDN but with configurable impulse response parameters.

Most of the other tools implement synthetic reverberation using parametric shoebox models, allowing the user to configure the dimensions of a rectangular room and walls materials and reflection coefficients, usually processing separately early reflections and late reverberation tail (SPAT, Oculus Spatializer, Rapture 3D, Virtual Acoustics and Slab 3D). The solution adopted by RealSpace 3D Audio is based also on the shoebox model but allowing to build more complex room geometries by dividing the geometry into multiple shoeboxes. Some tools go even further, allowing configuration of arbitrary room geometry through physical models of sound propagation based on scene geometry (Steam Audio, Resonance Audio).

Some tools provide means for simulating occlusions and reflections on obstacles (Resonance Audio, Steam Audio, Virtual Acoustics and Rapture 3D), while others delegate this to the application level.

As for the direct sound simulation, it is also common to save resources in the reverberation process, by reducing the number of early reflections (implemented by Real Space 3D Audio), or precomputing room impulse responses for different points in the scene (Steam Audio).

2.5 Auditory models

An auditory model is a mathematical algorithm that has been implemented to predict a listener performance in a specific auditory task, trying to mimic one or several parts of the human auditory system. To understand how a model can do that, this section starts describing the *auditory signal processing* that is carried out by the auditory system

(the anatomy of the auditory system was described in the Section 1.1.1), i.e. the process by which the listener perceives the sound and converts it into data that can be handled by the brain. In addition, this section presents a general description of a *binaural auditory model* and some examples.

2.5.1 Auditory signal processing

Figure 21 shows the auditory pathway of the sound signal at the peripheral auditory system. Note that these processes are carried out in parallel at both ears. Once the sound reaches the ear, the pinna modifies the sound waves (reflecting, attenuating or amplifying) before it enters the middle part of the ear through the *auditory canal*. These modifications to the signal, together with previous modifications performed by the listener's head and body, result in a set of auditory cues that will later help in the sound localization and intelligibility. The auditory canal gathers the different sound waves and leads them to the *eardrum* at the end of the canal, which vibrates and makes the three chained *ossicles* oscillate synchronously. In this part of the ear, the sound is captured through the different pressures at the windows of the cochlea (*oval* and *round windows*), where the transmission of the sound from air into vibrations of the fluid in the cochlea is carried out. In this way, the middle ear technically acts as an impedance-matching device which transform from the relatively low impedance airborne sounds to the higher impedance fluid in the inner ear (B. C. Moore, 2012).

The cochlea is formed by two membranes, *the basilar membrane* and the *vestibular membrane*. The organ of Corti is attached to the basilar membrane and contains an array of sensory *hair cells* that contact with the tectorial membrane. The cochlea is in charge of separating sounds according to their spectrum, performing a frequency selection and creating what is called the tuning curves. In this way, each point of the basilar membrane corresponds to a specific value of the stimulating frequency. The distribution of frequencies to places is called the tonotopic organization of cochlea. An uncoiled representation of the cochlea can be seen in Figure 21. The basilar membrane is widest (0.42–0.65 mm) at the *apex* of the cochlea, where high frequencies are captured and narrowest (0.08–0.16 mm) at the base (near the round and oval windows), which detects the low frequencies. However, one incoming single frequency is not limited to one point on the basilar membrane, so this part of the signal processing is often modeled by a set of adjacent bandpass filters (Ashmore, 2008). Then the hair cells act as a mechanoelectrical transduction, transforming the mechanical stimulus into electrical nerve signals, which will travel through the auditory nerve to the brain. All this process, performed by the peripheral auditory system, is often compared to a Fourier analysis of the sound wave (Roginska & Geluso, 2017) and the whole basilar membrane can be

described as a bank of overlapping filters (Meddis & Lopez-Poveda, 2010). The activation of the auditory nerve is analyzed in the form of a spectro-temporal response pattern (Jens Blauert, 2013).

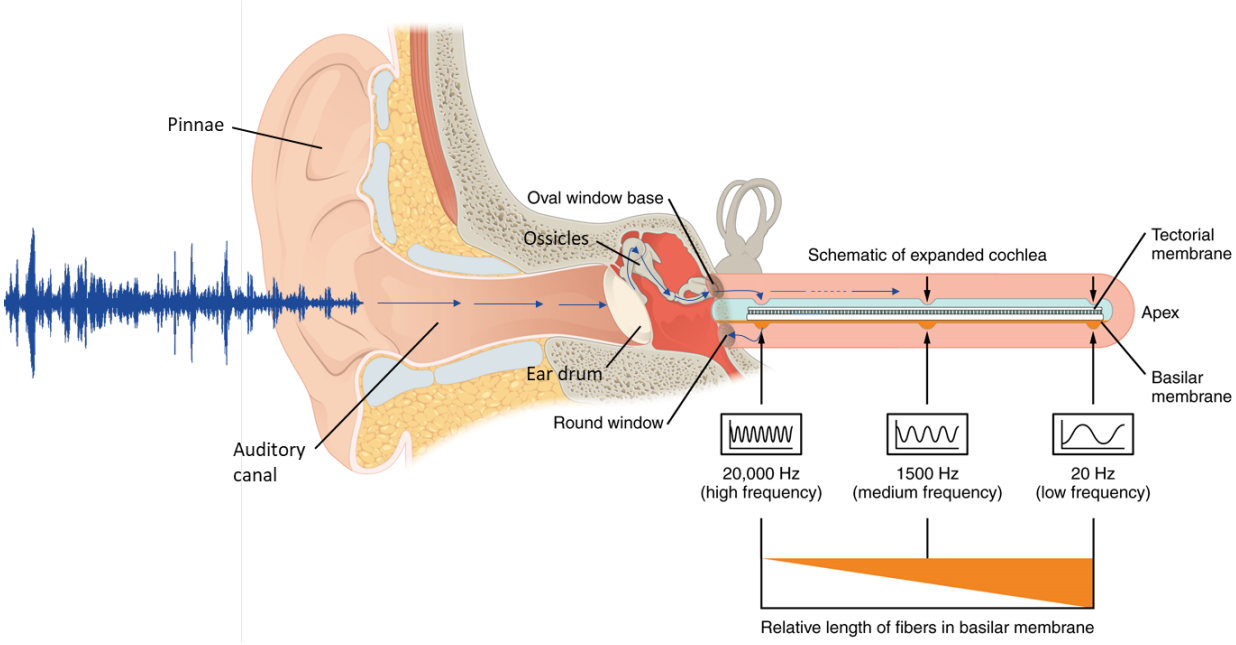


Figure 21. Peripheral auditory processing. Original image from Wikipedia (CC BY 2.5), with modifications.

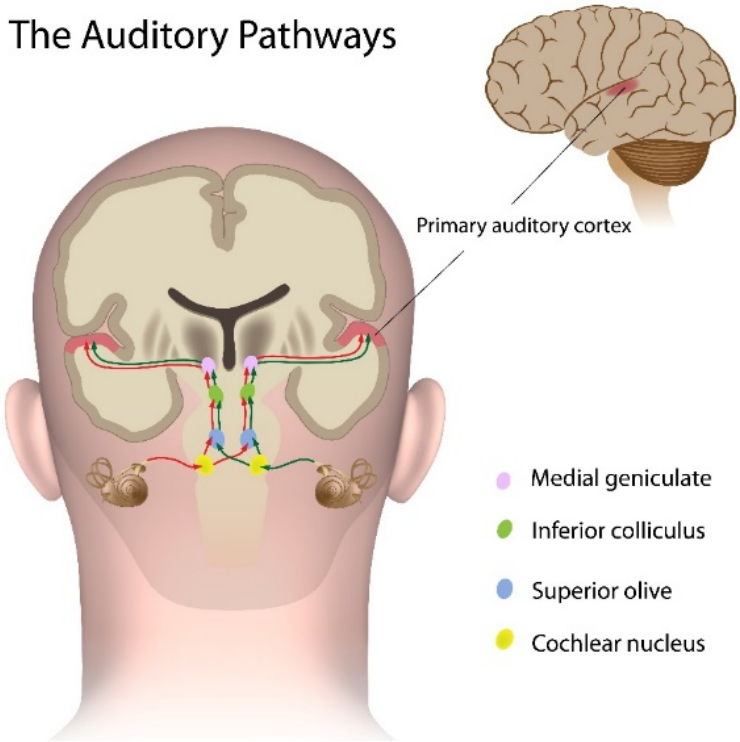


Figure 22. The auditory pathway from the cochlea to the primary auditory cortex. Image from *www.shutterstock.com* (ID: 229583680)

The ascending auditory pathway from the cochlea to the brain is carried out at the primary auditory cortex (Figure 22). The *superior olivary complex* is the first place where the information from the two ears interacts, processing the interaural level, time and phase differences and using inputs from the ipsilateral and the contralateral *cochlear nucleus.* Then, at a higher level of the pathway, the *superior olive* and the *inferior colliculus* receives input not only from the auditory pathway, but also from the nearby superior colliculus which processes visual inputs, leading to a multi-sensory integration by this part of the brain. The inferior colliculus relays auditory information to the *medial geniculat*e nucleus, which send the information finally to the auditory cortex.

2.5.2 Binaural modeling

A binaural auditory model is a mathematical algorithm that tries to replicate the previous mentioned parts of the human auditory systems and their specific relation with the binaural hearing, in order to predict the listener response to one or some sound stimulus. Models can be classified regarding the part of the auditory system they try to replicate or which kind of human performance they try to predict. According to the last classification, several models can be classified in two groups: *localization models* and *detection models* (Jens Blauert, 2013). The localization models predict the position of a sound source, while the detection models try to predict if a given sound source is perceived or not. Both group of models will be briefly described in the next sections.

Auditory models can be used in multiple scenarios. One of the most common ones is to use them as an alternative to a real subject in a perceptual evaluation of an audio system. Performing a perceptual evaluation requires the design of a controlled listening experiment and many subjects to carry it out, which is a laborious task. According to the AabbA group (Jens Blauert et al., 2010), who have an extensive experience in binaural modeling, other potential application areas are: audio technology (such as evaluation of the quality of audio hardware), audiology and hearing aids (such as assessment of speech intelligibility, hearing disorders or configuring hearing aids and aural virtual environment (e.g. experiment for identification of virtual sources or auditory-scene mapping). Models are also very useful for learning about the functioning and the structure of the human auditory system. Modeling the system and performing different modifications and configurations can help to evaluate and understand each part of the system and how a change in one or more component affects the performance of the whole system.

2.5.2.1 Localization models

Localization models utilize binaural cues (interaural and monoaural) to predict the position of a sound source. Regarding interaural cues, the first model can be found in the work of Jeffress (1948), which is an ITD-based model. Jeffress model tries to describe the interaction of neural activity from both ears by the auditory nerve fibers, implementing a combination of delay lines (one for each ear) and a set of coincidence cells. Signals from the two ears travel along both delay lines until they meet each other in a coincidence cell. Depending on which cell both signals meet, the model predicts a specific lateralization angle for the source.

Another method to estimate ITDs, which is currently widely used, is called the interaural *cross-correlation* method (IACC), introduced by (Colin Cherry & Sayers, 1956). In this kind of methods, the peak of the cross-correlation between the left and right channels determines the ITD of the sound source and then its lateral position. In addition, J. Blauert & Cobben (1978) worked on comparing left and right channels in frequency bands (called auditory bands) to simulate the mechanism of the basilar membrane of the inner ear. As well as applying a half-wave rectifier and a low-pass filter to simulate the behavior of the hair cells.

Reed, M. C., & Blum (1990) and Zacksenhouse et al. (1992) presented models to simulate the ILD processing by the lateral superior olive. These models follow the idea of (Jeffress, 1948), presenting an algorithm based on a set of excitation/inhibitions cells that correspond to different ILD and react to different values of power differences between left and right ear.

Regarding monaural cues, models based on templates are worth mentioning. These models consist on a comparison of the representations of the sound signal with an internal template of a listener, stored by the algorithm. An example can be seen in the model of Langendijk & Bronkhorst (2002), which predicts the localization of a sound source in the sagittal plane by comparing the spectrum of the signal presented to the listener with a measured directional transfer function (template). R. Baumgartner et al. (2013) also presented a model based in templates, which analyses the inter-spectral differences (ISDs) for various angles and predicts the localization of the sound source by finding the best ISD match between the analysed sound and the templates.

Many localization models are available today, from models that localize in the horizontal planes (Braasch & Hartung, 2002; Dietz et al., 2011) to the ones that localize in vertical planes (Robert Baumgartner et al., 2014; Majdak et al., 2014) or on the whole sphere (J. Reijniers et al., 2014).

2.5.2.2 Detection models

The second group of models, the *detection models,* try to mimic the human ability to distinguish signals in spatial auditory scenes, which is possible thanks to detectable dissimilarities between signals. These dissimilarities must be defined by the model to mimic a specific human behavior. Many of these models are based on a threshold detection of signal levels or signal-to-noise ratios.

It is very well known that binaural cues help in the detection of signals. In binaural conditions, many detection models follow the Equalization Cancelation (EC) theory (Durlach, 1963). This approach tries to quantify the binaural advantage in signal detectability, predicting what is known as Binaural Masking Level Difference (BMLD). EC theory suggest that first attempt to equalize the signal at both ears separately, based on 'a priori' information obtained from data collected in several experiments, and then subtracts both signals (cancelation). John Culling & Summerfield (1995) improved the EC model adding more components, such as a gammatone filter bank to perform the frequency bands separation to apply a different EC to each band and a method of compensation/rectification of the signal to simulate the hair cells behavior. This work was followed by (Breebaart et al., 1999) which added parallel processing in filter sub-bands, and cross-correlation models, among other components (Jens Blauert, 2013).

Among the detection models are those for speech intelligibility and the cocktail party phenomenon. Adelbert W. Bronkhorst (2015) shows an overview of binaural speech perception models. These models try to predict the threshold where a speech (target) can be separated from one or more interferes (maskers). Jelfs et al. (2011) presented a model to predict the SRT (Speech Reception Threshold) for speech in a noisy environment considering reverberant conditions, computing the increase in speech intelligibility when the target and noise interferer are spatially separated. The approach of this model consists of decomposing different effects to predict them individually. In this way, this model predict firstly the BMLD (from a formula of Culling et al. (2004, 2005)), together with a gammatone filter bank and an X-correlation, and secondly the benefit of the best-ear listener, simulated by computing the speech-to-noise ratio at the two eras by frequency bands. Finally, to obtain the SRT a simple adding method of both paths is used. This model will be described and used in the experiment described in Chapter 5.

2.5.3 Auditory toolboxes

There is a set of toolboxes that offer implementations of different auditory models. The *auditory toolbox* was published in 1994 as one of the first collection of auditory models that implements several popular models. The last version of this MATLAB Toolbox was presented in (Slaney, 1998) and the development seems to be stopped at that point[20]. The *auditory-image-model toolbox* (AIM) (Patterson et al., 1995) is a software package with a modular architecture that contains three main modules for spectral analysis, neural encoding and time-invariant stabilization. It is distributed in two formats, as a MATLAB package and a C code package[21], with a last activity reported in 2016. Another toolbox of auditory models was created by a research group of the University of Eseex[22]. They provide a *development system for auditory modeling* (DSAM), with a set of modules written in C and an application called AMS (*Auditory Modelling System*), which can be called from MATLAB and allows users to use the modules without handle the C code.

Finally, the *auditory modeling toolbox* (AMToolbox) is an open-source collection of auditory models available as Matlab/Octave toolbox, maintained by Piotr Majdak[23] and offers the most recent models. An extensive description of the AMToolbox can be found in (Søndergaard & Majdak, 2013). This toolbox contains models for many stages of the auditory system, from HRTFs modeling the outer- and middle-ear acoustics, various cochlear filters, inner-hair cell models, spatial models (lateralization, median-plane localization, sound externalization, spherical sound localization, etc.), up to speech perception models (intelligibility and spatial unmasking). This toolbox also offers a large amount of data from psychoacoustic experiments and acoustic measurements.

[20] https://engineering.purdue.edu/~malcolm/interval/1998-010/, retrieved January, 2022
[21] https://code.soundsoftware.ac.uk/projects/aim, retrieved January, 2022
[22] https://www1.essex.ac.uk/psychology/models/, retrieved January, 2022
[23] http://amtoolbox.sourceforge.net/, retrieved January, 2022

Chapter 3

The 3DTI Toolkit-Binaural Spatialiser

This chapter describes one of the main contributions of this PhD thesis, the 3DTI Toolkit-BS open-source tool, a 3D binaural audio spatialiser for virtual auditory spaces. The chapter starts with an introduction of the Toolkit in Section 3.1. Section 3.2 presents the architecture of the tool and briefly describes each of its components. Next sections introduce each component in detail, describing the techniques and algorithms implemented within each component. After the description of all the implementations, Section 3.7 presents some information about the current distributions of the Toolkit and describe some additional tools that are also available. Finally, Section 3.8 includes a discussion where the features of the implemented tool are compared with the ones offered by other existing tools.

3.1 Overview

The 3D Tune-In Binaural Spatialiser (3DTI-BS) described in this chapter is called 3DTI Toolkit-BS (3D Tune-In Binaural Spatialiser), named after the project in which it was developed, the 3D Tune-In project[24], as presented in Chapter 1. The 3DTI Toolkit-BS is an open-source and multiplatform C++ audio renderer that enables the design and creation of VASs. The renderer provides a set of methods and algorithms to achieve a high level of realism and immersion within headphones-based 3D audio simulations.

The 3DTI Toolkit-BS enables the integration of binaural audio spatialisation into a virtual environment, simulating in real time any moving sound source coming from a specific place. The direct sound simulation is carried out through convolutions with

[24] http://www.3d-tune-in.eu/. Retrieved January, 2022.

HRIRs. The HRTF can be selected from a previously loaded SOFA file and then interpolated to get the HRIR in the specific direction of a sound source. In addition, the ITD can be re-computed taking into account the listener's head circumference. Furthermore, distance simulation can also be performed for near and far sources. The renderer simulates near-field sources adding an extra shadow in the contralateral ear for sources very close to the listener's head. In addition to the anechoic spatialisation, the 3DTI Toolkit-BS integrates binaural reverberation capabilities (using different methods than for the anechoic spatialisation) by convolving sources with BRIRs using Virtual Ambisonics. This approach, together with an efficient convolution algorithm in the frequency domain, allows the 3DTI Toolkit-BS to compute large reverberating scenes, with virtually unlimited number of moving sources, maintaining high spatial accuracy for the direct sound (spatialized using direct-HRIR convolution). The listener head movements are also taken into account in the sound spatialisation.

The following sections will present a detailed technical description of the 3DTI Toolkit-BS components and the digital processes implemented to create the VAS, focussing on the technical innovations integrated in the processing chain. Early descriptions of the Toolkit were presented in 2017 (María Cuevas-Rodriguez et al., 2017) and 2018 (Maria Cuevas-Rodriguez et al., 2018). These publications included only high-level descriptions of the first prototype releases of the 3DTI Toolkit-BS. An in-depth literature review, overview and analysis of the various algorithms and components of the 3DTI Toolkit can be found in (Cuevas-Rodríguez et al., 2019).

3.2 3DTI Toolkit-BS components and structure

The approach followed by the 3DTI Toolkit-BS decouples the simulation of the anechoic path (direct sound) from the reverberation path (environment), to carry out the binaural spatialisation. Figure 23 shows the high-level scheme of the process, performed for the simulation of multiple sources at different positions within the VE.

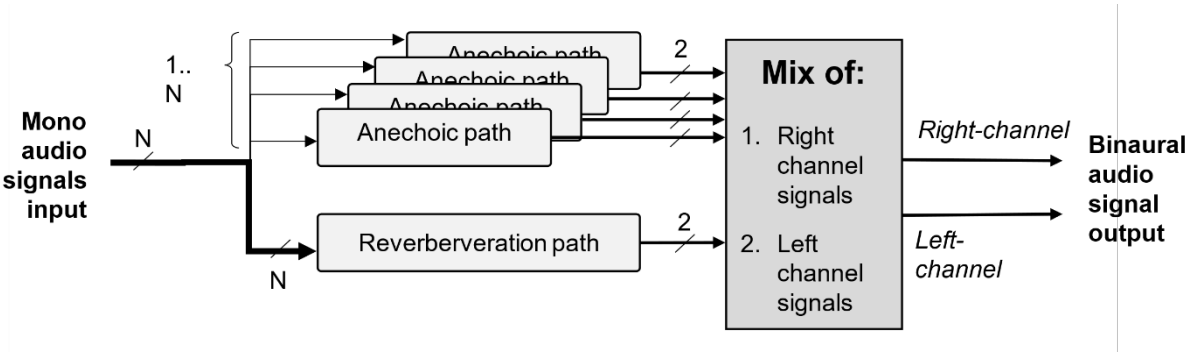


Figure 23. High level 3DTI Toolkit-BS process structure.

The input of the system is a set of mono audio source signals. The simulation of the *anechoic path* is independently computed for every source, which allows to maintain high spatial accuracy, using direct HRIR convolution for each source at their specific direction. The implementation of the *reverberation path* is based on BRIR convolution. However, a Virtual Ambisonic approximation is used to reduce the number of convolutions, as the BRIR may be very long. In this way, all sources are computed at the same time in the reverberation path, but thanks to the Ambisonic codification they keep certain location-dependent characteristics. This structure allows a very high spatial resolution and accuracy for the anechoic path while an efficient simulation, although rendered with less fidelity, for the reverberation path. The process of each path generates two signals, one for the left channel and another one for the right. Both, left-channel and right-channel signals are mixed separately, creating the *binaural audio signal output,* and sent to the headphones to be delivered to the listener.

Components of each path can be seen in Figure 24, where the process chain for a single source (N=1) is shown.

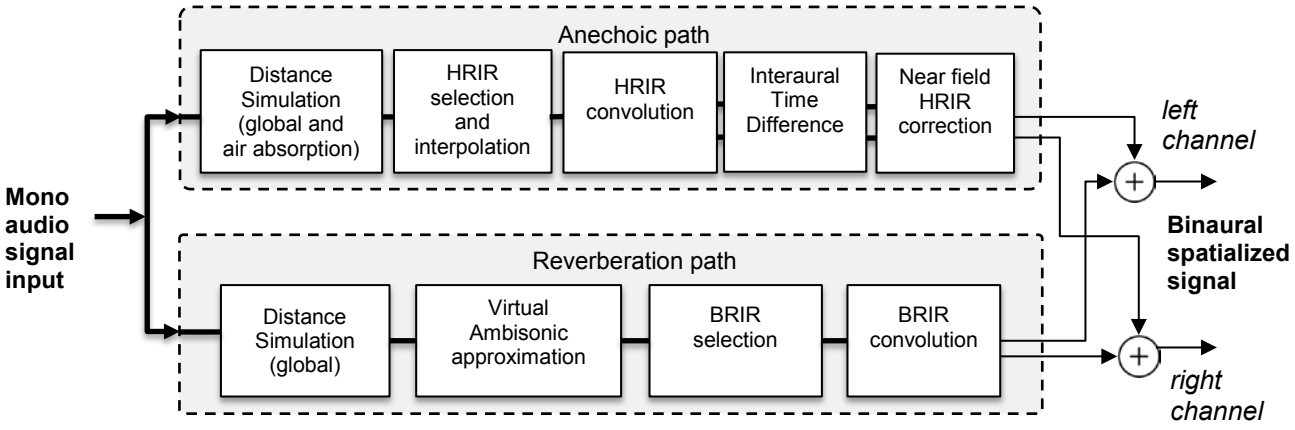


Figure 24. Low level 3DTI Toolkit-BS process structure for a single sound signal.

The ***anechoic path*** simulation is carried out for each source independently, through the following components, which will be described in detail in the following sections of this chapter:

- *Distance simulation.* A global distance-dependent attenuation is applied to the source, based on the acoustic power law. It also includes a frequency-dependent air absorption simulation for sources further than 15 meters.
- *HRIR selection and interpolation.* The HRIR for the specific source direction is obtained by a barycentric interpolation among a set of HRIRs selected from a full set of a previously loaded file.
- *HRIR convolution.* This module convolves the previously calculated HRIR with the sound signal, using a uniformly partitioned convolution.

- *ITD simulation.* ITDs, which has been previously removed from each HRIR, is added after the convolution. This allows reducing the comb filtering effect and using a customized ITD according to the listener's head circumference.
- *Near field HRIR correction.* Sound sources in the near field are simulated performing some additional filtering to the previously convolved signal. These filters simulate the specific ILDs that are caused by the head shadow at short distance.

The ***reverberation path*** simulation is also carried out in a set of steps different from the anechoic path. As previously mentioned, this path computes all the sources at the same time. This path is composed of the following components, which will be described in detail in the following sections of this chapter:

- *Distance simulation.* In the same way as the anechoic path, a global distance-dependent attenuation is applied to the source. This is independent from the anechoic attenuation, which allows the configuration of the direct-to-reflected signal ratio in the final binaural stereo output.
- *Virtual Ambisonic approximation.* The scheme of Figure 24 shows the example of a single source, however, as can be seen in Figure 23, all sources are computed together in "one reverberation path". To do so, this component encodes all sources into 1st Order Ambisonic B-Format signals (W, X, Y and Z channels), which allows to partially keep the spatial information of each source.
- *BRIR selection.* The BRIR are selected according to the configuration of the virtual Ambisonic approximation and then encoded into Ambisonic B-Format signals.
- *BRIR convolution.* Finally, the Ambisonic-encoded sound signals are convolved with an Ambisonic-encoded version of the BRIRs.

Other components included in the 3DTI Toolkit-BS that are not part of the process chain but support all previously mentioned processes are:

- *Geometric calculation.* Both anechoic and reverberation paths consider 3D locations for the sources, and the location and orientation of the listener. A set of classes have been implemented to support geometric calculations regarding the position of sources and listener.
- *Coordinate system conventions.* The 3DTI Toolkit-BS handles both polar and inter-aural spherical coordinates to manage the position of each source relative to the listener. It also provides classes for handling conventional transformations, including location and orientation. See Chapter 1, Figure 10 to see the coordinate system used by default for the 3DTI Toolkit-BS.

- The 3DTI Toolkit-BS supports *configurable frame sizes and sample rates.* This, in addition to increasing the configurability of the system, allows very low latency when a small frame size is selected, assuming a higher computational cost. Chapter 4 will show more details regarding the 3DTI Toolkit-BS performance.

This flexible and modular structure allows the use of each component independently, as well as the integration of new rendering methods or the substitution of some modules by others. In addition, a special effort has been put in removing artefacts related to dynamic scenes, where sources and listener are free to move. This is a particularly important requirement for interactive VR applications where the sound designer cannot easily predict scene changes in advance. The mechanism used to reduce these artefacts are described in each component.

3.3 Distance simulation

The ability to perceive the sound source distance from a listener is considerably worse than the ability to perceive directions. According to (Zahorik et al., 2005), humans perform a set of rather complex processes based on multiple cues for estimating the distance of a sound source. These involve a large set of parameters which result in the modification of the sound input into the auditory system.

The 3DTI Toolkit-BS implements a simplification of these processes, considering a global attenuation of the signal, with different attenuations for the anechoic and the reverberation paths. In addition, the Toolkit simulates air absorption and near-field effects for the anechoic path that will be described in sections 3.5.1 and 3.5.4 respectively.

To compute the global attenuation, a general rule is applied: doubling the distance between the source and the listener causes a 6 dB reduction of the sound level. For a given distance d, the global attenuation $A(d)$ is computed as:

$$A(d) = A_{ref}^{\log_2(\frac{d}{d_{ref}})}, \tag{3.1}$$

where A_{ref} is a configurable parameter, and d_{ref} is the distance at which the HRTF was measured, where we assume a 0 dB attenuation. The diagram of Figure 25 represents the global attenuation applied to a sound signal in decibels, following the previously described Equation (3.1).

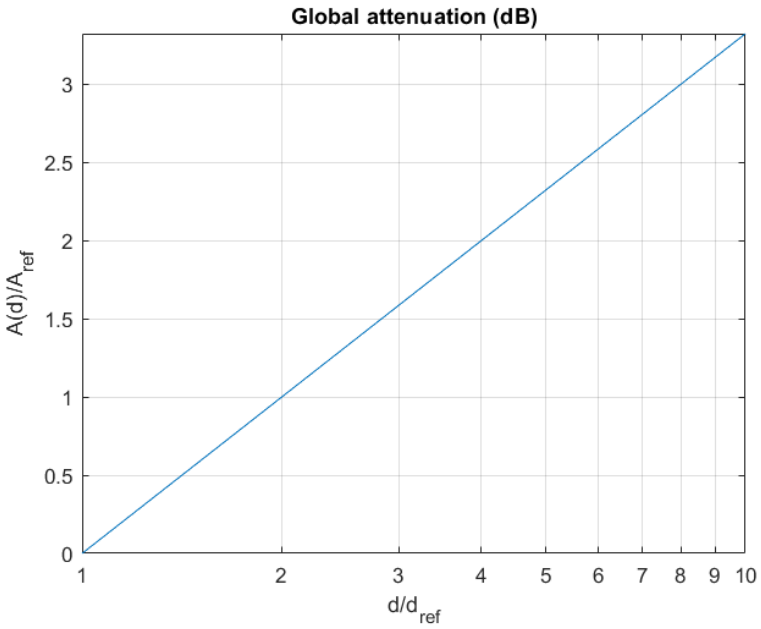


Figure 25. Global attenuation applied to the sound in decibels. Diagram that represents the inverse square law followed by the Equation (3.1).

3.3.1 Global attenuation smoothing mechanism

When the distance between source and listener varies, a smoothing mechanism avoids sudden changes in the signal level which could produce artefacts in the sound. An adaptive attenuation value a_i is applied to each sample of the audio buffer, asymptotically approaching the desired new attenuation $A(d)$, using the following law:

$$a_i = (1 - \rho) \cdot a_{i-1} + \rho \cdot A(d) \tag{3.2}$$

$$\rho = 1 - \exp\frac{\log 0.01}{t_a \cdot f_s} \tag{3.3}$$

where t_a(attack time) is the time for the envelope to reach 99% of the change, and f_s is the sampling frequency. By default, the 3DTI Toolkit-BS uses $t_a = 200ms$.

This attenuation is computed separately (but using the same method) for the anechoic and the reverberation path. If different A_{ref} values are chosen for the two components, the direct-to-reflected sound changes according to the distance of the sound source, emulating what happens in real-life conditions. The default A_{ref} values are -6dB for the anechoic path, and -3dB for the reverberation, in line with the work of Chowning (1971) when he was working on the simulation of moving sound sources in reverberant spaces.

3.4 Convolution with HRIR and BRIR

The 3DTI Toolkit-BS performs convolutions in both anechoic and reverberation paths. Anechoic spatialisation is performed by convolving the input signal with the HRIR, while the reverb simulation follows a Virtual Ambisonic approach, which is an optimized Ambisonic-based solution in which a set of BRIRs (measured at specific positions) are convolved with the Ambisonic codification of the input signals. HRIRs and especially BRIRs are long impulse responses signals. Rendering real-time spatial audio by convolution methods is challenging and requires careful implementations. To carry out a real-time simulation, the input signal is divided into frames; the whole signal processing is carried out for each frame, which means that the implementation must be fast enough to perform all the process before the next frame arrives.

As described in Chapter 2, convolutions with HRIRs and BRIRs usually are carried out in the frequency domain, since employing FFTs and complex multiplications reduces the cost of convolutions, comparing with the cost in the time domain. However, FFT operations are very problematic when having very large impulse responses. The optimum performance happens when the size of the impulse response is the same as the frame size. When performing convolutions with BRIRs the computational load can be really high due to the large size of the BRIR filters compared with the frame size. In the case of HRIRs, the impulse response sizes are more similar to the frame size, however, it must take into account that all the rendering processes are doubled for two ears and they scale linearly with the increase of the number of sources (as it is shown in Figure 23).

To solve this problem, different methods have been proposed by different authors (Välimäki et al., 2012). One of the solutions is to use the Graphic Processing Unit (GPU) to perform the convolution processes, thanks to its highly parallel programable processors (Lauri Savioja et al., 2011). An example can be seen in (Belloch et al., 2013), where they use the GPU to execute multiple convolutions simultaneously, within an application that simulates the movement of a source between different positions. The application was able to render up to 240 moving sources at the same time with an HRIR of 512 samples.

Another solution can be found in techniques that compute *convolutions in blocks*. These techniques are based on the idea of splitting the IR into several segments. Then, compute the convolution with the input signal in blocks. And finally sum the convolutions of each block to get the output signal. This technique presents a main advantage (Wefers, 2015): (1) when the frame size and the IR length differ strongly (which is usually the case of the BRIRs), it reduces the zero-padding necessary to match the size of both signals and (2) convolutions can be performed in blocks of short size, requiring short FFTs. In this technique, IRs can be divided into blocks of the same

lengths (Wefers, 2015; Wefers & Vorländer, 2011) or different lengths (Gardner, 1994; Wefers, 2015). The *one-block-size* methods, which use blocks of the same length, allow to do the partitions and FFTs off-line before the real-time processing) while the *different-block-size* methods involves that partitions and consequently FFTs of different length to be executed during real-time processing. Garcia (2002) proposed a method that considers using different block lengths which are previously configured. This allows the FFT processes to be performed off-line and to have different pre-configured options during the real-time processing.

Torger & Farina (2001) presented the computational cost of the unpartitioned and partitioned convolution (with equal and different sizes). In a standard personal computer, according to their measurements, the partitioned convolution algorithm outperforms unpartitioned convolution and allows for a small overall latency, if an appropriate number of partitions is selected. In addition, their measurements showed that the different-block-size methods were never faster than the one-block-size convolution. Following this idea and after studying the work of (Wefers, 2015), the 3DTI Toolkit-BS implements a one-block-size FFT and circular convolution, but with a modification that allows to simulate moving sources. Wefers calls this method the *uniformly partitioned convolution* and will be described in detail in the following section.

3.4.1 The Uniformly Partition Overlap-Save (UPOLS) convolution

In both anechoic and reverb paths, the 3DTI Toolkit-BS employs the Uniformly Partition Overlap-Save (UPOLS) convolution in the frequency domain presented in (Wefers, 2015; Wefers & Vorländer, 2011). This FFT-based algorithm splits the IR into a set of blocks with the same length as the frame size (N). Each of these blocks is treated as a separate IR and convolved by a standard overlap-save process. As mentioned before, it allows convolving with long IRs in a very efficient way, since this technique decomposes them into better manageable shorter blocks. It is especially convenient in the case of long IR, as the BRIRs.

Figure 26 shows the whole process of the UPOLS in the case of the HRIR, but it is the same for the BRIR. The partitions and FFTs of the HRTF are computed offline, where the whole HRTF is partitioned in blocks of length N (same length as the frame size) and stored. At run-time, for every audio frame, a new input buffer arrives, the content of the input block buffer is shifted N samples to the left and the new input of N samples is then placed on the right (see bottom-left part of the diagram). Then, the whole input buffer is transformed into the frequency domain using a 2N-point real-to-

complex FFT and stored in a *delay-line,* following an overlap-save scheme. In this way, in each audio frame, the input signal stored in the delay line is shifted up by one frame slot. During the audio frame, each delayed input buffer is convolved with each HRIR segment, by a multiplication in the complex domain. Finally, all the multiplications results are mixed and transformed back into the time-domain. To do so, the Toolkit implement a 2N-point complex-to-real IFFT, where the first N points are discarded, as we are using an overlap-save scheme. These sample removal and the size used for the FFT and IFFT of 2N-point are implemented to avoid aliasing in the time domain (Oppenheim, 1999). In addition, zero-padding is used to complete the signal buffer in case it is needed (Oppenheim, 1999).

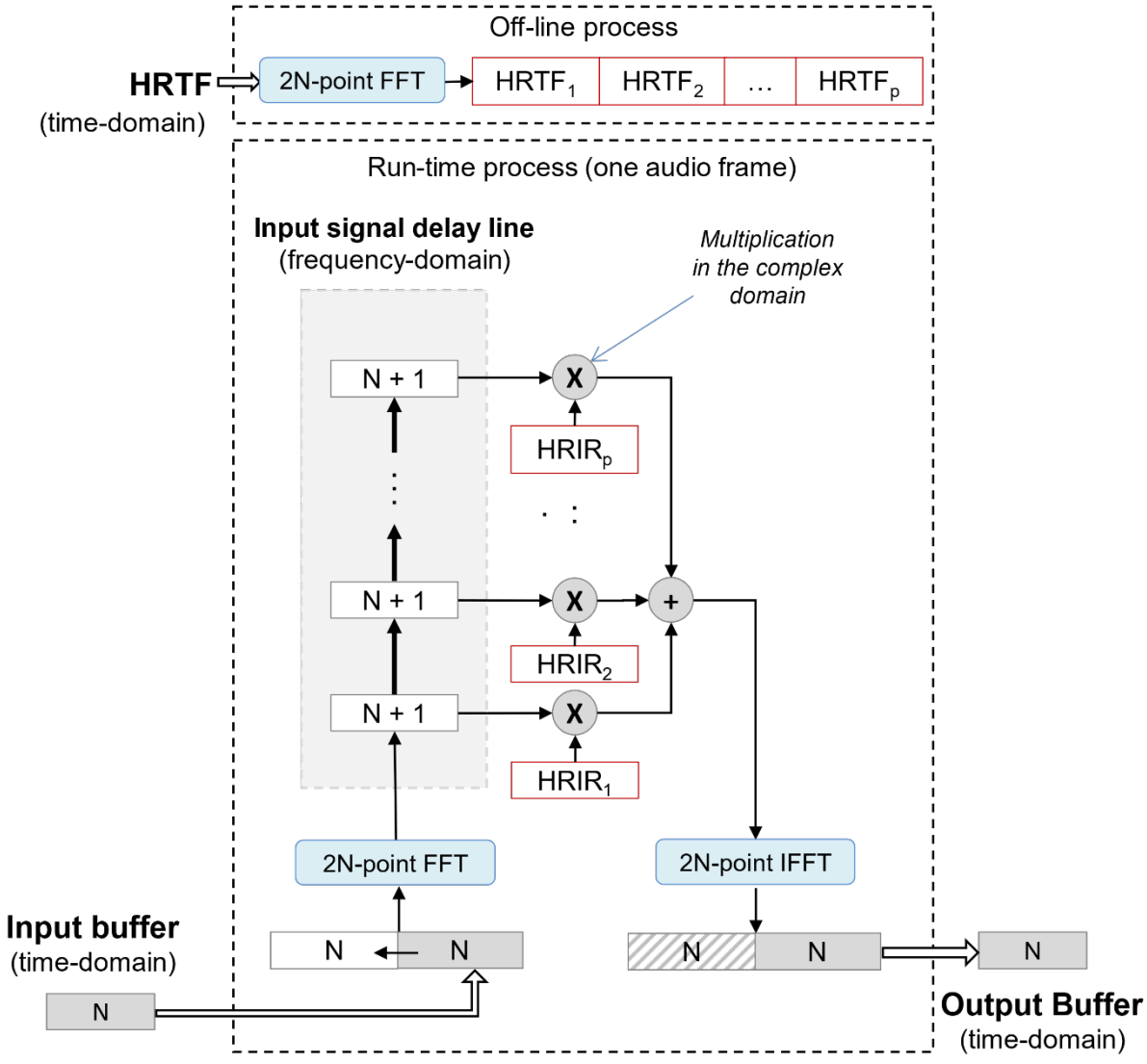


Figure 26. UPOLS convolution algorithm of Wefers (2015a) for static sources and/or listener in a real-time processing. This scheme shows an example of a convolution between the sound source input signal and the HRIR, for the specific position of source and listener.

3.4.1.1 Convolution smoothing mechanism

In scenarios where sources and listener are moving, the HRIRs are constantly changing. The UPOLS convolution can be modified to support these changes, if we add a new delay line to manage the HRIR partitions, as shown in the left red box of Figure 27. This improvement has been included only in the convolution with HRIR. The BRIR convolution does not present this problem since, in this case, the BRIR is convolved with a set of Ambisonic channels that are always the same regardless of the position of the source and the listener.

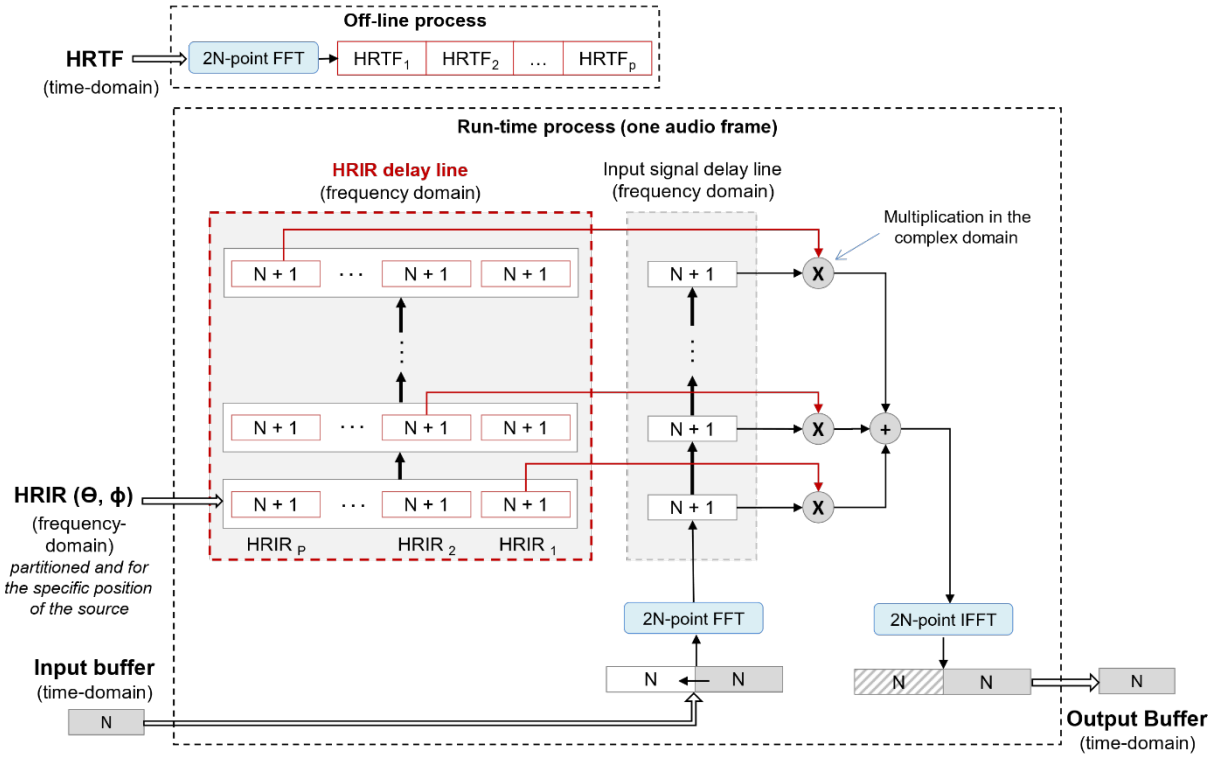


Figure 27. **Modification of the original UPOLS convolution algorithm** by Wefers (2015a) for moving sources and/or listener in a real-time processing.

In this new delay line, for every audio frame, and in the same way as the input signal, a new HRIR, corresponding to the direction of the source at this time, is introduced in the first slot of the HRIR delay line. Then, in each audio frame, both delay lines (HRIR and input buffer signal) are shifted up by one frame slot. In this way, the input signal is convolved with the new HRIR while the remaining slots of the input signal are convolved with the previous HRIRs. As a result, the number and artefacts due the movement of the sources can be significantly reduced, which will be shown in the chapter of the evaluation (Chapter 4).

3.5 Anechoic path

The process implemented by the 3DTI Toolkit-BS to simulate the spatialisation of the direct sound from the sound source to the listener, labeled in Figure 23 as the anechoic path, is composed of a set of components shown in Figure 24. Since the distance simulation and the convolution components are used in both paths, those algorithms have been described in the previous sections. This section describes the specific components and algorithms only implemented in the anechoic path, such as near and far field distance simulation, HRTF interpolation and ITD simulation. Finally, an alternative spatialisation process (high performance) is described.

3.5.1 Air absorption simulation

Another auditory cue well known for distance perception is the high-frequency attenuation caused by the air absorption. In the free field and for distances larger than 15m, the air absorption acts as a low pass filter, modifying the spectral characteristics of the sound. The 3DTI Toolkit-BS implements a set of filters to simulate sources placed at very far distances. Those filters have been implemented as two cascaded second order Butterworth low pass filters, designed to match the data presented in the ISO9613-1 standard (ISO 9613-1, 1993). Figure 28 shows the cut-off frequency (vertical axis) used for each distance (horizontal axis). A cut-off frequency of 20kHz has been selected as a reference for a distance of 15 meters, and it is exponentially decreased as distance increases, again to match data presented in ISO9613-1. The air absorption is shown in dB with a color map, being the white color the values that correspond to an absorption of -6dB.

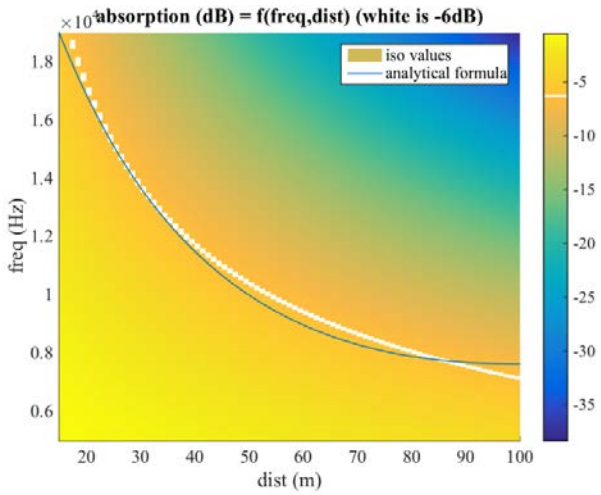


Figure 28. Air absorption and cut-off frequency from ISO9613-1 and the analytical formula used by the 3DTI Toolkit-BS.

The Butterworth filters implemented by the 3DTI Toolkit-BS result in a roll-off of 24dB/octave, which is a good approximation to the data reported in the ISO standard, as in Figure 29, where the filter is shown for a cut-off frequency of 10 kHz and a distance of 50 meters.

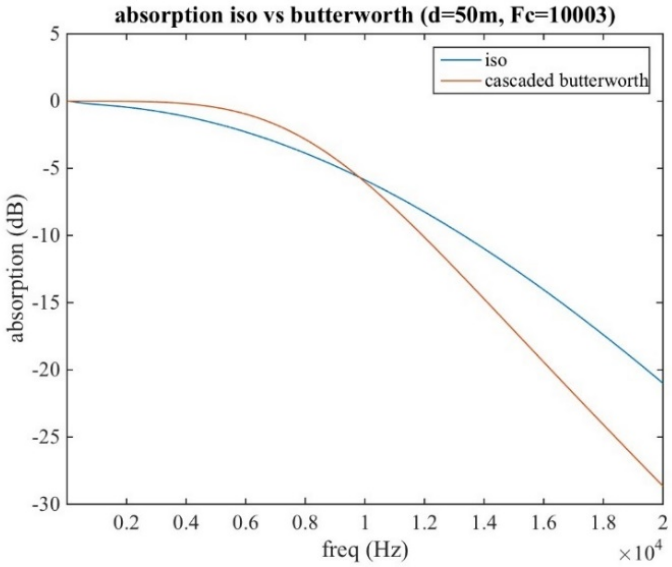


Figure 29. Air absorption as a function of frequency for 50 metres distance obtained from ISO9613-1 and the two cascaded second order Butterworth filter implemented in the 3DTI Toolkit-BS.

3.5.2 HRIR interpolation

The 3DTI Toolkit-BS allows the use of any HRTF saved in the SOFA format (Majdak et al., 2013). See Chapter 2, Section 2.3.3.2.4 for more information regarding this format. One of the advantages of reading standard SOFA files is the wide range of existing HRTFs databases that can be loaded (*SOFA General Purpose Database*, 2017), either directly provided as SOFA files, or by using the SOFA Matlab/Octave API (*SOFA Matlab/Octave API (Github)*, 2007) to convert different formats into SOFA. The 3DTI Toolkit SOFA reader is based on the Lib-Sofa library (Pompidou, 2014) for Linux/MacOS, which has been ported also to Windows.

As previously explained, an HRTF is composed by a set of HRIRs defined by the direction, azimuth (left-right) and elevation (up-down), where they were measured. The 3DTI Toolkit-BS allows for the use of any arbitrary distribution of HRIR measurement directions. This means that the renderer does not assume a regular or complete HRTF direction distribution and does not require any minimum density. Therefore, to spatialise sources located in the 3D space, the 3DTI Toolkit-BS needs to estimate HRIRs at the

source specific direction and distance, which may or may not be included in the set of measurements of the HRTF included in the SOFA file. This estimation is performed by interpolation with some known HRIRs in the surroundings of the desired direction. The interpolation of HRIRs is a widely investigated matter in binaural spatialisation and many mechanisms have been already proposed to perform HRIR interpolation.

Since the HRIRs are usually measured at a fixed distance, many methods use 2D (azimuth and elevation) interpolation for HRIR measurements, which considerably simplifies the interpolation mechanism. An example of 2D interpolation can be found in (Elizabeth M. Wenzel & Foster, 1993), which shows a perceptual comparison between non-interpolated and linearly interpolated HRIRs. Begault (1994) presented a bilinear interpolation among the HRIR in the four surrounding directions, assuming a regular distribution of measured positions. Later, Freeland et al. (2004) used a simplified version using three directions. Gamper (2013) presented a framework for interpolating HRIR measurements in 3D, taking into account both direction (azimuth and elevation) and distance of the source. He described and evaluated an algorithm which implements a tetrahedral interpolation with barycentric weights, obtaining good results for near-field HRIRs.

The interpolation can be performed in the time domain or in the frequency domain. Sodnik et al. (2005) presented an evaluation of the CIPIC HRIR library, carrying out a low-pass interpolation in the time domain to determine the smallest angle that can be distinguished by the listener. An example of interpolation in the frequency-domain can be seen in (Nishino et al., 1999), where linear interpolation and a spline interpolation were evaluated and compared. They applied the interpolations in the median plane, suggesting both methods where effective. Hartung et al. (1999) also carried out a comparison of different 2D HRTF interpolation methods using interpolation in the frequency domain with spherical splines, which provides much better results than interpolation in the time domain.

Other methods looked at decompositions based on principal component analysis (Carlile et al., 2000) and spherical harmonics (Alon et al., 2018; Romigh et al., 2015). Romigh et al. (2015) presented a method where the HRTFs are broken down into spherical harmonics, which allows to implement source movements and soundscape rotations directly in the spherical harmonic domain. This method requires a series of careful choices (e.g. the spherical harmonic order) in order to avoid aliasing, and other phase and frequency related problems (Brinkmann et al., 2017), which could create complications when allowing users to import their own HRTFs, with custom spatial resolution and non-uniform distribution. This kind of interpolation struggles when a large part of the HRTF sphere is missing. The bottom part of the HRTF is usually missing from measured HRTFs due to physical constraints. This can be mitigated with

regularization techniques (Duraiswaini et al., 2004) but it is still an important disadvantage comparing with bilinear interpolation. In addition, the advantage of linear interpolation over more advanced approaches, such as spherical splines or the use of spherical harmonics, is the reduced complexity regarding implementation and computation, which can be a decisive factor in simulation of real-time VAS.

In a similar way as (Gamper, 2013), but using 2D spherical coordinates (azimuth and elevation), the 3DTI Toolkit-BS carries out a barycentric interpolation in the frequency domain. The proposed approach is based on finding the closest three HRIR directions forming a triangle that encloses the desired HRIR direction and performing a barycentric interpolation. This allows the algorithm to deal with irregular distributions of HRIR but assuming that all measurements are at the same distance to the listener. In addition, this interpolation is carried out separately for both ears considering the parallax correction presented in the next section. Furthermore, this correction allows to assume that the source is located at a different distance from the fixed distance where the HRIR was measured.

Apart from the cross-ear parallax correction, in order to implement the interpolation approach, there are some concepts and methods that must be taken into account, which have been implemented and are presented below.

3.5.2.1 Cross-ear parallax correction

HRTFs are commonly measured at a single distance from the listener. In this way, the 3DTI Toolkit-BS considers that the loaded HRTF table contains a limited set of arbitrary azimuths and elevations but at a fixed distance, forming what we call the *HRTF sphere*. This distance (i.e. the radius of the HRTF sphere) is measured from the center of the listener head to the source and it is always indicated in the loaded HRTF SOFA file. When the source is not located at the distance where the HRIR was measured, the *acoustic parallax effect* occurs, as shown in Figure 30. The angle between the center of the head and the source (θ_C) differs from the angles between the sound source and each of the two ears (θ_L and θ_R). Therefore, when obtaining the HRIR from the indexed table, we have to take into account a different angle for each ear. This effect is present in near and far distances, being even much more relevant for near-field sources (D. S. Brungart & Rabinowitz, 1999).

To perform the binaural rendering of a source at any place in the 3D space and in order to select the most appropriate HRIR, the 3DTI Toolkit-BS implements a *cross-ear parallax correction* (Romblom & Cook, 2008) for each ear separately. This correction is based on calculating the projection of the vector from the ear to the source on the

HRTF sphere. Those projections are indicated in Figure 30 as $HRIR_L$ (left-HRIR) and $HRIR_R$ (right-HRIR). To calculate those directions, we used the Equations (3.4) and (3.5), written in cartesian coordinates, as it is shown in Figure 31.

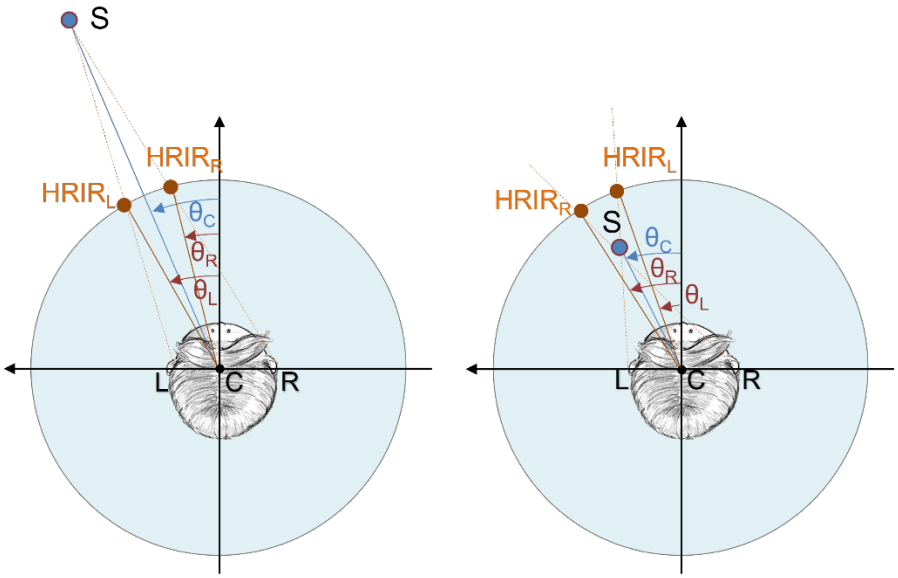


Figure 30. The acoustic parallax effect for a source further (left) and nearer (right) than the HRTF sphere. The coloured circle around the listener indicates the HRTF sphere. S shows the source position, located in the horizontal plane. θ_C the azimuth angle from the centre (C) of the head, θ_L the azimuth angle from the left ear (L) and θ_R the azimuth angle from the right ear (R).

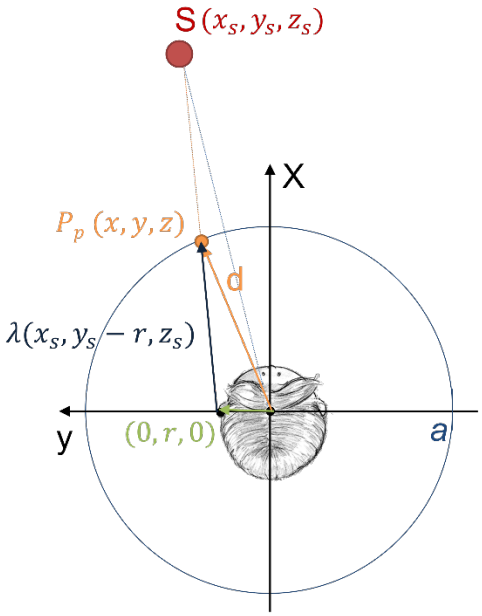


Figure 31. Parallax effect for the left-HRIR in cartesian coordinates.

S indicates the real direction of the source and P_p its projection on the HRTF sphere as seen from the left ear:

$$P_p = (x, y, z) = (0, r, 0) + \lambda(x_s, y_s - r, z_s) = (\lambda \cdot x_s, r + \lambda \cdot (y_s - r), \lambda \cdot z_s) \tag{3.4}$$

$$x^2 + y^2 + z^2 = a^2 \tag{3.5}$$

where $P_p = (x, y, z)$ is the direction of source projection on the HRTF sphere; r is the listener head radius; λ is the parameter that defines the line between the ear and the sphere, passing by the source; (x_s, y_s, z_s) is the real position of the source and a is the radius of the HRTF sphere. The figure and the equations show the example of the position of the $HRIR_L$ for a source placed in the far-field (distance larger than 2 meters), but it is valid for the right ear and sources inside the HRTF sphere as well.

To obtain λ, we match the two previous equations, where all the variables are known but λ, obtaining following quadratic equation:

$$(\lambda \cdot x_s)^2 + \big(r + \lambda \cdot (y_s - r)\big)^2 + (\lambda \cdot z_s)^2 = a^2 \tag{3.6}$$

where we can solve λ as follows

$$\lambda^2(x_s^2 + (y_s - r)^2 + z_s^2) + \lambda \cdot \big(2 \cdot r \cdot (y_s - r)\big) + (r^2 - a^2) = 0 \tag{3.7}$$

$$\lambda = \frac{-2 \cdot r \cdot (y_s - r) + \sqrt{-4(x_s^2 + (y_s - r)^2 + z_s^2)(r^2 - a^2)}}{2(x_s^2 + (y_s - r)^2 + z_s^2)} \tag{3.8}$$

And finally

$$P_p = (x, y, z) = (\lambda x_s, r + \lambda(y_s - r), \lambda z_s) \tag{3.9}$$

3.5.2.2 HRIR and initial delay interpolated separately

Using nearby HRIRs to get the interpolated HRIR means that we have to perform operations between HRIRs with similar modules but different phases (different initial delays). This can cause what is known as the *comb filter effect* (Elizabeth M. Wenzel & Foster, 1993; F. Wightman et al., 1992). Comb filtering is produced when adding two audio signals with similar magnitude but different phase, which causes that the resulting signal is either amplified or attenuated at certain frequencies. In this way, the frequency

response consists of a series of regularly spaced notches, giving the appearance of a comb. This effect results in audible coloration, reducing the rendering quality.

To reduce the comb filter effect, the 3DTI Toolkit-BS perform interpolations between synchronized HRIRs, which means that the renderer carries out separately the interpolation of the initial delay and the HRIR, using HRIRs where the initial delay has been previously extracted. There are multiple ways to calculate the initial delay or ITD (when considering both ears) and remove it from the HRIR (Katz & Noisternig, 2014). The 3DTI Toolkit-BS does not implement this extraction, assuming instead that the loaded HRIRs are synchronized. How the 3DTI Toolkit-BS manage the initial delay and adds it after the interpolation process will be described in Section 3.5.3.

3.5.2.3 The barycentric interpolation

First of all, to perform the interpolation, we need to know the three nearest directions to the desired direction where the source is located. To do so, we need to calculate the distance between two points on a sphere surface. The Haversine Formula (C. C. Robusto, 1957) computes the distance between points $P_p(\theta^{P_p}, \phi^{P_p})$ and $P_i(\theta^{P_i}, \phi^{P_i})$, expressed in spherical coordinates.

$$\sigma_{P_p,P_i} = 2r \cdot arccos\left(\sqrt{\sin^2\left(\frac{\phi^{P_p} - \phi^{P_i}}{2}\right)} + \cos\phi^{P_p} \cdot \cos\phi^{P_i} \cdot \sin^2\left(\frac{\theta^{P_p} - \theta^{P_i}}{2}\right)\right) \tag{3.10}$$

Using this formula, we will sort all HRIRs from nearest to farthest from the desired point, according to the values of σ_{P_p,P_i}, where P_p indicates the direction of the desired HRIR and P_i the direction of each HRIR contained in the HRTF table (i.e. $i \in [1, number\ of\ HRIR\ measurements]$). The three nearest HRIRs will be the first three in the list, but these three points also have to fulfill the following condition: they must form a triangle around the desired point, as will be explained later.

The barycentric interpolation consists in calculating the desired HRIR at P_p using the barycentric coefficients ($\alpha, \beta\ and\ \gamma$) calculated with the position of three nearest HRIRs ($P_1, P_2\ and\ P_3$). These three points have to form a triangle such that the target position P_p is inside it, as it is shown in Figure 32.

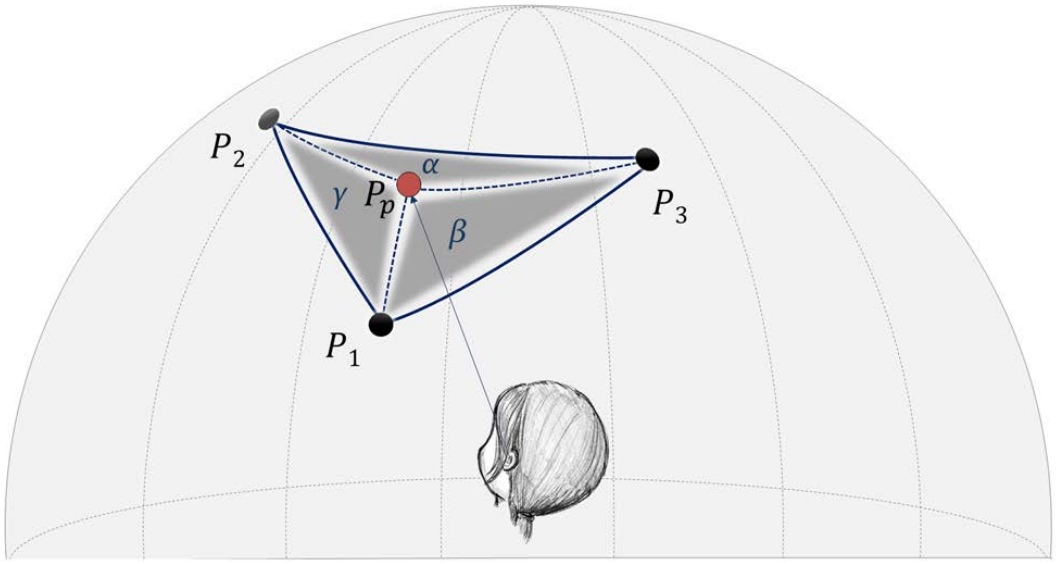


Figure 32. Example of barycentric interpolation in the HRTF sphere surface for the left ear.

The barycentric coefficients are calculated with the following equations (3.11):

$$\begin{aligned}\alpha &= \frac{(\phi^{P_2}-\phi^{P_3})\cdot(\theta^{P_p}-\theta^{P_3})+(\theta^{P_3}-\theta^{P_2})\cdot(\phi^{P_p}-\phi^{P_3})}{(\phi^{P_2}-\phi^{P_3})\cdot(\theta^{P_1}-\theta^{P_3})+(\theta^{P_3}-\theta^{P_2})\cdot(\phi^{P_1}-\phi^{P_3})}\\ \beta &= \frac{(\phi^{P_3}-\phi^{P_1})\cdot(\theta^{P_p}-\theta^{P_3})+(\theta^{P_1}-\theta^{P_3})\cdot(\phi^{P_p}-\phi^{P_3})}{(\phi^{P_2}-\phi^{P_3})\cdot(\theta^{P_1}-\theta^{P_3})+(\theta^{P_3}-\theta^{P_2})\cdot(\phi^{P_1}-\phi^{P_3})}\\ \gamma &= 1-\alpha-\beta\end{aligned} \qquad (3.11)$$

where $(\theta^{P_p},\phi^{P_p})$ represents the direction of P_p (azimuth and elevation respectively), and $(\theta^{P_1},\phi^{P_1})$, $(\theta^{P_2},\phi^{P_2})$ and $(\theta^{P_3},\phi^{P_3})$ the directions of the nearest HRIRs, of P_1, P_2 and P_3.

To guarantee that P_1, P_2 and P_3 form a triangle around P_i the barycentric coefficients have to meet the following condition: $\alpha \geq 0$, $\beta \geq 0$ and $\gamma \geq 0$

In this way, the implemented procedure was as follows:

1. Get the three first positions of the sorted list calculated with the Haversein formula presented in equation (3.10).
2. Calculate the barycentric coefficients using the equation (3.11)
3. Check if these points meet the condition $\alpha, \beta, \gamma > 0$
4. If yes, these are the barycentric coefficients and the three points used for the interpolation in the Equation (3.12).
5. If not, the next point of the sorted list is selected, together with all the previous ones. With this list of candidate's points, we do combinations of three points and calculate again the barycentric coordinates until we meet the condition $\alpha, \beta, \gamma \geq$

0. If we cannot meet the condition, we introduce a new point in the candidate point list and repeat this step 5 again, until we meet the condition $\alpha, \beta, \gamma \geq 0$.

Once we have the three nearest points that form a triangle around the desired position, we use the following expression (3.12) to get the interpolated $HRIR^{P_i}$, the calculation is carried out in the frequency domain where $HRIRs$ are stored as complex numbers:

$$HRIR^{P_p} = \alpha \cdot HRIR^{P_1} + \beta \cdot HRIR^{P_2} + \gamma HRIR^{P_3} \quad (3.12)$$

And evaluation of this technique, where some results of the interpolation are represented, is described in Chapter 4.

3.5.2.4 HRIR interpolation process implemented by the 3DTI Toolkit-BS

In practice, finding the nearest HRIR in an arbitrary set is an expensive process, due to the high number of operations that must be carried out (calculations, memory accesses, etc.). To reduce this problem during the run-time processing, the 3DTI Toolkit-BS divided the interpolation process into two separate parts. A first process performed *off-line* and a second one performed *on-line*. The *off-line* process is in charge of generating a resampled-HRTF table where all HRIRs are equally spaced in azimuth and elevation. The main purpose of this first process is to get a complete and regular HRTF table to simplify and accelerate the second part of the process, which is performed *on-line* in real-time and takes into account the position of the sound source.

- **Off-line interpolation**

The whole off-line process is shown in Figure 33. First, an off-line interpolation is performed, resulting in a resampled HRTF table in the time domain, where all the HRIRs are distributed in a regular grid with a configurable spacing step in elevation and azimuth. Secondly, the partitions and FFTs needed to perform the UPOLS convolution are performed. The result of this process will be a regular and complete table where the HRIRs are distributed uniformly.

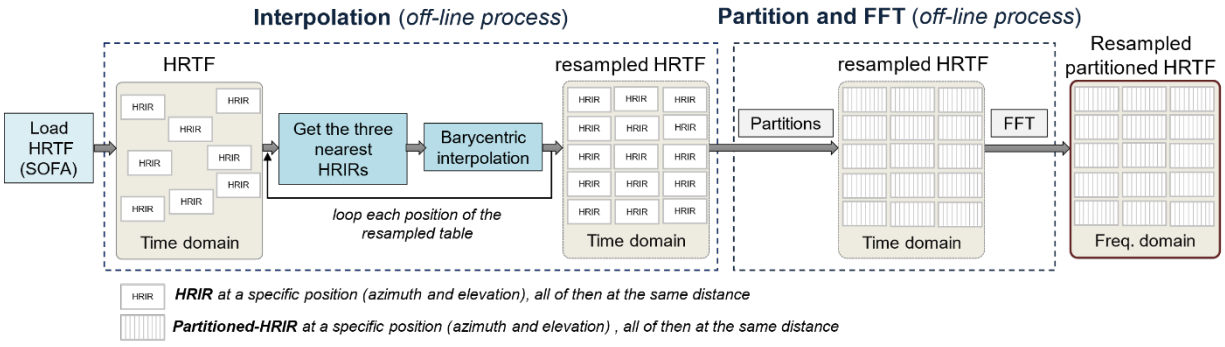


Figure 33. Process to create the resampled HRTF table in the frequency domain with partitioned HRIRs. Each box of the HRTF table represents the HRIRs for different directions (azimuth and elevation).

When measuring HRTFs on a spherical grid, the measurement directions are typically distributed more sparsely toward the poles than at the equator, in accordance with the decreasing localization accuracy of humans toward extreme elevations. In addition, measurements at the sphere poles ($\phi = 90°$ and $\phi = -90°$) are sometimes missing. In this way, if no HRIR measurement is available for the pole direction, a linear interpolation is used to obtain an *HRIR estimation in the missing polar direction*. After the HRIR is calculated at the poles, the algorithm starts an iterative process where it goes through the whole HRTF table step by step, calculating the HRIR at the regularly distributed directions to complete the regular table. For each new HRIR direction, firstly it is necessary to *get the three nearest HRIR around the new direction*. To do this, the method calculates the distance from the direction of the new HRIR to all other points in the HRTF table using the algorithm described in Section 3.5.2.2. Finally, the algorithm takes the three HRIRs found at the minimum distance and calculates the new HRIR with a *barycentric interpolation*, explained in Section 3.5.2.4.

After the interpolation process, each HRIR is partitioned in blocks to match the input buffer length (N), in order to use the modified UPOLS convolution described in Section 3.4. Finally, the FFT is applied to each of the HRIR partitions.

- **On-line interpolation**

The on-line interpolation process starts calculating the specific direction of the HRIR (P_p), taking into account the *parallax effect* explained Section 3.5.2.1. Once the direction on the HRTF sphere for each ear is obtained, if there is no available HRIR for that specific direction on the resampled HRTF table, the 3DTI Toolkit-BS performs again the interpolation. This is now a much simpler operation, as the resampled HRTF table has a regular basis where the HRIR directions are known in advance. Therefore, it is not necessary to run any algorithm to calculate the distance between different directions in the table. We just need to get the quadrant where the desired point is positioned. To do so, we calculate $P_A(\theta^{P_A}, \phi^{P_A})$, $P_B(\theta^{P_B}, \phi^{P_B})$, $P_C(\theta^{P_C}, \phi^{P_C})$ and $P_D(\theta^{P_D}, \phi^{P_D})$ of Figure 34 with the following equations, where k is the step used to calculate the resampled table.

$$\theta^{P_C} = \left\lfloor \frac{\theta^{P_p}}{k} \right\rfloor \cdot k \tag{3.13}$$

$$\phi^{P_C} = \left\lfloor \frac{\phi^{P_p}}{k} \right\rfloor \cdot k \tag{3.14}$$

$$\theta^{P_A} = \theta^{P_C} \tag{3.15}$$

$$\phi^{P_A} = \phi^{P_C} + k \tag{3.16}$$

$$\theta^{P_B} = \theta^{P_C} + k \tag{3.17}$$

$$\phi^{P_B} = \phi^{P_A} \tag{3.18}$$

$$\theta^{P_D} = \theta^{P_B} \tag{3.19}$$

$$\phi^{P_D} = \phi^{P_C} \tag{3.20}$$

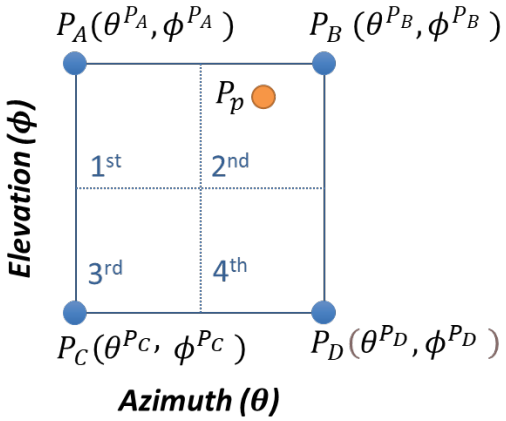


Figure 34. A portion of the resampled table, divided in four different quadrants, where the desired position (P_p) is surrounded for 4 different points (P_A, P_B, P_C and P_D where the HRIR are known.

According to the values of azimuth and elevation of $P_p(\theta^{P_p}, \phi^{P_p})$, we can know in which quadrant the desired point is placed, and use the three already known points that form a triangle around it. With these three points we calculate the barycentric coefficients using Equation (3.11) and finally the $HRIR^{P_p}$ using Equation (3.12).

Finally, the on-line method selects the three nearest points from the resample HRTF table (i.e. form a triangle containing P_i) and performs a barycentric interpolation (presented in Section 3.5.2.4) among the HRIRs corresponding to these three directions. Again, this interpolation is performed using partitioned HRIRs in the frequency-domain

and obtained the HRIR partitions that will be convolved with the input signal as it was explained in Section 3.4.

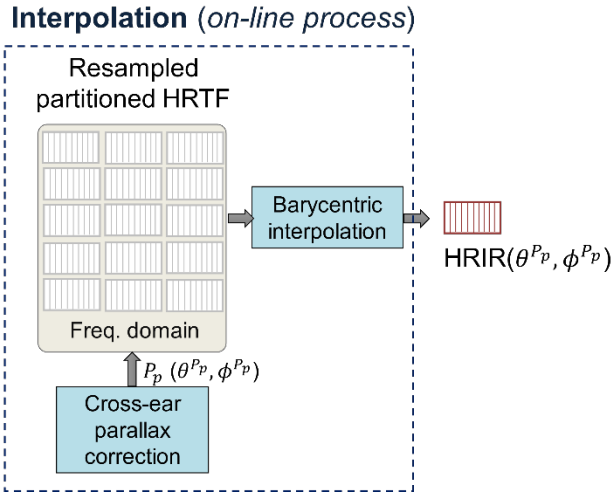


Figure 35. Run-time processing to obtain the HRIR for the specific position of the source ($HRIR(\theta^{P_p}, \phi^{P_p})$)

3.5.3 ITD simulation

As explained in Section 3.5.2.3, the 3DTI Toolkit-BS manages the initial delay or ITD (when considering both ears) separately from the interpolation processes. This allows to simulate a customized ITD, different than the one included in the HRIR and closely to the individual ITD of the listener. Figure 36 shows the implemented approach.

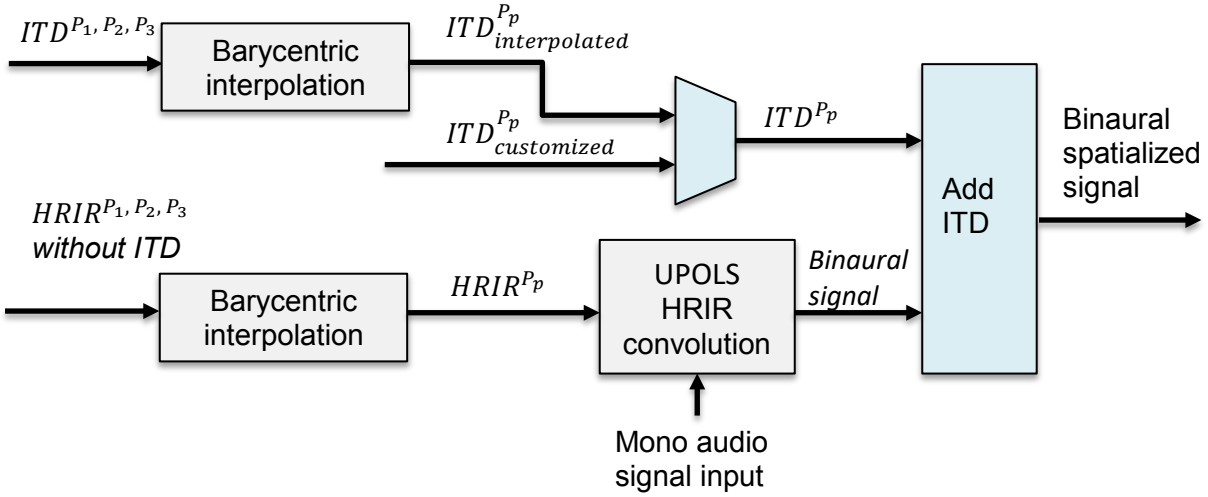


Figure 36. Interaural Time Difference simulation process.

As it is shown in Figure 36, the ITDs for the specific position of the source (ITD^{P_p}) can be estimated in two different ways. In the first case, the 3DTI Toolkit-BS uses the

barycentric interpolation among those corresponding to the three nearest HRIRs, described in Section 3.5.2.4, but this time employing the exact azimuth and elevation of the sound source (i.e. without taking into account the acoustic parallax effect). In the second case, the ITD is synthesized using data about the direction of the sound source and the head circumference of the listener. The ITD is customized for the specific listener and it is calculated using Equation (3.21), originally developed by Woodworth (Woodworth et al., 1954):

$$ITD_{customized} = \frac{a}{c}(\theta_I + \sin\theta_I) \tag{3.21}$$

where a is the listener's head circumference, c is the speed of sound and θ_I is the interaural azimuth. The interaural azimuth is calculated using the following equation, where θ and ϕ are the polar azimuth and elevation respectively.

$$\theta_I = \text{asin}(\sin(\theta) \cdot \cos(\phi)\,) \tag{3.22}$$

Once the ITD^{P_s} has been estimated, it is added to the binaural audio that results from the UPOLS convolution (explained in Section 3.4.1) between the input audio signal and the $HRIR^{P_p}$ for both ears.

To add the ITD^{P_p} to the $HRIR^{P_p}$, the 3DTI Toolkit-BS works adding a delay only to the $HRIR^{P_p}$ of the contralateral ear, in the following way.

$$D_{left} = \begin{cases} ITD^{P_p}, & -\frac{\pi}{2} < \theta_I < 0 \\ 0, & 0 < \theta_I < \frac{\pi}{2} \end{cases}$$
$$D_{right} = \begin{cases} 0, & -\frac{\pi}{2} < \theta_I < 0 \\ ITD^{P_p}, & 0 < \theta_I < \frac{\pi}{2} \end{cases} \tag{3.23}$$

3.5.3.1 ITD smoothing mechanism

The ITD depends on the relative position between the listener and the source, so the delay added to the signal (D_{left} *and* D_{right} of equation (3.23)) is not always the same. To add the corresponding delay to the audio signal, the 3DTI Toolkit-BS implements a method to squeeze or stretch the signal, i.e. the difference between the delay of a previous frame and the next frame will be solved expanding or compressing the signal samples.

Figure 37 scheme the process in a most generic way, showing the buffer sizes. In this example, the current delay to be added to the frame (D_i) is larger than the delay of the previous frame (D_{i-1}). This caused the input buffer to be stretched in order to fill out part of the output buffer ($N - D_{i-1}$) and the new delay (D_i), which will be stored to be added in the following frame. The extended buffer is re-sampled by linear interpolation among samples, as explained below.

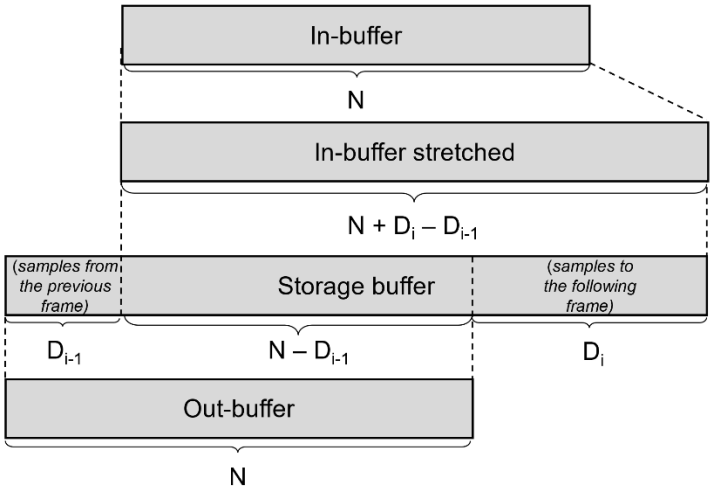


Figure 37. Behavior of the samples in a frame during the stretching algorithm for adding a delay to the contralateral ear.

Figure 38a shows an example when the new delay is larger than the previous delay, and thus the input frame has to be expanded, by distributing its samples between the output frame and the new delay (which will be output in the next frame). Figure 38b shows an example when the new delay is smaller than the previous delay, and thus the input frame has to be compressed, squeezing its samples to fit in output frame and the new delay.

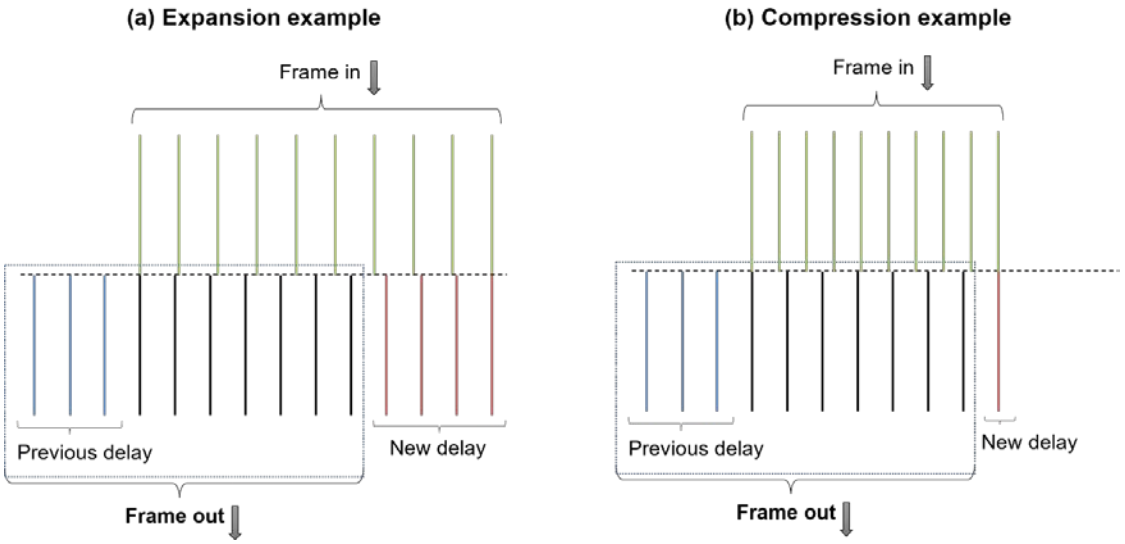


Figure 38. Examples that show when the 3DTI Toolkit-BS applies an expansion or a compression algorithm. The sticks represent samples.

As shown in the example, the first sample of the input signal and the last sample are copied in the new buffers with any modification. The rest of the samples will be calculated with a lineal interpolation of the two nearest samples. The interpolation will be based on a weighted algorithm where the new value will be calculated with the weighted sum of the two nearest samples. The weight will be calculated with the distance between the new samples to each of the two nearest samples. Figure 39 shows an example of how to get the value of a specific output-sample.

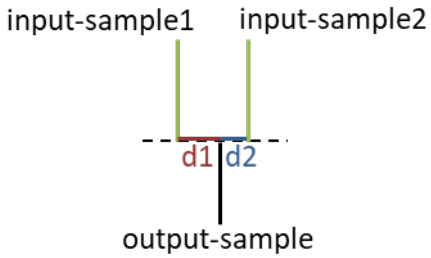


Figure 39. Lineal interpolation to get the output-sample value

The output-sample (y) is calculated with the following lineal interpolation expression, using the values of the input-sample 1 (x_1) and input-sample 2 (x_2):

$$y = \frac{d2 \cdot x_1 + d1 \cdot x_2}{d1 + d2} \tag{3.24}$$

3.5.4 Near-field HRTF compensation (ILD correction)

The sound of a nearby source contains a set of special cues that are interpreted by the brain in order to locate these types of sources (Shinn-Cunningham, 2000). There are databases that provide HRTF measured at the near field. Some of them were presented in Chapter 2, Section 2.3.5.1.2. However, these measurements are not very common since the procedure and later storage of HRTFs at multiple distances is rather impractical and time-consuming. Instead, some corrections to far field HRTF are usually carried out to simulate sources in the near field.

The 3DTI Toolkit-BS simulates sources in the near-field implementing a *compensation for the HRTF* conventional processing, where the main algorithm is the convolution with an HRTF measured at a fixed distance. The Toolkit considers sources in the near fields when they are located at *distances lower than 2 meters* to the listener's head. The implemented approach uses the model presented by Romblom & Cook (2008). It relies on a *difference filter* that predicts the spectral differences between a near-field source and a source placed at the same direction (in this case the interaural azimuth is used, which depends on the spherical azimuth and elevation as it is shown in Equation (3.22)), but at the distance where the HRTF was measured.

This *difference filter* is based on a Spherical Head Model (SHM) presented in (Duda & Martens, 1998). They solved the analytic problem of an incident wave scattered by a solid sphere (which simulates the listener's head). When we take into account both ears we can consider that we are talking about the ILD (Interaural Level Difference). ILDs are caused by the head shadow on the contralateral ear. This effect, which is included in the HRTF, occurs at all distances and it is more relevant for high frequencies. However, for sources at near distances, the effect of the head's shadow on the source spatialisation is larger and affects the whole range of frequencies. The ILD for a SHM (ILD_{SHM}) for a specific direction (horizontal plane, azimuth 100 degrees) is shown in Figure 40 from the work of Duda & Martens (1998). These graphs show how the low-frequency ILD increases when the source approach to the listener (the value of source distance approaches the head radius, $\rho = 1$ in Figure 40).

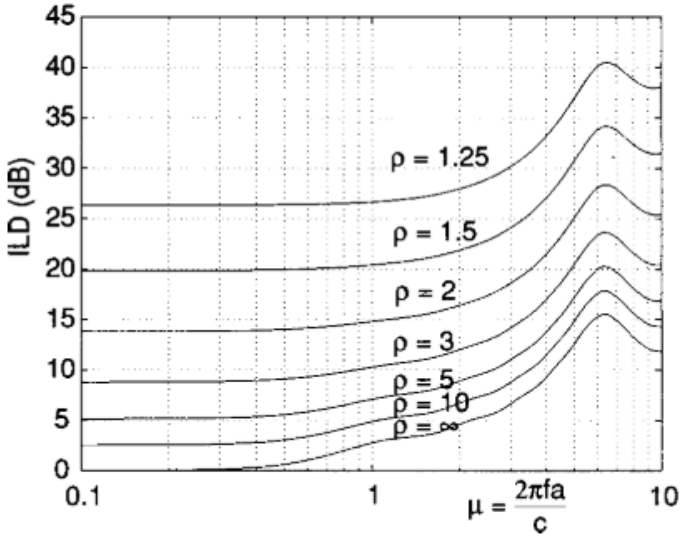


Figure 40. ILD_{SHM} for a sound source at $\theta = 100°, \phi = 0°$. This image is from Duda & Martens (1998), where ρ, which is calculated as $\frac{d}{a}$, is the distance from the source to the centre of the listener head (d), normalized with the radius of the listener head (a). The horizontal axis is the normalized frequency using the radius of the listener head (a) and the sound speed (c).

In this way, the difference filter (called $ILD_{difference}$) is calculated as the ratio between the filter given by the model at the source distance (d) and the filter given by the same model in the distance where the near-filed is considered to star, in our case at 2 meters, as it is shown in Equation (3.25).

$$ILD_{difference}(d, \theta_I) = \frac{ILD_{SHM}(\theta_I, d)}{ILD_{SHM}(\theta_I, 2m.)} \tag{3.25}$$

These difference filters are implemented as IIR filters adjusted to match the described transfer function. The 3DTI Toolkit-BS includes two biquad filters for each ear, where the coefficients for these filters depend on both the distance of the sound source (d) and its interaural azimuth (θ_I). These filters are pre-calculated and stored in a file as a look-up table. Each entry of the look up table provides 10 coefficients that will be applied to the two biquad filters of each ear. The table is indexed by parameters (d) and (θ_I) with range and step configurable. By default, it is configured with distance ranges from 10 cm to 2 m stepping by 1cm and the azimuth angle ranges from 0° to 355° stepping 5°. This process can be considered as an HRIR correction since it is applied in series with the HRIR selected and interpolated in the previous stages of the pipeline.

3.5.4.1 Near-field smoothing mechanism

When the binaural spatialisation is performed, and as mentioned in previous described algorithms, a problem arises when the source or the listener are moving, since some audible artefacts can appear in the signal. In this particular case, those artefacts can be caused as the near-field correction filters have to change from frame to frame. In order to minimise this problem, at every frame each biquad filter is applied using both the previous and the new coefficients, and a *linear cross-fading* is performed to produce the output. This approach is not particularly expensive, as these filters, which are implemented in the IIR canonical form, require only two delay cells and a minimum number of operations. The evaluation of this mechanism is shown in Section 4.3.

3.6 Reverberation path

Chapter 2, Section 2.3.5.4, described different methods to simulate sources in enclosed spaces, i.e. to simulate the acoustic reverberation caused by a room wall, floor, ceiling, etc. The process implemented by the 3DTI Toolkit-BS in the reverberation path is outlined in Figure 41. The chosen approach is based on *virtual Ambisonics* and *convolution with BRIRs*.

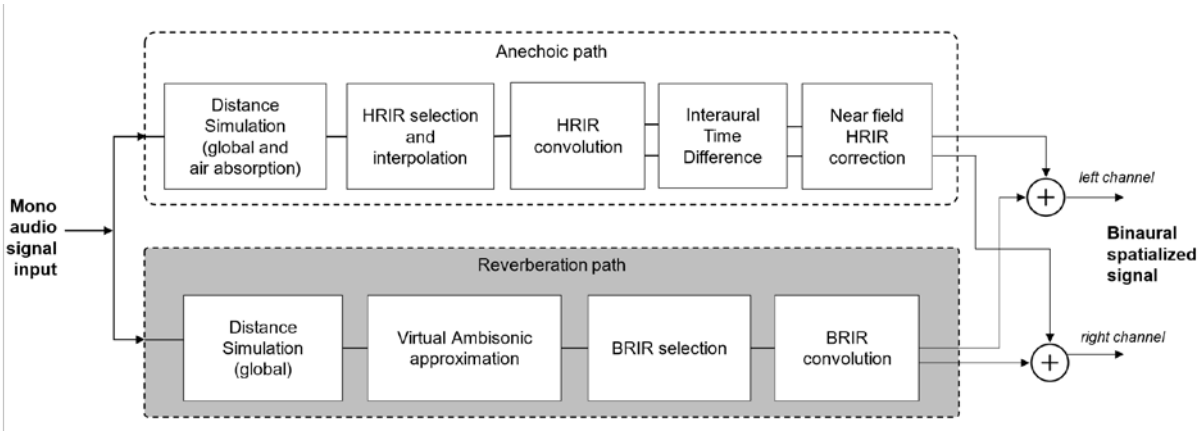


Figure 41. Low level 3DTI Toolkit-BS process structure

The *virtual Ambisonic* solution was presented by McKeag & McGrath (1996a), when they were investigating the problem of HRTF interpolation and the computational load caused for real-time audio rendering. A similar approach was also followed by Noisternig et al. (2003). Their approach consists in spatialising the sound sources using Ambisonics and a reduced set of BRIR convolutions. First, all input signals are encoded into a set of Ambisonics channels, determined by the Ambisonics order configured. Then, the Ambisonics channels are decoded in a set of virtual speakers (from SPK_1 to SPK_k) distributed around the "virtual listener" in different directions. Finally, the SPK signals

are convolved with the BRIR corresponding to the SPK directions, creating the binaural signals that will be delivered to the listener by the headphones. The process is illustrated in Figure 42:

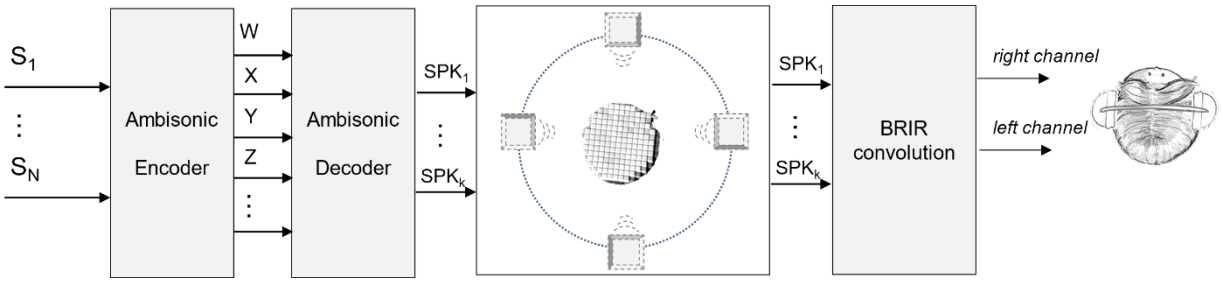


Figure 42. Virtual Ambisonic approach

For multiple moving sources, this is an efficient way to create spatialised sounds for enclosed spaces, since it does not depend on the number of sources. All sources are encoded together using the Ambisonic approximation, which maintain a certain characteristic regarding the location of the sources, as explained in Section 1.2.1.2. In addition, the required BRIRs do not depend on the location of the source and therefore there is no need of interpolation.

As described in Section 3.2, in the 3DTI Toolkit-BS the anechoic and reverb paths are simulated in two parallel process lines. Since the anechoic path already simulates the direct sound by HRTF convolutions, BRIRs used in the reverberation process should not contain the direct sound, but only the reflections. In this way, BRIRs are preceded by a set of zeros in order to maintain the appropriate timing between the beginning of the impulse response and the appearance of the first reflections. The number of zeros will depend on the type of environment where the BRIRs were measured/synthesized, and on the position of both the source and the listener's microphones. A simple method for removing the direct sound from the BRIRs is to geometrically estimate the delay of the arrival of the first reflection, using this information to remove the signals that appears before that within the BRIRs, and adding an equivalent number of zeroes.

The reverberation process implemented by the 3DTI Toolkit-BS is shown in Figure 43. The process performs a 1st order Ambisonic approach and uses six virtual speakers. In addition, it presents a further optimization (explained below) to the previously described virtual Ambisonic approach.

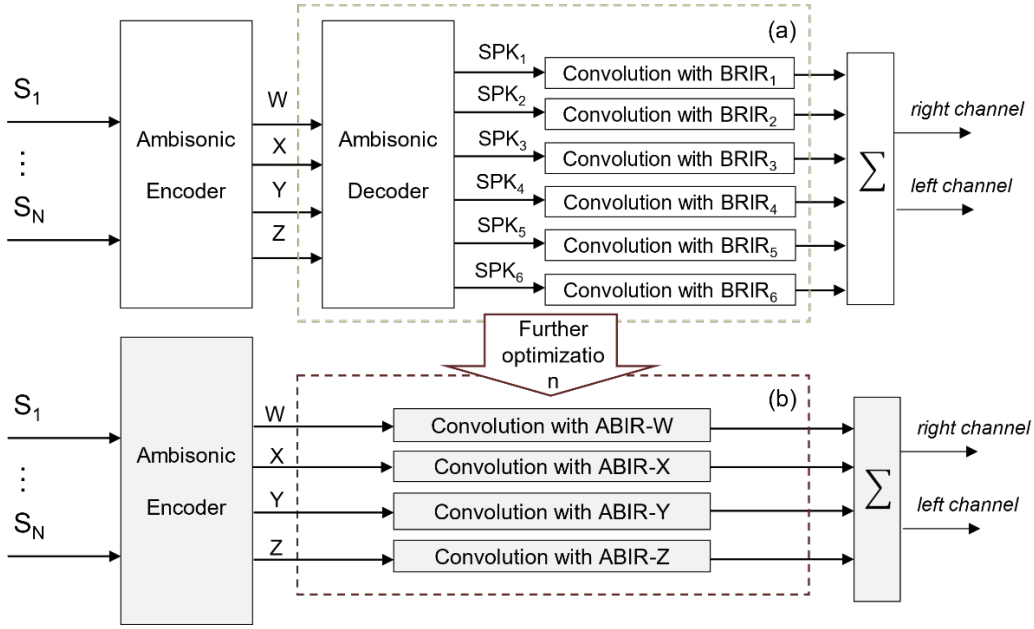


Figure 43. **Reverberation sound simulation process.** The top box (a) represents the Ambisonic decoding and the six convolutions needed for the six virtual speakers' signals (SPK_1-SPK_6). The bottom box (b) shows the optimization performed by the 3DTI Toolkit-BS, which reduces the number of needed convolutions to four.

The upper box (a) in Figure 43 follows the same approach showed in Figure 42. In this process every sound source (S_1-S_N) is encoded into a 1st order Ambisonic format, where the directional information of the entire sound-field is included into B-format Ambisonic channels (W, X, Y and Z). These channels are then decoded into six virtual speakers located at the vertices of an octahedron, at the directions shown in Table 4. The equations used for the Ambisonic encoder and decoder are shown in Table 5. Finally, the virtual speakers' signals (SPK_1-SPK_6) are converted to the reverberant binaural domain by convolving them with the BRIR corresponding to each of the speaker's directions. Each of the convolutions with the BRIR are actually two convolutions, one with the left-BRIR and one with the right-BRIR. As with the HRTF, here we will talk in general about convolution with the BRIR.

Table 4. Virtual speakers' location (azimuth and elevation) around the listener.

$\boldsymbol{SPK_1}$	$\theta = 0°, \phi = 0°$	$\boldsymbol{SPK_3}$	$\theta = 180°, \phi = 0°$	$\boldsymbol{SPK_5}$	$\phi = 90°$
$\boldsymbol{SPK_2}$	$\theta = 90°, \phi = 0°$	$\boldsymbol{SPK_4}$	$\theta = 270°, \phi = 0°$	$\boldsymbol{SPK_6}$	$\phi = 270°$

Table 5. 1st order Ambisonic encoder and decoder equations.

Encoder equations	Decoder equations
$W = S \cdot \frac{1}{\sqrt{2}}$ $X = S \cdot \cos\theta \cdot \cos\phi$	$SPK_1 = W + X$ $SPK_2 = W + Y$

$Y = S \cdot \sin\theta \cos\phi$ $Z = S \cdot \sin\phi,$ where $S = \sum_{j=1}^{N} S_j$	$SPK_3 = W - X$ $SPK_4 = W - Y$ $SPK_5 = W + Z$ $SPK_6 = W - Z$

Both Ambisonic decoding and convolution are linear processes and can be combined in order to simplify the previously described process. The 3DTI Toolkit-BS introduces the optimization presented in the bottom box (b) of the Figure 43. The B-Format channels (W, X, Y and Z) are directly convolved with what is called ABIRs (Ambisonic-to-Binaural Impulse Responses). ABIRs are obtained through the off-line process shown in Figure 44.

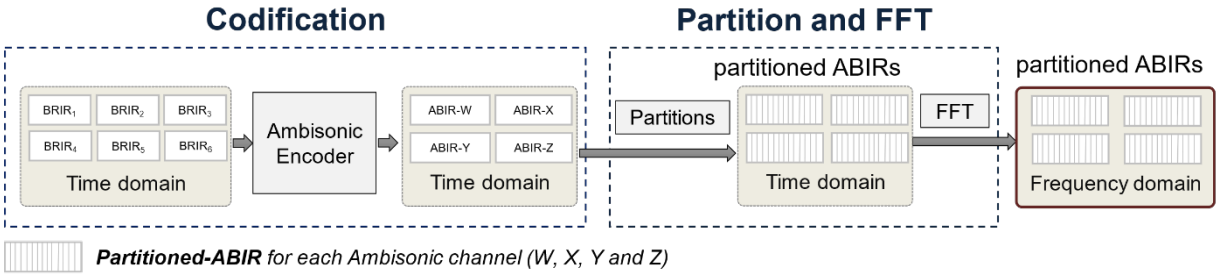


Figure 44. Off-line process to get the ABIRs ready to convolve with the input signal with the UPOLS convolution method.

First, in the codification process, the six BRIRs, corresponding with each position of the six virtual speakers, are encoded into B-format Ambisonic channels, in the following way:

Table 6. B-format Ambisonic channels of the ABIRs

Encoder equations to create ABIRs
$ABIR\_W = \frac{1}{\sqrt{2}}(BRIR_{SPK_1} + BRIR_{SPK_2} + BRIR_{SPK_3} + BRIR_{SPK_4} + BRIR_{SPK_5} + BRIR_{SPK_6})$
$ABIR\_X = BRIR_{SPK_1} - BRIR_{SPK_3}$
$ABIR\_Y = BRIR_{SPK_2} - BRIR_{SPK_4}$
$ABIR\_Z = BRIR_{SPK_5} - BRIR_{SPK_6}$

Using this approach, the number of stereo convolutions is reduced from six to four (from the number of virtual speakers to the number of Ambisonic channels). This approach was originally introduced by (McKeag & McGrath, 1996), but in that case it was used to compute anechoic binaural spatialisation, and not BRIR-based reverberation. In (Picinali et al., 2014, 2017) a similar technique was used for computing real-time binaural reverberation, but no separation between the direct sound and reflections was considered.

Since ABIRs are usually rather large, this process uses the convolution method based on partitions presented and described in Section 3.4.1, the Uniformly Partition Overlap-Save (UPOLS) convolution in the frequency domain. To do so, the second part of the off-line process consists of performing the partitions of each ABIR and the conversion into the frequency domain.

Using the presented technique based on virtual Ambisonics and ABIR convolution, the 3DTI Toolkit-BS can handle reverberation for an unlimited number of moving sources in a full three-dimensional space, performing only four ABIR convolutions. Using first order Ambisonics reduce the number of real-time convolutions but presents the limitation of inaccurately simulating sources positions. However, a perceptual study presented by Picinali et al. (2017) suggested that first-order Ambisonics was able to simulate the reverberation of a room which were indistinguishable from higher order simulations. The performance of this part of the process will be evaluated in detail in Section 4.4.

The 3DTI Toolkit-BS implements two additional configurations to simulate the reverberation path, which are differentiated according to the number of Ambisonic channels used:

- The *three-channel configuration* have been implemented to support those situations where the BRIRs are measured only on the horizontal plane, which is a rather common case and means that we can only use four virtual speakers ($SPK_{1\text{-}4}$). In this case, only the W, X and Y B-format channels are computed, resulting in a reduction of the number of ABIR convolutions to three. When the source to be spatialised is located outside the horizontal plane, i.e. with an elevation different to 0 degrees, in order to avoid the loss of power due to the absence of virtual speakers above and below the listener ($SPK_{6,7}$), the elevation, that should be encoded in the Z-channel, is computed in the X-channel using the following equations, where θ and ϕ are the azimuth and elevation of source S_j:

$$W = \sum_{j=1}^{N} S_j \frac{1}{\sqrt{2}} \tag{3.26}$$

$$X = \sum_{j=1}^{N} S_j \cdot (\cos\theta_j \cdot \cos\phi_j + \sin\phi_j) \tag{3.27}$$

$$Y = \sum_{j=1}^{N} S_j \cdot \sin\theta_j \cdot \cos\phi_j \tag{3.28}$$

Then, the ARIRs are obtained using the following equations. The W channel is also multiplied by 1.5 due to the absence of two loudspeakers (the overall power of six loudspeakers must be distributed in four speakers).

Table 7. B-format Ambisonic channels of the ABIRs for the three-channel configuration

Encoder equations to create ABIRs
$ABIR\_W = \frac{1}{\sqrt{2}} \cdot 1.5 \cdot (BRIR_{SPK_1} + BRIR_{SPK_2} + BRIR_{SPK_3} + BRIR_{SPK_4})$
$ABIR\_X = BRIR_{SPK_1} - BRIR_{SPK_3}$
$ABIR\_Y = BRIR_{SPK_2} - BRIR_{SPK_4}$

- The *one-channel configuration* is included to allow efficient reverberation simulation reducing the number of convolutions to one only. In this case, the 3DTI Toolkit-BS uses only the W channel (the omnidirectional one), convolving it with a single BRIR, obtained by averaging all the available BRIRs.

Due to the fact that a fixed number of BRIRs are used for the reverberation simulation, and that these do not change depending on the location of the sound source, in order to simulate sources located close to boundaries within a reverberant environment (e.g. close to a wall) user-imported BRIRs need to be measured/synthesized from those locations/conditions. In addition, the current implementation of the 3DTI Toolkit-BS reverberation path is limited to 1st order Ambisonic, but further implementations of the same method are possible increasing the Ambisonic order, and consequently the spatial resolution. As mentioned before, this method allows to simulate multiple moving sources, however, the main drawback of this method is the fact that the relative position between the listener inside the room is blocked. Since the virtual speakers are always situated at the same location with respect to the listener, this implies that, when the listener changes their position or orientation, the room will change in the same way, keeping the relative position with the listener. This means that when the listener rotates their head, all the virtual environment is rotated in the same way, while the relative position between the sources and the listener changes. However, these "differences" are not noticeable by the listener (Engel et al., 2021), thanks to the fact that this movements are taking into account to simulate the direct path.

3.7 Releases and additional tools

The 3DTI Toolkit is available on a public GitHub repository (https://github.com/3DTune-In/3dti_AudioToolkit) under a GPLv3 license. All algorithms described before have been implemented from scratch, creating a C++ library

which uses only platform-independent code from the standard template library. The only third-party library integrated in the 3DTI Toolkit-BS is the Takuya Ooura General Purpose FFT (Ooura, 2001). All file reading and writing operations are provided in an optional separate package. The 3DTI Toolkit provides an optional library for allowing a simple management of the data resources (HRTFs, BRIRs and ILD models). The library allows loading HRTF and BRIR data from SOFA format (described in Section 2.3.3.2.4).

In order to allow simple access to the various features available in the 3DTI Toolkit-BS for testing and evaluation purposes, a *demonstrator test application* has been created using OpenFrameworks[25]. The application is not open-source, but an installable package for Windows, Mac and Linux is available in the library repository (https://github.com/3DTune-In/3dti_AudioToolkit/releases). In addition, sources and listener positions, as well as audio playback, source level, record options, etc. can also be controlled remotely via Open Sound Control (OSC [91]). This allows the test application to be used as an audio renderer, fully and remotely controlled by other applications, such as VR visual renderers, motion tracking systems, etc.

A snapshot of the user interface can be seen in Figure 45. This interface is divided into two panels: "audio spatialisation configuration panel" on the left, and the "sources and listener layout chart" on the right.

The *audio spatialisation configuration panel* can be seen in more detail in Figure 46. The different controls of the panel are listed below (following the numbering of the Figure 46):

1. Buttons for setting, calibration and OSC. The settings button allows changing many audio configurations, such as frame size, sampling frequency, HRTF resampling step, audio interface, etc. The OSC button provides access to the parameters needed to use the OSC protocol.
2. Listener. This panel allows to load a HRTF file, setup the radius of the listener's head and select the position of the listener.
3. Environment. This panel allows to load a BRIR file, setup the overall gain of the reverb effect and configure the channels of the Ambisonics reverb processing.
4. Source spatialisation. This panel allows to enable or disable all the spatialisation algorithms available to simulate the direct and the reverberation paths, for all the sound sources, separately or all of them at the same time.
5. Source control. This panel allows the user to adjust the position and the volume of the currently selected sound source.

[25] https://openframeworks.cc/ (retrieved January, 2022)

6. Output signal. This panel shows the output signal wave graph and a level meter for each channel (left and right).

The *sources and listener layout chart* allows to set the position of the sources relative to the listener. Sources are represented with spherical coordinates decoupled (one control for the azimuth and a different control for the elevation). Each source includes a volume control, a button to remove it from the chart, a stop, play, pause and mute button and a progress bar for the clip. On the top-left part of the chart a set of buttons are organized on three groups. First group of buttons allows to add a sound source and load and save a specific scenario. The second group is for the audio clip control, which allows to play, pause, stop and record all the sound sources at the same time. The third group of buttons controls the display and the information shown for each source in the chart.

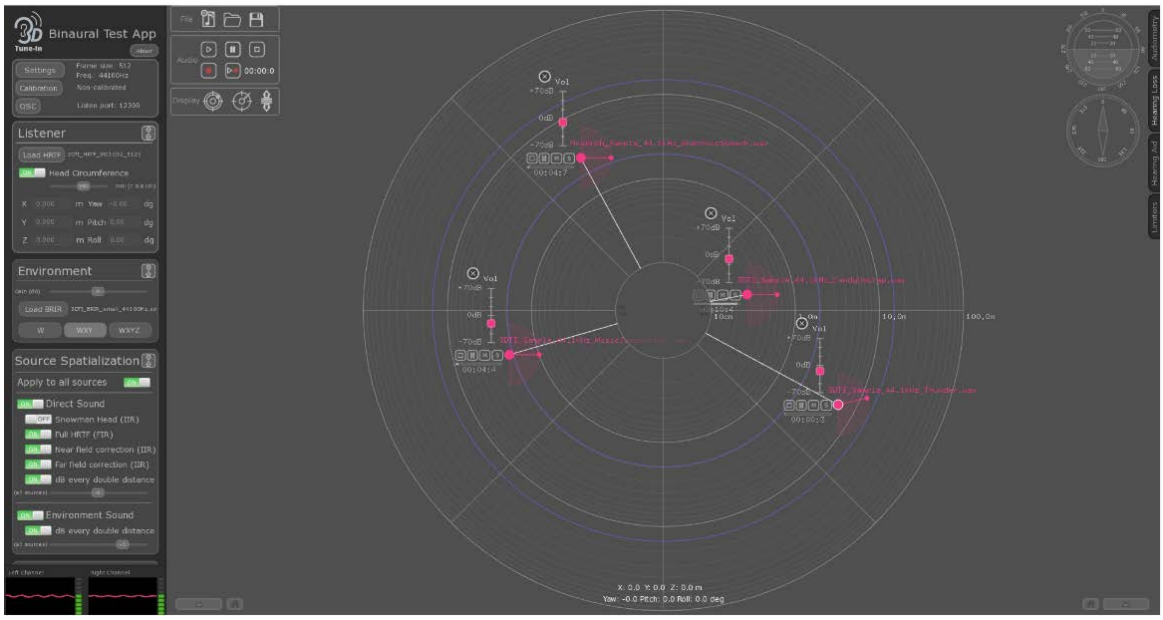


Figure 45. 3DTI Toolkit-BS test application snapshot

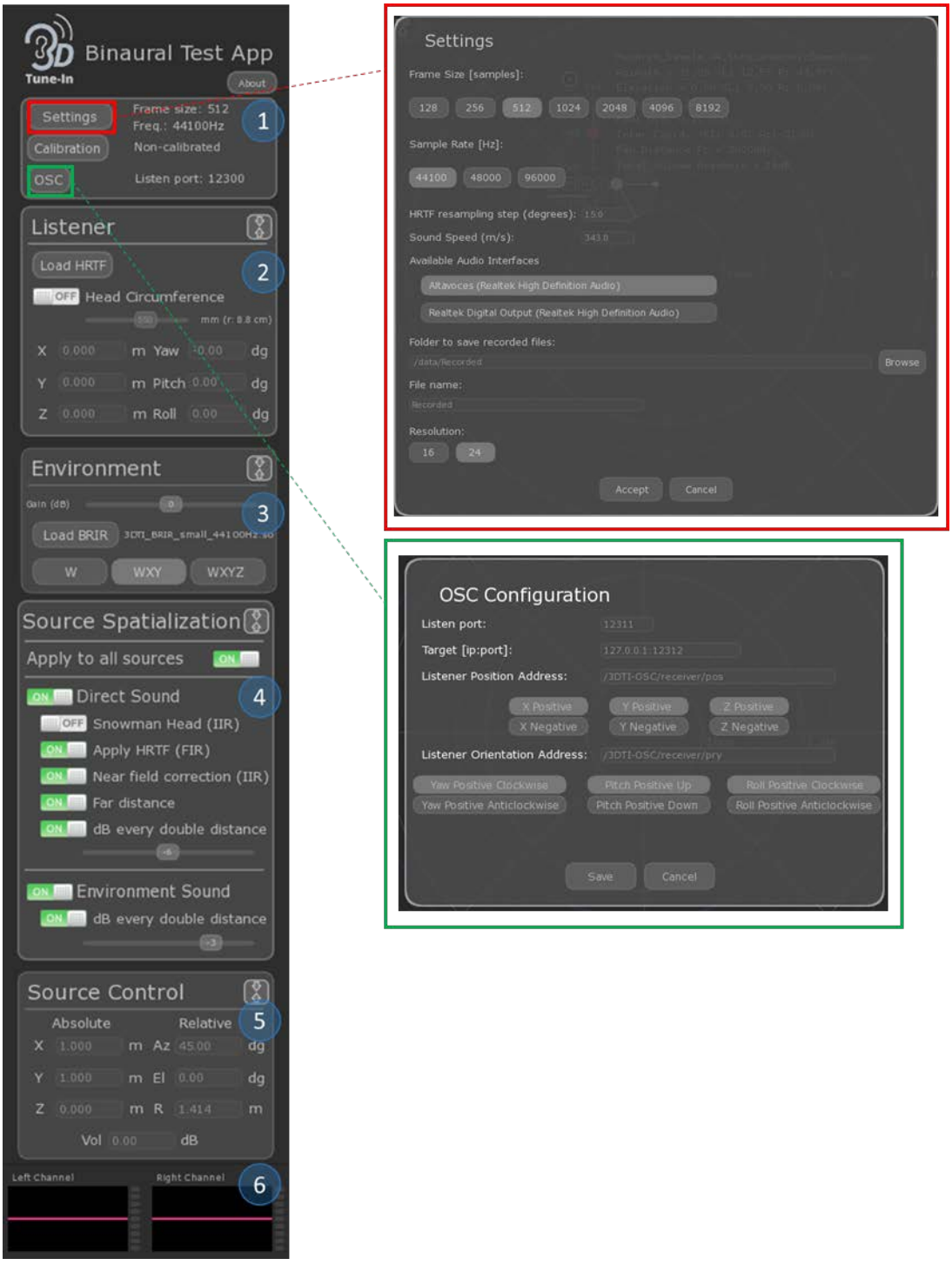


Figure 46. 3DTI Toolkit-BS test application configuration panels

In addition, two complete examples projects implemented in C++ on how to build a simple application which uses the 3DTI Toolkit to spatialise audio is shared under GPLv3 as well. Code and documentation with further details can be found at https://github.com/3DTune-In/3dti_AudioToolkit_Examples.

Other additional tools that make use of the 3DTI Toolkit are listed below. Their development is out of the scope of these PhD Thesis, but it is worth mentioning all the tools currently available that allow the library to be used on different platforms:

- **Unity wrapper.** The 3DTI Toolkit is available as Unity package, allowing integration of some components of the library within a Unity 3D environment. Code and documentation with further details can be found at https://github.com/3DTune-In/3dti_AudioToolkit_UnityWrapper
- **VST plugin.** A plug-in for Virtual Study Technology (VST). All the binaural spatialisation features of the 3DTI Toolkit have been integrated in a VST plug-in for Mac and Windows. Code and documentation with further details can be found at https://github.com/3DTune-In/3dti_AudioToolkit_VST_Plugins
- **3DTI JavaScript Wrapper.** This wrapper allows the integration of the library on web-based platforms, exposing some of the features of the 3DTI Toolkit. Code and documentation with further details can be found at https://github.com/3DTune-In/3dti_AudioToolkit_JavaScript.

3.8 Discussion and comparison with existing tools

As it was presented in Chapter 2, Section 2.4, there are many tools available for binaural spatialisation, both commercial and open-source. In this section, we briefly compare the 3DTI Toolkit-BS with the most representative open-source tools.

General architecture

The architecture of the 3DTI Toolkit presents the direct and reverberation components as independent modules. The direct sound path is rendered separately from the reverb through convolution with discrete HRIRs and BRIR respectively. This architecture allows for insertion of processes only in the direct or reverberation components. An example is the near-field ILDs modification, which is applied only to the direct path signal, and not to the reverberation. Among the tools currently available as open-source, the 3DTI Toolkit is the one allowing most configurability, making it a very appropriate instrument for 3D audio research. Furthermore, it is important to emphasize that, having been implemented according to the C++ 14 standard, the 3DTI Toolkit-BS is highly portable. As an example, the demonstrator test application is currently available for Windows, Mac and Linux (shown in Section 3.7).

As described in previous sections, in the 3DTI Toolkit-BS, a special effort has been put in removing artefacts related to dynamic scenes, where sources and listener are free to move. Several *switching mechanisms* have been implemented to remove these artefacts, which will be evaluated in Chapter 4, 3DTI Toolkit-BS Evaluation. This is a

particularly important condition for interactive VR applications where the sound designer cannot easily predict scene changes in advance.

Direct sound simulation

All tools presented in Chapter 2, Table 2 can be divided into two groups, according with the technique used to simulate the direct sound: HRTF-based and Ambisonics. The 3DTI Toolkit-BS supports *HRTF-based convolution* as the main technique for binaural spatialisation of the direct sound. It supports HRTF files from the standard file formats SOFA (AES69-2015 standard), an owner-format (called 3dti format) or directly by loading an array of floats for the impulse responses. One of the advantages of reading standard SOFA files is the wide range of existing HRTFs databases that can be loaded, either directly (when they provide the SOFA files), or by using the SOFA Matlab/Octave API. Other tools which are able to load any HRTF from this standard format are: *OpenAlSoft* and *SOFAlizer*. *OpenAlSoft* offers a SOFA file reader since the version released in 2019. The *SOFAlizer* is a quite new spatialisation tool developed by the same group that defined the SOFA standard (first release was in 2018). This is a very useful but basic Unity-based spatialisation engine that processes HRTFs in real time with head-tracking, allowing to use and switch on-the-fly between different HRTFs.

3DTI Toolkit-BS allows loading HRTF with any HRIR distribution. To minimize discontinuities and artifacts when HRIR data is not available for a specific direction, a barycentric interpolation algorithm has been implemented, presented in Section 3.5.2 and evaluated in Chapter 4 Section 4.2. In the implemented algorithm, the HRTF table is resampled in an offline process and, if necessary, HRIRs are interpolated in runtime. Many available tools perform HRIR interpolation, however almost none of them (only *Slab 3D*) follows the same approach, dividing the process in an on-line and off-line process, allowing for a better performance in real time. *Slab 3D* uses an external implementation that consists of a biharmonic spline interpolation to create the resampled HRTF table with a uniform grid. However, this tool works with its own HRTF format.

As mentioned in the introduction, *Google Resonance Audio* has become one of the most used renderers nowadays, for both commercial applications and research in VR. This tool uses an Ambisonic approach to simulate the direct path, which allows to render a sound field more efficiently but limiting its spatial resolution. Using convolution for each source separately to reder the direct path, as the 3DTI Toolkit-BS does, offers a higher level of accuracy.

ITD and ILD simulation

The 3DTI Toolkit-BS allows customization of listener head radius for *ITD* processing. In most tools, the delays for ITD are implicit in HRTF and are not handled separately,

except for *Slab 3D* and *OpenALSoft*. *Slab 3D* extracts the ITD to reduce filter length and to simplify real-time interpolation, using a modified version of the Nam/Abel/Smith algorithm described in (Nam et al., 2008), where ITDs are derived from the HRTF group-delay. It also generates ITD data using a spherical head model (Woodworth & Schlosberg, 1954), but they use a fixed head radius of 0.09m. *OpenALSoft* also performs a pre-processing of the HRTF, generating a HRTF dataset in their own format, where the ITD is separated from the HRTF but it is not customizable. The 3DTI Toolkit-BS implements the interpolation of the HRIRs and the ITD separately, to avoid the comb filtering effect described in Section 3.5.2.3 and evaluated in Chapter 4 Section 4.2.

Another feature of the 3DTI Toolkit-BS, again very relevant for VR applications, is the simulation of near-field effects. The user is completely free to move and approach sound sources in the virtual environment, which is rendered by simulating both directional and distance cues to a level of accuracy which cannot be found in other available tools. Other open-source tools offer simulation of near field sources, as *OpenALSoft*, *SoundScape Renderer*, *Resonance* and *Virtual Acoustics*, however, no other tool allows personalization of near field as the 3DTI Toolkit-BS, which allows loading *personalized near-field filter data* for HRTF correction.

Distance simulation

Regarding distance simulation, as the 3DTI Toolkit-BS, most existing tools simulate the effect of attenuation with distance or sound propagation through air, which follows the inverse square law (attenuation of 6dB with every double distance). Only *OpenALSoft* provides customization of the distance attenuation curve. The 3DTI Toolkit-BS follows the inverse square law as well but allows for customization of the attenuation slope (in decibels), which is a solution also found in *SoundScape Renderer*. In addition, the 3DTI Toolkit-BS implements a smoothing mechanism when distance change in real time, described in Section 3.3.1 and evaluated in Section 4.5.

BRIR-based reverb simulation

Environment simulation in the 3DTI Toolkit-BS consists in the convolution of an Ambisonics sound field with BRIRs. This solution provides a high quality and fully spatialized reverb, in contrast with other models where only early reflections are spatialized (e.g *Slab3D* and *Csound Binaural Processing*). *SoundScape Renderer* is the only tool that incorporates BRIR-based convolution but does not allow reading BRIR data from standard SOFA files, as the 3DTI Toolkit-BS. In addition, the fact that 3DTI Toolkit-BS manages the direct and reverberation components as independent modules, allows the selection of the HRTF and BRIR separately, which cannot be done in *SoundScape Renderer*. Furthermore, considering that real-time convolution with BRIRs can become an issue in terms of computational costs, the use of Uniformly Partitioned

Convolution is a unique feature that, to our knowledge, has not yet been implemented in other available open-source tools. Most tools use other rendering methods, such as ray- or image-based methods or network delay lines and digital filters, which are not as accurate but can be very efficient.

Released as open source

3D audio is increasingly being used in VR and gaming applications, and a large amount of research has been conducted in the recent years on this topic. This resulted in several 3D audio rendering tools to be released, with various characteristics and integrating different features. However, not all of them are available as open-source. As Ince et al. (2012) argue in a recent editorial in Nature, the rise of computational science has added a new layer of inaccessibility. This should be overcome by releasing of computer programs as open-source, allowing clarity and reproducibility. Furthermore, it is important to mention that the 3DTI Toolkit is an alive project, which is being continuously improved and assessed. This is obviously facilitated by its open-source nature, which allows for external contributions and bug reporting.

Chapter 4

3DTI Toolkit-BS Evaluation

This chapter is focused in a quantitative and objective evaluation of the implemented 3DTI Toolkit-BS. It starts with and introduction in Section 4.1, explaining the way the tool has been evaluated. Each of the next sections is devoted to the evaluation of a specific feature, including a description of the used technique the obtained results. Finally, Section 4.7 includes the conclusions and perspectives of this chapter.

4.1 Introduction

To assess the new features introduced by the 3DTI Toolkit for binaural spatialised audio rendering, a series of tests have been conducted and an objective evaluation is presented in this chapter. The results of the objective evaluation are organized in different sections, which correspond to the Toolkit feature that is being evaluated. A discussion of each evaluation result is presented in each section. In the first section the proposed technique to interpolate HRIRs separately from the ITD is evaluated, looking at spectral variations and comb filtering. Then, the simulation of source in the near field is shown and discussed. In addition, it is evaluated how Virtual Ambisonics, together with BRIRs, works for reverb simulation. Then, measurements of non-linear distortion are presented to assess how well the 3DTI Toolkit behaves in dynamic situations, testing all implemented "switching mechanism" for a moving source. Finally, real-time performance indicators are presented to report on how many sources can be simultaneously rendered and how large can the simulated environment be (i.e. how long can the BRIR be before real-time performance is affected). All the tests have been done using the test application presented in Section 3.7.

We carried out an objective evaluation, based on repeatable measures, to assess the quality of the rendering and the correctness of the implemented algorithms. However, performing an objective evaluation presents limitations. When simulating spatial audio, we provide a set of cues that produce an illusion, i.e. "an instance of a wrong or misinterpreted perception of a sensory experience" in the listener. That is, we manipulate real stimuli to produce a false perception. Perception is a cognitive process which cannot be evaluated objectively. Subjective evaluation allows to assess this perceptual part, for example by comparing real sounds with virtually spatialised sources, but this is not a trivial task, of which an example can be found in (Adelbert W. Bronkhorst, 1995). Multiple ways for assessing the quality of experience of spatial sound simulation are presented in (Rozenn et al., 2014). Both evaluations, objective and subjective, are useful and complement each other. The subjective evaluation of the 3DTI-BR Toolkit was outside the scope of this thesis.

4.2 Evaluation of the HRIR interpolation

As described in Section 3.5.2, the 3DTI Toolkit-BS obtains the HRIR to be used in the convolution process using a barycentric interpolation of the three nearest aligned HRIRs, which means that the initial delay (ITD if we take into account the difference between both ears) was removed. The latter operation is needed in order to minimize comb filter effects, which are produced when two signals with the same amplitude, but different phases, are superposed. Extracting the initial delay from the HRTF is not a trivial task, and most databases include this delay in the HRTF. In this section we will evaluate the benefits of performing the interpolation without this initial delay and show how the comb filtering effect is indeed reduced.

The interpolation technique implemented by the 3DTI Toolkit-BS is evaluated taking an HRTF from a database and removing one of the already known position. The HRIR for this position is compared with an HRIR in the same position but calculated using the interpolation process. The interpolation process has been carried out with HRIRs that include the initial delay (called non-aligned HRIRs) and with ones that do not include the delay (called aligned HRIRs). An example of non-aligned and aligned HRIRs is shown in Figure 46.

The procedure was as follows. The stimulus of the input signal was a 30-second logarithmic sine sweep, ranging in frequency from 200 Hz to 20 kHz at a sampling frequency of 44100 Hz. The sound source was placed at three different directions: (15°, 0°), (45°, 0°) and (75°, 0°), as it is shown in Figure 47. The HRTF used was the number 1008 (raw version) from the IRCAM LISTEN database. We have compared three

different conditions, for the three different directions, which are distinguished from each other by the HRIR used:

- Condition 1: HRIR from the database.
- Condition 2: HRIR interpolated using non-aligned HRIRs. The HRIR was obtained by interpolating the three nearest HRIRs, which are not aligned since they include the ITD.
- Condition 3: HRIR interpolated using aligned HRIRs. The HRIR was obtained by interpolating the three nearest HRIRs, where the HRIRs have been aligned by removing the ITD. The extracted ITD has been included in the "Delay" field within the SOFA file.

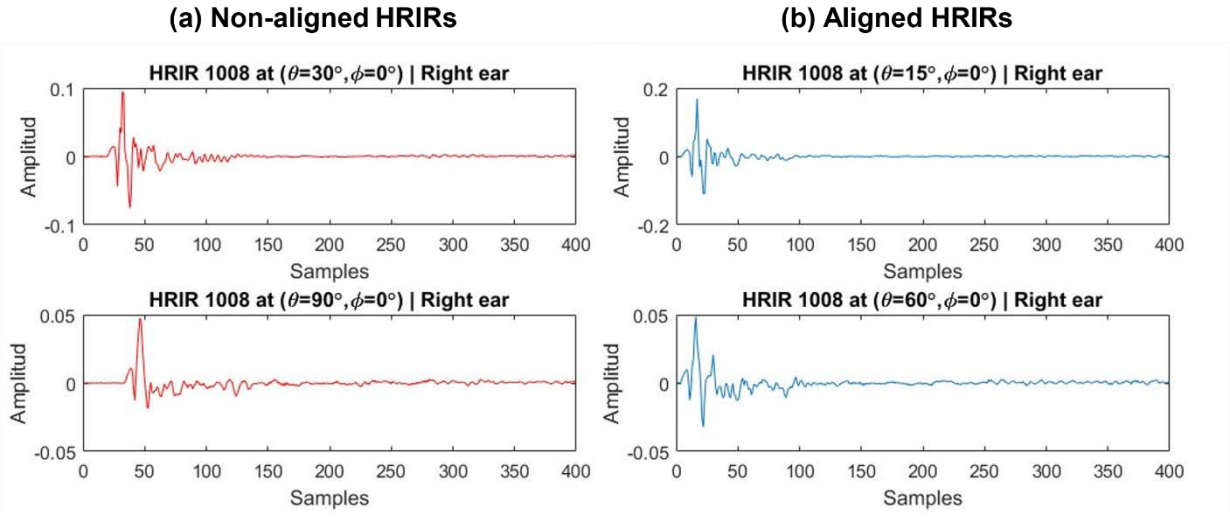


Figure 47. HRIR 1008 in the time domain in two different positions when they are non-aligned (a) and aligned (b).

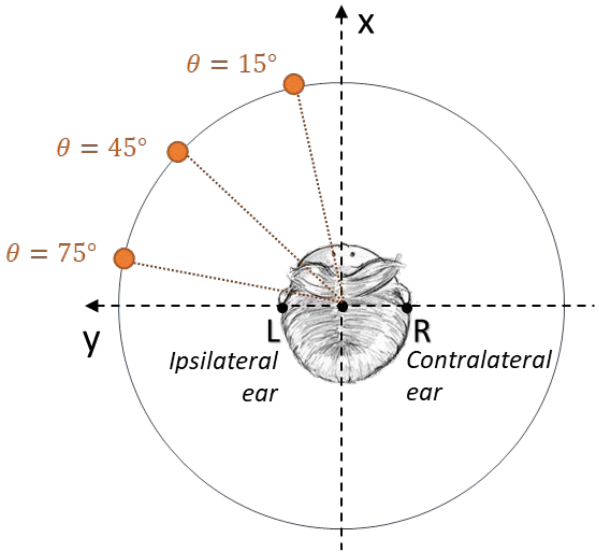


Figure 48. Positions of three directions of the source in the horizontal plane ($\phi = 0$)

The output signal of the system was recorded for the three different conditions. Only the anechoic path processing was activated, disabling reverberation and distance simulation. It should be noted that the HRTF used from the LISTEN database is measured with a step of 15 degrees. Therefore, when performing the interpolation using the three nearest HRIRs, there will always be a minimum difference of 15 degrees between the desired position and the HRIRs used in the interpolation.

Figure 48 shows the results of the evaluation of the HRIR interpolation. Each sub-figure shows the spectrum of the signal for the three different conditions presented above. The y-axis shows the value of the HRIR module and the x-axis the frequency in kHz. Results are shown for left ear (sub-figures (a), (c) and (e)) and right ear (sub-figures (b), (d) and (f)). In addition, three different directions have been evaluated.

Results reveal that, estimating the interpolated HRIR using non-aligned HRIRs (red dashed line – Condition 2) produces some important coloration due to comb filtering effect, causing additional notches to appear in different frequencies. In the ipsilateral ear (left ear) first notches can be found in frequencies between 3 kHz and 8 kHz for directions 15° and 45° but not for the direction of 75°. The absence of a notch at 75° can be due to the fact that at this direction the signal is less affected by the pinnae (which modifies the signal in this frequency range), since it arrives more directly to the auditory canal due to its direction. For high frequencies (larger than 10 kHz), the comb filtering can be observed in all directions. Regarding the interpolated HRIR using aligned HRIRs (green line – Condition 3), the comb effect does not appear and the spectrum is very similar to the original HRIR (blue line – Condition 1). Finally, the phase of the HRTFs in the three conditions and has been found to be not affected by the interpolation process.

In addition, we have computed the Spectral Difference (SD) between the interpolated versions ($Y_{HRIR^*}(f)$) and the original one ($Y_{HRIR}(f)$), in this case for all the available azimuth from 0° to 180° with a step of 15°, as follows:

$$SD(f) = 10 \cdot log_{10} \frac{|Y_{HRIR^*}(f)|^2}{|Y_{HRIR}(f)|^2} \tag{4.1}$$

Results are shown in Figure 49 for HRTF 1008 and 1013 from LISTEN database (raw versions).

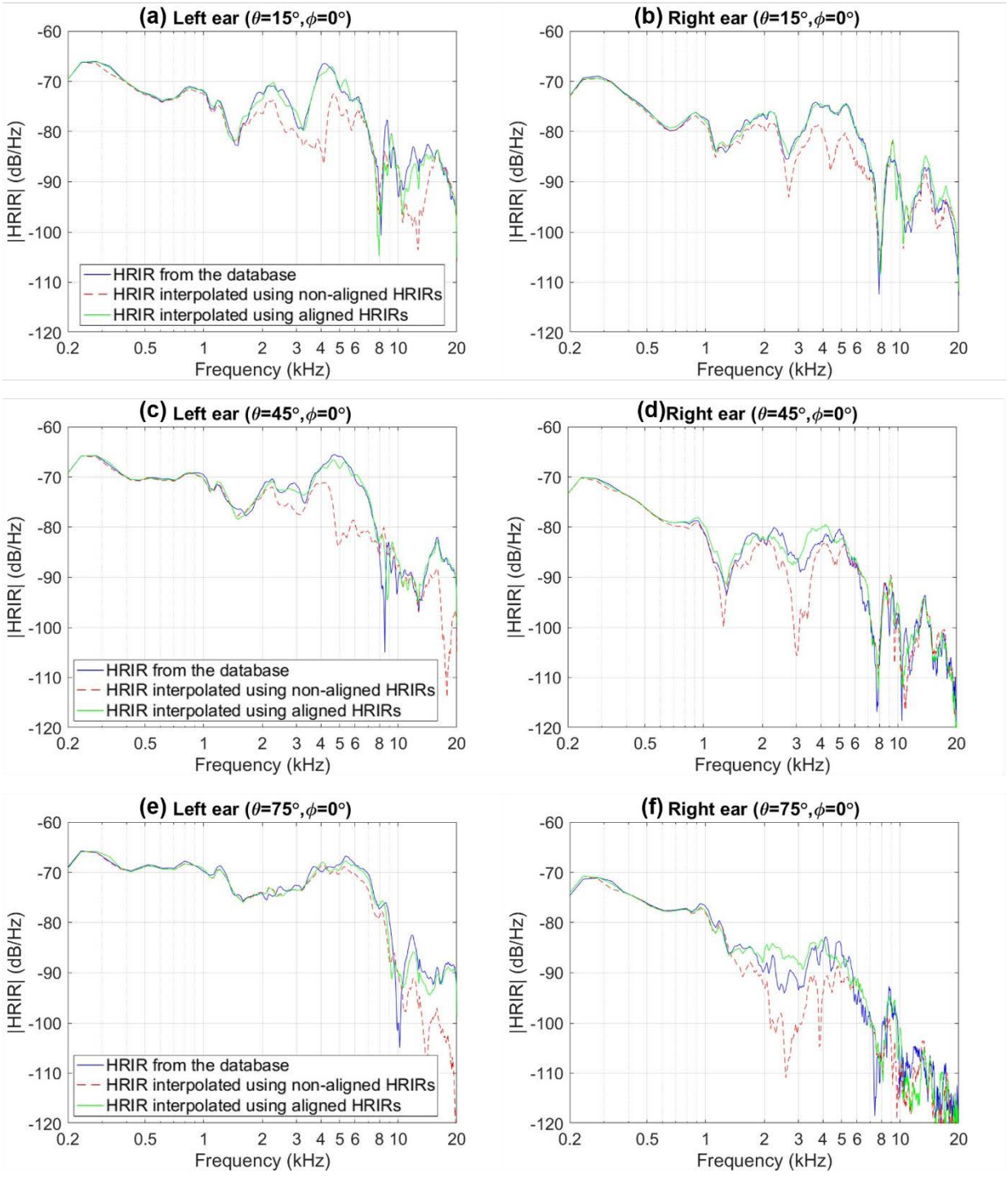


Figure 49. Power spectral density for a sweep signal comparing three conditions; (1) original HRIR from the database (blue line), (2) interpolated HRTF using non-aligned HRIRs (red line), and (3) interpolated HRTF using aligned HRIRs (green line). HRTF 1008 from LISTEN database.

The behaviour is very similar for both HRTFs. The first and third rows show the SD between the original HRIR and the one calculated by interpolating the nearest HRIR where the HRIRs were not aligned. The reddish colours show the highest values of SD,

which means that at these positions and frequencies the comb filter effect is more pronounced. We can observe how, in the ipsilateral ear (left ear), the first comb filter effects appear from frequencies above 3kHz and for azimuths below 45° and above 135°. In the contralateral ear, comb filter effects appear earlier in frequency, around 2 kHz but in both cases, it takes its highest value at frequencies of 3 kHz. In this case the largest SD values can be seen between 45° and 120°, which is when the head more affects the signal in the contralateral ear.

The highest values of SD can be found for frequencies between 3 kHz and 8 kHz and from 10 kHz to 20 kHz, for many directions. The notches that help in elevation perception are located in the frequencies around 5-6 kHz, which may suggest that an interpolation using non-aligned HRIRs may alter elevation perception. Moreover, we can observe how, in the ipsilateral ear, by approaching the direction of 90 degrees, the differences decrease, reaching their minimum at 90 degrees. This is due to the fact that in this case the sound arrives in a more direct way to the auditory canal, without a big modification due to the body and pinna of the listener. In this way, we understand that in this area, HRIRs for nearest directions have a smaller delay difference and therefore it can cause less comb filtering effect.

Regarding interpolation using aligned HRIRs (sub-figures of rows two and four of Figure 49), the SD are much lower for all azimuths and frequency ranges than the interpolation using HRIRs with ITD. The SD between the original HRIR and the one interpolated without ITD, but it should be considered that the comparison is being carried out between a HRIR measured in a given location, and another HRIR estimated through interpolation between HRIRs measured from three adjacent locations. A certain level of spectral discrepancies is therefore to be expected. In addition, we should mention that the SD is greater for the contralateral ear since it must be taken into account that these are the HRIRs which produce the lower output signal, because of ILD.

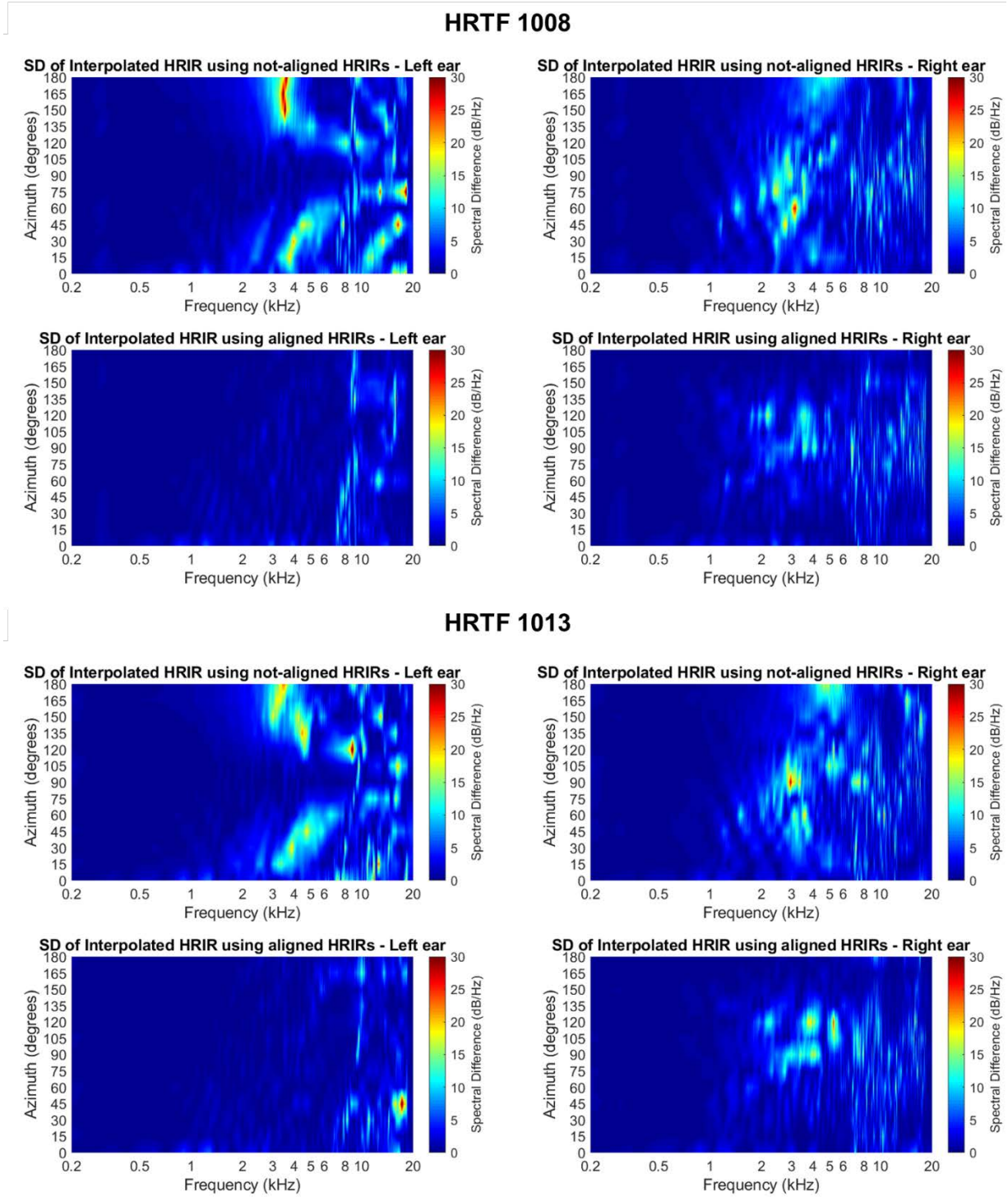


Figure 50. Spectral Differences (SD) of HRTF 1008 and HRTF 1013 for multiple directions in the horizontal plane. Vertical axis shows the azimuth value and horizontal axis shows the frequency in a logarithmic scale. First row uses the interpolated method using non-aligned HRIRs and second row the interpolated method of the HRIRs using aligned HRIRs. The SD is calculated with respect to the original HRIR. First column shows the left ear and second column the right ear.

4.3 Evaluation of the near field simulation

To simulate sources located in the listener's near field, the 3DTI Toolkit-BS implements a correction of the HRTF, simulating the ILD that occurs at these distances, presented in Section 3.5.4. A set of filters have been implemented based on a Spherical Head Model (SHM) of Duda & Martens (1998). These filters simulate the ILD increments when the source approaches to the listener, for the whole range of frequencies. The ILD for the SHM is presented in Figure 51, where ρ, which is calculated as $\frac{d}{a}$, is the distance from the source to the centre of the listener head (d), normalized with the radius of the listener head (a). The horizontal axis is the normalized frequency using the radius of the listener head (a) and the sound speed (c), which means that 1 is when the wavelength is equal to the head radius.

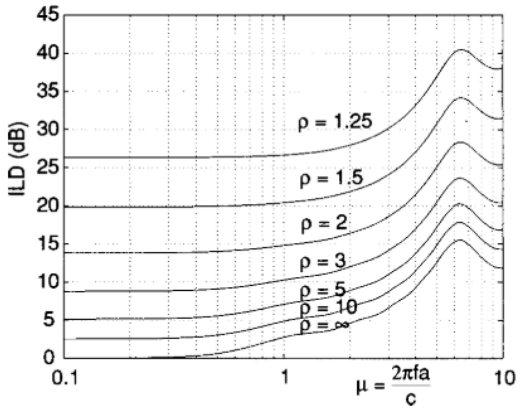


Figure 51. Image from (Duda & Martens, 1998). ILD for a sound source at (100°, 0°) using the SHM as HRTF.

The performance of the 3DTI Toolkit-BS without and with near field sources simulation is presented in Figure 51, for the HRTF 1008 and HRTF 1013 from the LISTEN database. These graphs show the output of the Toolkit for a sound source placed at 100 degrees of azimuth and 0 degrees of elevation and different distances indicated with ρ. The sound stimulus is a 30-second logarithmic sine sweep, ranging in frequency from 200 Hz to 20 kHz at a sampling frequency of 48000 Hz.

Figure 51 shows how both HRTFs present the same performance. When the near field are enabled (sub-figures of the second column) we can observe the effect reported by Duda with the SHM (Figure 50), where the ILD increases for all frequencies as the source approaches the listener (ρ close to 1). In addition, it can be seen the same shape in the curves for μ values between 6 and 8, which correspond to 3.7 kHz and 5 kHz. This effect also appears in the graph where the near field is disabled (graphs on the first column), since it is an effect caused by the filtering of the head characterized in the HRTF.

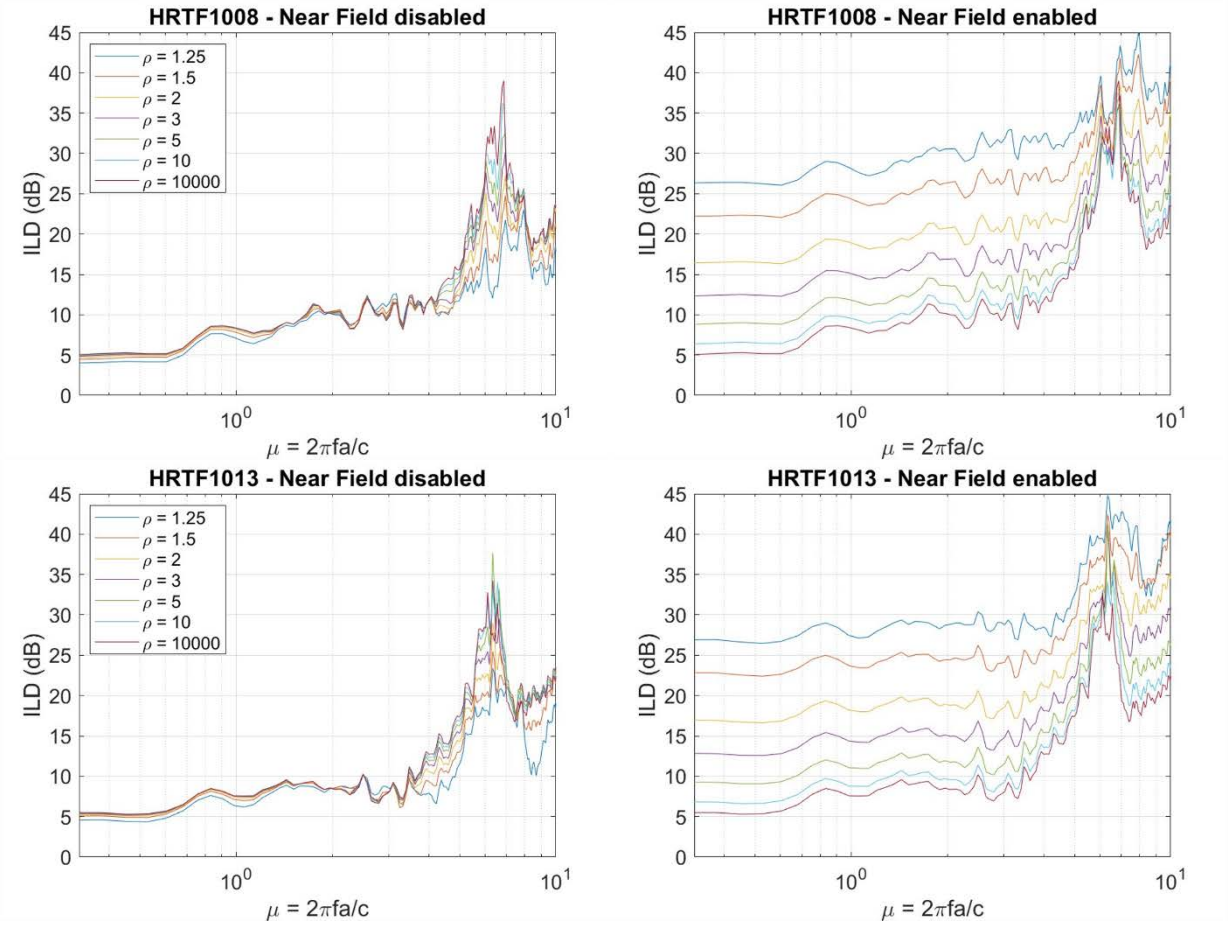


Figure 52. IDL (dB) for HRIR 1008 (first row) and HRIR 1013 (second row) with the sound source placed at (100°, 0°) and the near field simulation disable (left graph) and enabled (right graph).

In order to see the near field effect in a "clearer" way, we eliminate the effect of the pinnae (which is also not included in SHM model of Duda), using a synthetic HRTF, which consists of a head model with no ears and only containing the two binaural cues, ITD and ILD. The ITD was modelled as a time delay function, using the Woodworhts's formula (Woodworth et al., 1954). This formula defines the difference in the arrival time of a wave sound as $r\,(\theta + sin\,\theta)/c$, where r is head radius and was set to 8,75 cm, θ is the azimuth of the source, and c is the speed of sound. The ILD was built as a simple one-pole one-zero model, based on the analytical model obtained by L. Rayleigh & Lodge (1904).

Results are shown in Figure 52. It shows again the effect of the near field simulation, which makes, in the same way as previously described, the ILD increments for all frequencies as the source is closer to the listener. This synthetic HRTF does not have the same shape than the measured HRTFs and SHM graphs for μ values between 6 and

8, which suggests that this is an effect of the spherical head shape presented by the SHM and the measured HRTFs, and therefore does not appear in the very simplified version of our synthetic head since the one-pole-one-zero model is too simple and does not capture that high-frequency part.

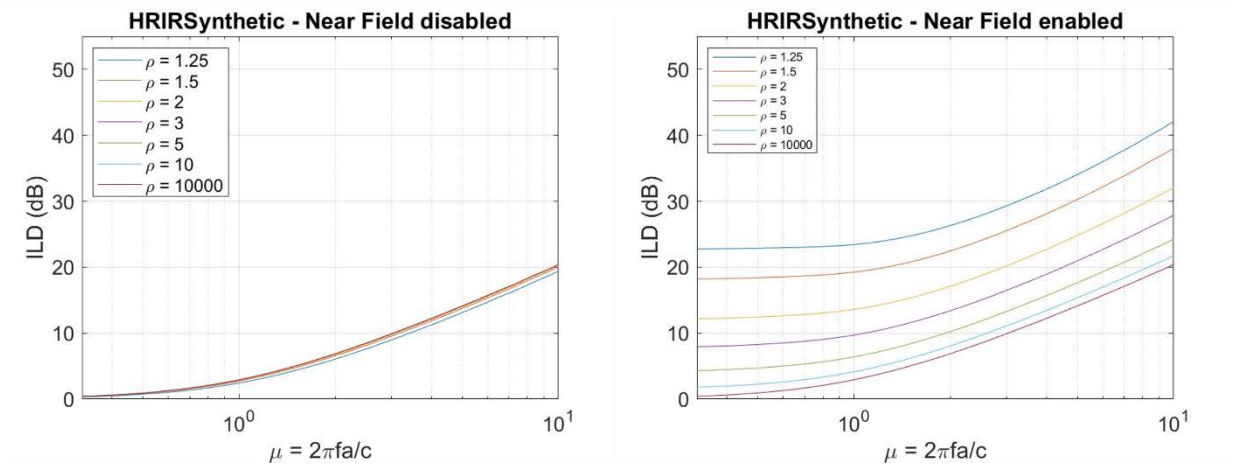


Figure 53. ILD (dB) for synthetic HRTF and the sound source placed at (100°, 0°). With the near field simulation disabled (left graph) and enabled (right graph).

In addition, when the source is placed at (0°, 0°) the ILD is 0 dB for all frequencies and distances, for the synthetic-HRTF, since it is totally symmetric (Figure 53 left sub-figure). For the measured HRTFs (1008 and 1013) we can see how the ILD values do not increase for the different distances, as it happened for other positions of the source seen in previous figures. This is because, being the ILD the difference of level between the two ears, the value must be zero in the centre, regardless of the distance of the source. However, we can see that the curves are not zero, which may be simply due to an asymmetry in the performance of HRTF measurements or an asymmetry in the listener's head.

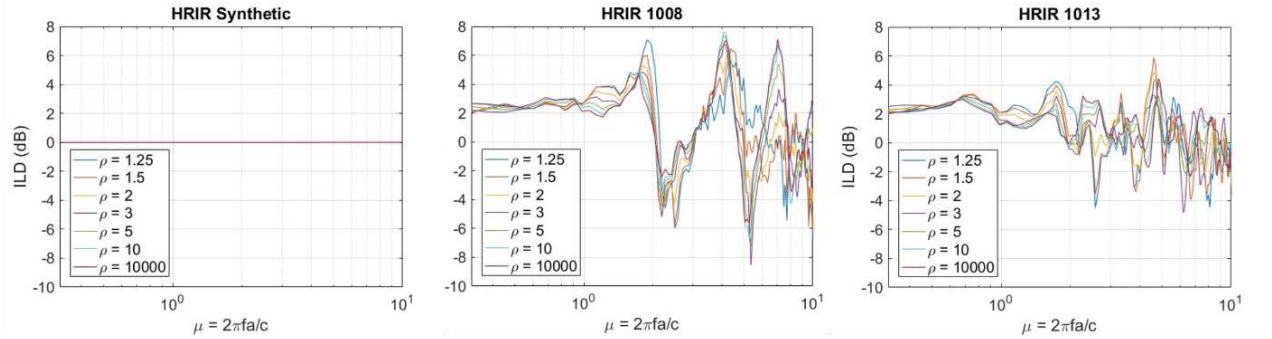


Figure 54. ILD for HRTF-1008 and 1013 from LISTEN Database with the sound source placed at (0°, 0°).

4.4 Evaluation of the BRIR simulation

The 3DTI Toolkit-BS simulates the reverberation of a set of sources within an enclosed space by implementing a virtual Ambisonics approach (Section 3.6). In this approach, all sources are encoded together into a first-order Ambisonic format and then decoded into a set of virtual speakers, always placed in the same location in the virtual scene. The Ambisonic decoded signals of each virtual speaker is convolved with the BRIR that corresponds to the speaker direction. In this way, we only need to know the BRIRs in the positions of the virtual speakers, which in our case are at the front, back, right, left and above the listener. To simulate the reverberation of a source at a specific position of a virtual loudspeaker, the input signal will be encoded in a way that it becomes an output of several virtual speakers, even if the source is placed in the specific position of a virtual speaker, since the Ambisonic does not realistically decode a source that is at the position of one speaker. The Ambisonic approximation allows to preserve certain characteristics of the position of the source, but it is not as accurate as convolution with the BRIR of the exact direction of the source. This approach allows us to simulate the reverberation of a source at any position within a room (at a fixed distance), having the BRIR of the room only in few positions.

This section evaluates how a source is spatialized at different directions, which can coincide or not with the location of the virtual speakers. To do so, a set of BRIRs have been measured in some directions in the horizontal plane, with azimuths between 0° and 90°, in 10-degrees step. The BRIR has been measured in a 10m x 15m x 5m church with an estimated T30 of approximately 1s[26].

These measured BRIR have been compared to the impulse response of the Toolkit (called here Toolkit-BRIR), which has gone through an Ambisonic process. To get the Toolkit-BRIRs, the input of the Toolkit was a delta signal located at the same positions as the measured BRIRs, shown in Figure 54a with blue circles. To perform the virtual Ambisonic approximation the Toolkit needs to know a set of BRIRs (as explained in the Section 3.6). For the sake of comparison, the BRIRs have been obtained from the set of synthetic BRIRs, for the positions shown in Figure 55a with grey squares. For each pair of Toolkit-BRIR and measured BRIR, a cross correlation has been carried out and the results are shown in Figure 55b.

The maximum correlation is reached when we compare the measured BRIR with the Toolkit BRIR at the virtual speaker positions (0 and 90 degrees). As the azimuth value is moved away from these positions, the correlation coefficient decreases, taking its minimum value between 40 and 50 degrees, the positions farthest from the virtual

[26] Thanks to Lorenzo Picinali for the measurements.

speakers. Even so, these values are around 0.8. This suggests that, for this type of room, there is a good correlation between the measured BRIR and the one we get from the Toolkit. The use of an Ambisonic approach makes that the BRIR of the Toolkit in the positions of the virtual speakers and the measured BRIR at this position are not exactly the same (the correlation is lower than 1). This is because in the Ambisonics approach the signal is spread over the different virtual loudspeakers, even if the position of a source coincides exactly with that of a loudspeaker.

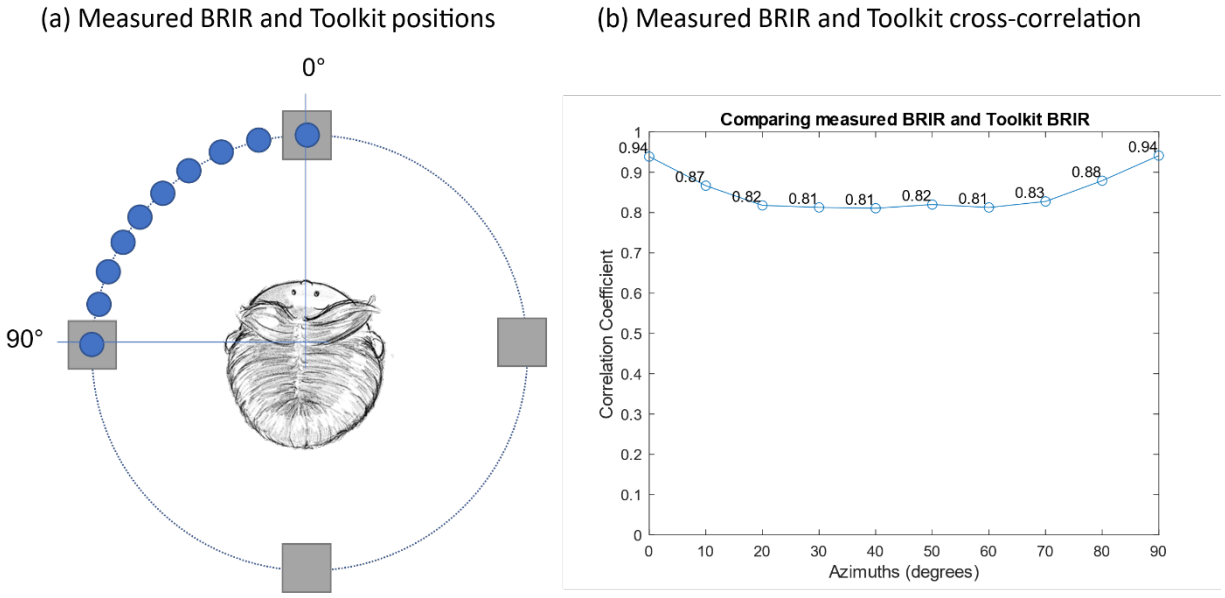


Figure 55. The diagram of the left (a) shows the direction of the measured with blue circles and the position of the virtual speakers with grey squares. Note that there are also speakers above and below the listener, but these are not shown in the diagram. The diagram on the right (b) shows the cross correlation between a measured BRIR and the one simulated by the 3DTI Toolkit-BS for the left ear.

The 3DTI Toolkit-BS allows to select different reverb configurations, according with the number of virtual speakers used. Usually, BRIRs are only measured at the horizontal plane (Section 3.6), for which the Toolkit offers an option called '2D'. When this configuration is selected, and the source is outside the horizontal plane, to avoid the power loss due to the absence of virtual speakers above and below the listener, the elevation is encoded in the Ambisonic channel X. For the same reason, the W channel in the de-codification process is amplified. Figure 55 compares the output of the Toolkit when the same source, placed at different elevations, is spatialized using both orders (3D and 2D). The graph shows the cross-correlation between a delta input signal with 2D reverb spatialisation and the same signal with 3D reverb spatialisation, both placed at different elevations in the median plane.

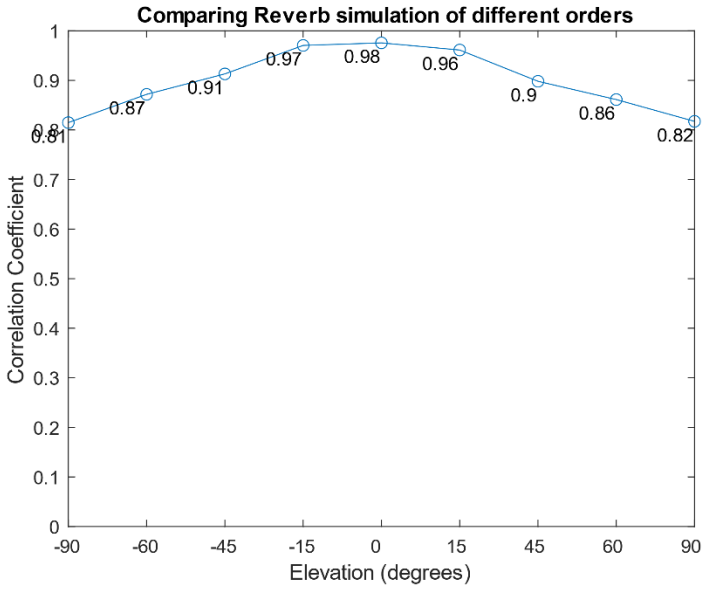


Figure 56. Cross correlation between a 3D reverb simulation and 2D by the 3DTI Toolkit-BS for the left ear for a source located at different elevations in the median plane.

The maximum correlation coefficient is obtained at 0° of elevation when the source is in the horizontal plane and the 2D approximation is equivalent to the 3D. In the same way as before, the correlation is not 1 due to the Ambisonic approximation and the absence of the W channel in the up and down speakers of the 2D case, although thanks to the power compensation we made the correlation achieved quite high. The further we move away from the horizontal plane, both for the upper and lower semi-hemisphere, the correlation coefficient begins to drop, being its lowest value 0.81, at -90°. Given this correlation coefficient we consider that the 2D approach is a good alternative for when BRIR measured at different elevations is not available.

4.5 Reduction of non-linear artefacts

The 3DTI Toolkit-BS supports real-time 3D audio spatialisation for dynamic scenarios (moving sources and listener). Whenever a source is moving or the listener moves relative to a source, filter coefficients, gains and delays can change in consecutive audio frames, resulting in discontinuities which may produce audible artefacts. The smoothing mechanisms to avoid audible artefacts in the resulting audio signal have been described in Chapter 3. These mechanisms are summarized in the diagram of the next Figure. This diagram shows the different components and algorithms implemented to simulate the anechoic path (presented in Section 3.5) and which smoothing mechanism has been implemented within each component.

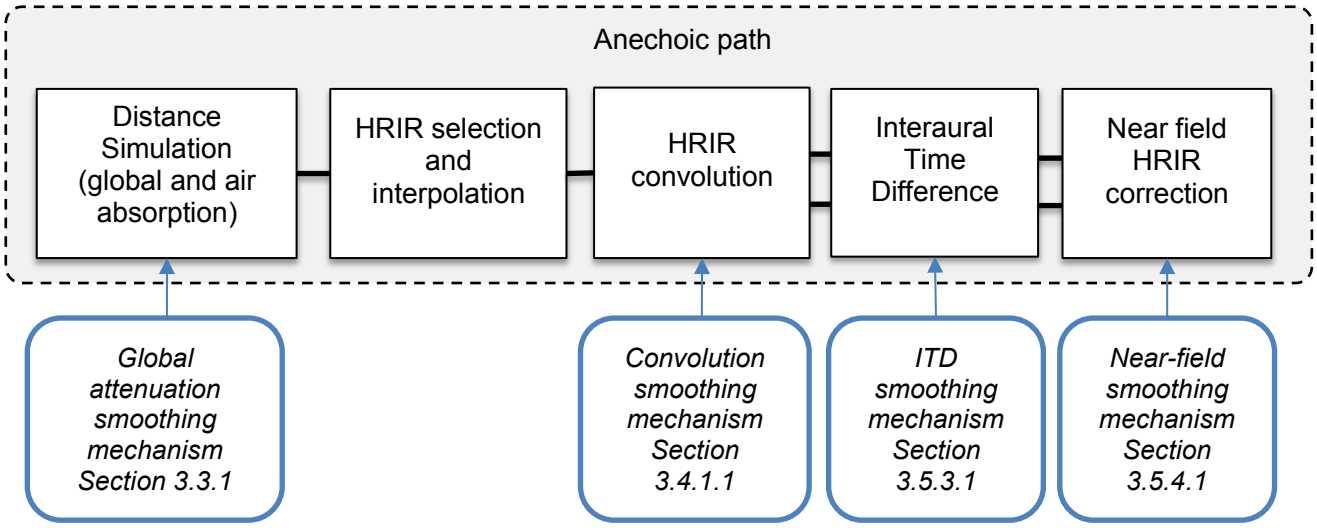


Figure 57. Components of the pipeline of the anechoic path, highlighting the different smoothing mechanisms implemented in each component.

This section presents an evaluation of the 3DTI Toolkit-BS dynamic behaviour, assessing the performance of the above-mentioned smoothing mechanisms. The analysis is based on measuring the non-linear distortion produced by the 3DTI Toolkit-BS when a source is moving at different speeds. For this purpose, the system has been tested with a signal composed by three representative tones (859.65 Hz, 4298 Hz and 8596 Hz), estimating the percentage of the Energy out of Band (EoB), following the approach described in (Belloch et al., 2013). They selected this kind of signal because «it has three equally spaced tones that are sufficiently separated in the audible spectrum». The frame size used in the test was 512 samples, and the sample rate 44.1 kHz. We used the HRTF number 1013 from the LISTEN database (raw version) and the simulation was only anechoic. The source was moved on the horizontal plane in a circle trajectory around the listener at different distances and with different speeds for 360 frames (4.18 s). Hence, we computed the FFT of an output signal of N = 184320 samples (360x512). The result was $Y[i]$, composed of N samples as well. Then we defined the energy of frequency f, as the energy contained in M= 361 samples around f, which is the frequency of the tones:

$$E(f) = \sum_{i=i_f-(M-1)/2}^{i=i_f+(M-1)/2} |Y[i]|^2 \qquad (4.2)$$

being i_f the index of the sample that correspond to frequency f, that's to say, $f = i_f \cdot f_s/(2 \cdot N)$, where f_s is the sampling frequency. We also define the total energy through

$$E_{tot} = \sum_{i=0}^{i=N-1} |Y[i]|^2 \qquad (4.3)$$

Then, for the case of our three tones, the EoB, in percentage, is computed as:

$$EoB(\%) = 100 \cdot \frac{E_{tot} - E(859.65) - E(4298) - E(8596)}{E_{tot}} \tag{4.4}$$

EoB was calculated for each combination of distance and speed and switching mechanism implemented, giving the results in Figure 57.

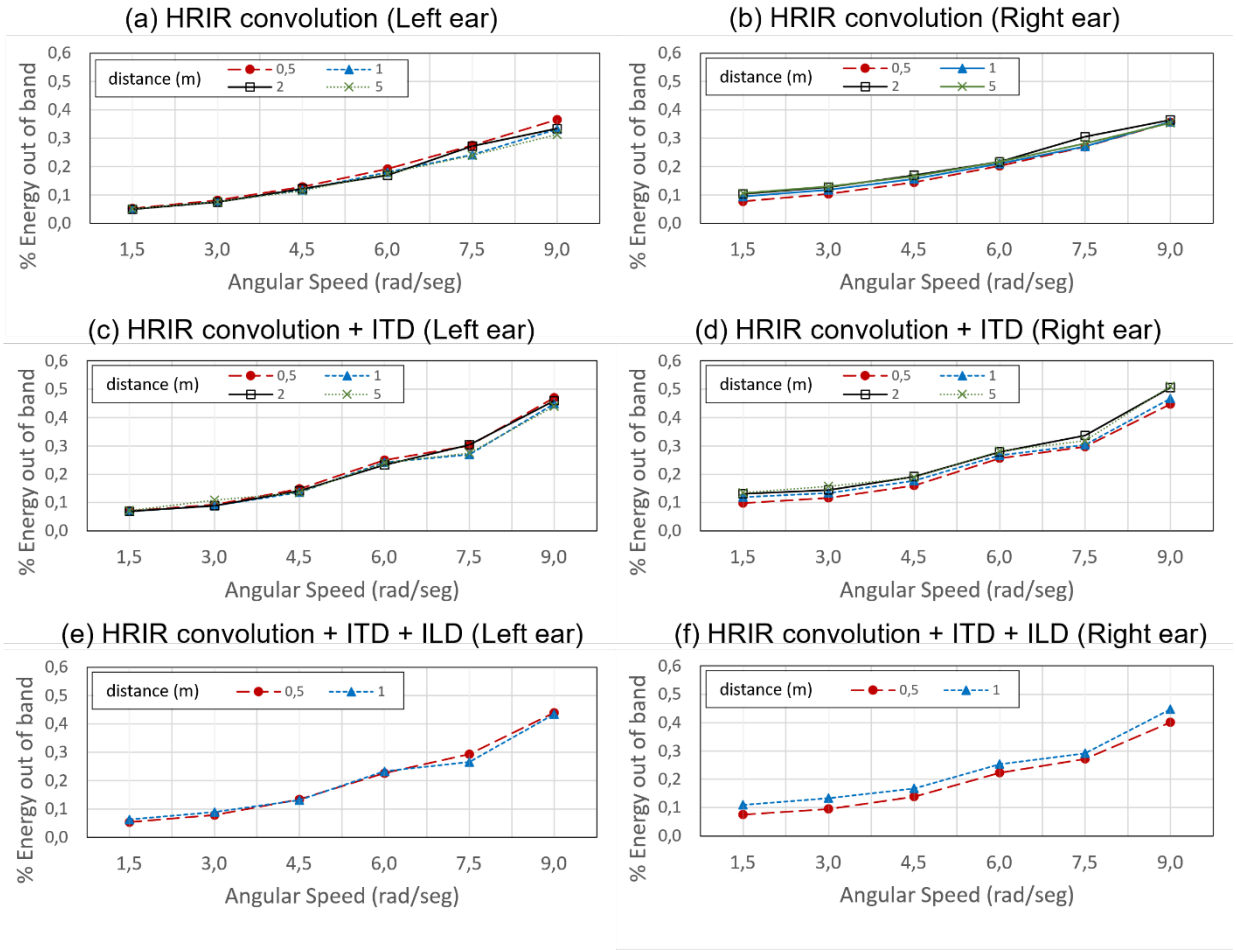


Figure 58. Energy out of band produced by the spatialisation process for different configurations. (a) and (b): only convolution of aligned HRIRs, (c) and (d): convolution with aligned HRIRs and computed ITD, (e) and (f) convolution with aligned HRIRs and computed ITD and ILDs. Every configuration for both ears.

First, we evaluate the *convolution smoothing mechanism*, disabling all other components of the pipeline. Figure 57a and Figure 57b shows the EoB, for the left and right ear respectively. It can be noted that increasing speed results in an increased EoB, as expected, but even at high speed (9 rad/s) the overall distortion is relatively small.

Then we activate the component where the ITD is added after the convolution process, so the *ITD smoothing mechanism* is enabled. The estimated distortion shown in Figure 57c and Figure 57d (for left and right ear respectively) reveals only a very slight increase if compared with the previous conditions, despite the fact that a delay of up to 30 samples is applied, which dynamically decreases down to 0 and up again on the other ear for every full lap. Finally, we add the component where the ILD is simulated, enabling in this case the *near-field smoothing mechanism.* Results are shown in Figure 57e and Figure 57f for the left and the right ear. Only distances under 2 meters are plotted here, as the near-field compensation is activated only in that distance range. In this case, the EoB does not increase when adding the near-field correction; on the contrary, a minor decrease of the overall distortion is noted. This is probably due to the fact that non-linear distortion is higher in the contralateral ear, where near-field correction filters apply higher attenuation, as can be seen in Figure 58. In any case, distortion introduced by the dynamic behaviour of these filters can be considered negligible in the light of Figure 57.

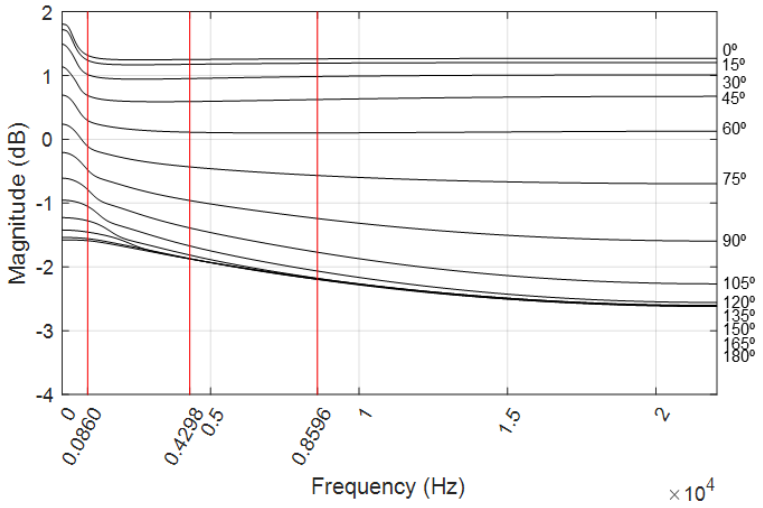


Figure 59. Difference filters implemented for Near-Field HRIR correction. Vertical lines indicate the frequency of the tones used for the evaluation. The vertical right axis indicates the interaural azimuth of each filter.

Finally, we evaluate the *global attenuation smoothing mechanism,* using a different trajectory than in the previous cases, since in the above evaluation the distance is fixed for each movement (circular trajectory around the listener). In order to study the EoB when distance attenuation is applied, we repeated the same procedure but moving the source in a line trajectory, away from the listener with a distance from 0.1 to 40 meters. In this case only the Global attenuation component is enabled. Results can be seen in Figure 59.

Results are in the same line as previously shown in Figure 57. As expected, when the lineal speed increases, the EoB increases, but even at the highest speed (9 m/s) the energy out of band is relatively small, so we can think that even for high speeds no big audible artifacts will be produced.

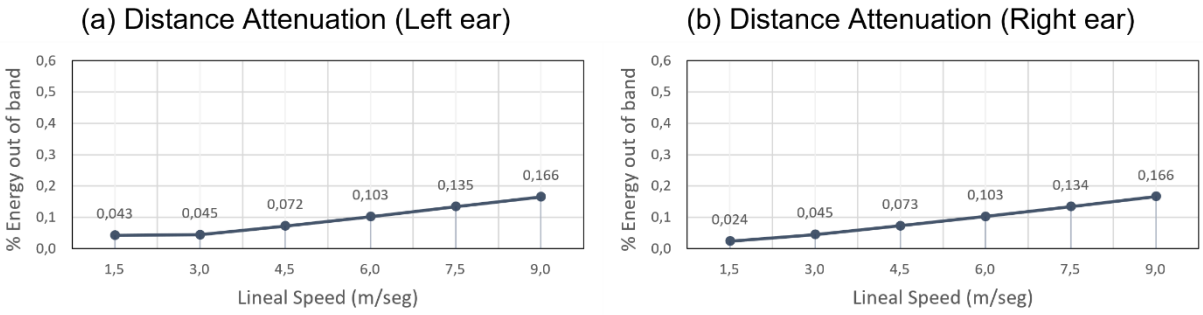


Figure 60. Energy out of band produced by the spatialisation process for only global distance attenuation for the anechoic path enabled, for different lineal velocities.

4.6 Real-time performance

The 3DTI Toolkit-BS is designed to work on Windows, MacOS and Linux, which are not real time operating systems. This means that its process has to share CPU time with other processes. For this reason, the time that the Toolkit takes to process one frame is the actual time devoted to produce the spatialised audio plus the time taken by the interruptions of the current process. If this total time exceeds the frame time, the whole audio frame will be dropped out, producing an audible artefact.

All tests presented in this section were performed on a desktop computer with Intel i7-6700 microprocessor, 3.40 GHz and RAM of 16 GB, working with Windows 10, 64 bits. A test application using the Toolkit was the only user process running in the computer besides the operating system processes. Time measurements were performed using the profiling tools included in the Toolkit, so they were integrated in the test application.

As described before, the 3DTI Toolkit-BS processes the anechoic path and the reverberation separately. These are independent process chains and follow different approaches (see Figure 23 of Chapter 3). While the anechoic path is rendered per source, reverberation is generated over an Ambisonic encoding of all sources, which are therefore merged together at the beginning of the process. For this reason, real-time performance is evaluated separately for both paths.

Figure 60 shows a set of box plots of the percentage of available time in a frame taken by the Toolkit process every frame for different number of sources and various frame sizes. This rendering was made using an HRIR of 512 samples and a sampling frequency of 44100 Hz. For each condition, 1000 frames were recorded. The boxes represent the inter-quartile range between Q1 and Q3, the median is represented with a central line, and the mean by a cross. As expected, this percentage, or duty cycle, mostly concentrates around their typical values, which are assumed to represent an estimation of the net processing time. It linearly increases with the number of sources, allowing to render a relatively large number of sources in an ordinary desktop computer. Moreover, the frame size has a significant influence on performance; the lower the frame size, the larger the duty cycle. There are also some outliers which we can attribute to the presence of those interruptions commented before. In these tests, no frame took longer than the available time. Hence there was no dropout.

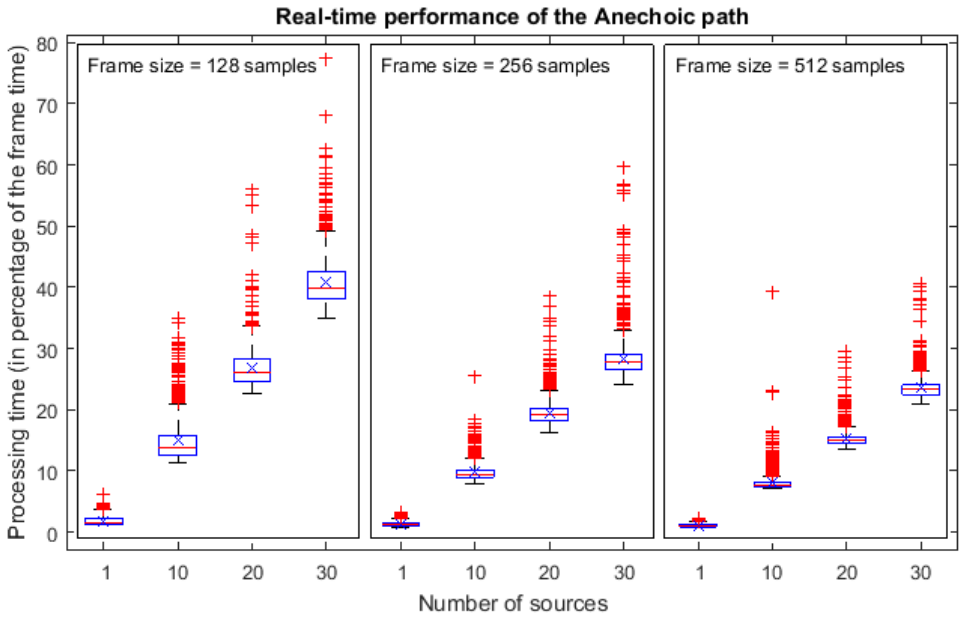


Figure 61. Performance of the anechoic process depending on the frame size. The horizontal axis shows number of sources and the vertical shows percentage of total frame size.

Similarly, the graphs presented in Figure 61 show the duty cycle for the reverberation process. This process is almost independent of the number of sources involved. In this test, 10 sources were used, and the duty cycle was measured for different BRIR lengths and various frame sizes. Again, 1000 frames were recorded for each condition. It can be noted that, as the frame size increases, the duty cycle decreases. For very large BRIRs and small frame sizes, it was not possible to perform the processes in real-time, as it required an amount of resources that were not available; these data points are therefore not reported in the chart. Moreover, for some conditions shown in Figure 61, as that for a frame size of 128 samples using the medium room, even having enough time to perform

the process (around 75% of duty cycle), there are some frames which took more than 100% of the available time in the frame, producing dropouts and thus yielding to a low quality audio experience.

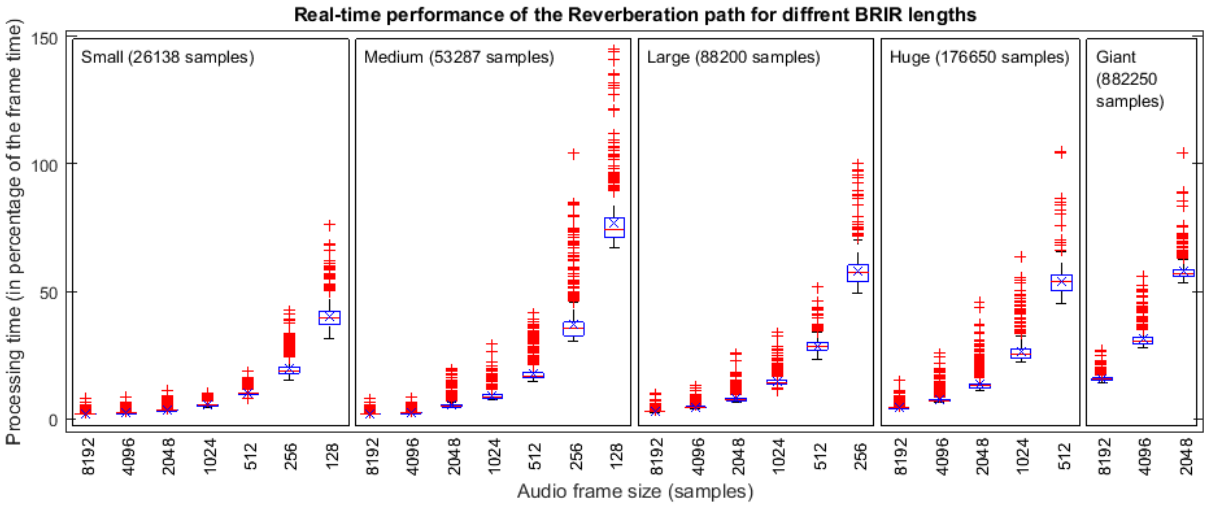


Figure 62. Performance of the reverberation process depending on BRIR length. The horizontal axis shows different frame sizes, and the vertical axis shows percent of total frame time.

A careful selection of the frame size provides the possibility of rendering with low latency very long reverberations. There is a trade-off between latency and supported computational cost. Taking a sample rate of 44100 Hz as a reference, a frame size of 128 samples implies a latency of 2.9 ms. With the computer used for this test, the 3DTI Toolkit-BS is able to render up to tens of sources with small and medium reverberation lengths (up to around 1s). On the other side, very long reverberation (we tested a very large room with a BRIR of 20 sec, i.e. 882250 samples) can be rendered at the cost of sacrificing latency, where with a frame size of 2048 the latency is 46 ms.

Both anechoic and reverberation processes are independent of each other, and the total computational cost can be estimated by simply summing the values for each of the two processes. As an example, referring to Figure 60 and Figure 61, and considering a scenario with 20 sources in the large room, with a frame size of 512 samples, the anechoic path takes 15% of the CPU time on our test computer, and the reverberation path takes 29% of the CPU time. Consequently, this scenario takes 44% of the total CPU time, producing no dropouts.

The low latency requirements can be reached thanks to the use of the uniformly partitioned convolution (presented in Section 3.4.1). These algorithms allow the use of small audio frame sizes by splitting the HRTFs and BRIRs filters into blocks with the same size as the audio frames. Wefers (2015) demonstrated that a lower latency is achievable with these techniques at a relatively low cost. Even if the latency is irrelevant

(for no real-time performances), the uniformly partitioned convolution is still advantageous due the improvement that offers in memory locality and cache utilization.

4.7 Conclusions

This chapter describes an objective evaluation of the different algorithms implemented in the 3DTI Toolkit-BS, especially those that are new in comparison to other existing tools. It starts evaluating the HRIR interpolation technique, showing that the extraction of the ITD prior to interpolation is crucial to avoid the comb filtering effect. This effect makes the timbre of the signal to be coloured, and when the stimulus is a speech, it can make the sound more "robotic". When comparing the interpolated HRIR (using aligned HRIRs) and the original HRIR we got very low spectral differences, which suggest that both signals are very similar. This will offer a more accurate spatialisation and less audible artefacts for a moving source.

The technique implemented for simulating sources in the near field has been also evaluated. In this case, we have extracted the response of the Toolkit for a source located in a distance very close to the listener, with different azimuths. We have compared these signals with the model that was followed during the implementation (Duda & Martens, 1998), observing how the Toolkit responses acts in accordance with the model and provide all the effects produced in the near field that we know from the state of the art. The novelty offered by the Toolkit is the fact that the ILD model used to simulate the near field can be individualized for each user, depending on their HRTF, thanks to the use of a look-at table that contains the coefficients of the implemented filters.

The 3DTI Toolkit-BS performs the reverb simulation of a sound source located at any place inside the environment with a BRIR-based convolution technique, thanks to a virtual Ambisonics approximation. This chapter has evaluated the BRIR used for many different positions of the source, considering that the Toolkit only contains the data of the BRIR for the positions of the virtual speakers. Comparing synthetized BRIRs with the Toolkit BRIR for different positions in the horizontal plane, we found a very good correlation between both (correlation coefficients higher than 0.81), which suggest that this implementation offers a good approximation for reverb simulation in any place of the 3D space. We have also evaluated the use of 2D order reverb simulation when no BRIR is available outside the horizontal place, obtaining a good correlation comparing with the 3D approximation used when all BRIRs are available.

Engel et al. (2021) used the 3DTI Toolkit-BS to study the perceptual implications of different Ambisonics approximations (with different spatial orders) to simulate reverberation for binaural listening. Their results suggested that, thanks to the fact that

the direct sound is processed separately to the reverb sound, the perceived quality of the spatialized sound stops improving from the third Ambisonics order, which is lower than the threshold known from previous studies. In a second experiment, they did a perceptual evaluation where the virtual Ambisonics approximation was compared with a first order Ambisonics which take into account the head rotations, to simulate the reverb path. Their results showed that the virtual Ambisonics approximation was not systematically rated lower than Ambisonics, suggesting that the simplifications in the virtual Ambisonics rendering, where the orientation of the listener head within the room is not considered, do not significantly impair subjective preference.

In dynamic rendering, i.e. real time simulation with moving source and listener, when the listener or the source changes its position, the system should update data (HRTF, ILD filter coefficients, etc.) while processing the audio signal. If the updating is instantly accomplished, the audio delivered to the listener suddenly change, which may generate audible artefacts. To obtain smooth transitions the 3DTI Toolkit-BS implements a set of smoothing mechanisms. This chapter showed the evaluation of the reduction of non-lineal artefacts when the sound signal change abruptly. Results show the percentage of the energy out of band when a delta signal is used as input signal, obtaining very low values (less than 0.5% in every situation). These values are lower than the ones reported by Belloch et al. (2013), which best values where obtained with a cross fading using different fading vectors, having a percentage between 1% and 20%, depending the position of the sound source. In addition, all mechanisms shown by Belloch are based on cross fading, which adds latency to the process.

Finally, the real-time performance of the 3DTI Toolkit is shown, demonstrating that for many different scenarios (varying number of sources and different room sizes) the Toolkit can process the whole auditory scene without any dropouts of samples or generating additional latency, using a conventional pc. For example, for a medium room with 30 sound sources with a frame size of 512 samples, the anechoic path takes 40% of the CPU time on our test computer, and the reverberation path takes 20% of the CPU time. Consequently, this scenario takes 60% of the total CPU time, freeing up 40 % of CPU time for other pc processes. This rendering, taking a sample rate of 44100 Hz, implies a latency of 11.6 ms, which is an acceptable latency for real-time applications according to (Brungart, D., Kordik, A. J., & Simpson, 2006). They suggested that a system latency lower than 60 ms is adequate for most applications, and a time less than 30 ms is difficult to perceive in highly demanding acoustic VAS. If the latency wants to be reduced, we should choose a lower frame size, taking into account that it will be more expensive, requiring a higher percentage of CPU time.

The possibility of having low latency and the reduction of audible artefacts in dynamic situations are considered to be fundamental conditions for real-time interactive

VR applications. These features offer more realism to the scene and allow the listener to be more immersive. This makes the 3DTI Toolkit very relevant for its uses in this kind of virtual environments.

Chapter 5

Study of the impact of non-individual HRTFs on speech intelligibility

Related with the goal of using the 3DTI Toolkit for psychoacoustical experiments, this chapter describes a study performed using the Toolkit, where the impact of a set of non-individual HRTF on speech intelligibility in a cocktail party situation is analyzed. The structure of this chapter is as follows. Section 5.1 introduces the research question of the experiment and some basic concepts needed to understand the study. Section 5.2 presents what is known about the topic of Speech intelligibility in an environment with spatially separated sound sources, with a description of the state of the art. A summary of the study and the hypothesis are outlined in Section 5.3. Then, Section 5.4 includes the description of the Material and Methods of the experiments. Finally, the obtained results, their analysis and some discussions are presented in Section 5.5 followed by Section 5.6 that summarizes the main conclusions.

5.1 Introduction

The 3DTI Toolkit-BS presented in previous chapters allows to carry out VR-based psychoacoustic experiments as it can render virtual acoustic environments with multiple sources, each with a different configuration and using any selected HRTF. Thanks to its extensive configurability, it can be used to perform many different types of studies, such as experiments to investigate human sound localization abilities, to study the effect of different virtual acoustic scenarios configurations on the listener or the ones related with speech-in-noise perception and intelligibility. This chapter presents an example of a virtual psychoacoustic experiment implemented using the 3DTI Toolkit-BS. The idea behind this experiment is described below.

Several techniques for binaural 3D audio simulation require the use of the individual characteristics of the listener, represented in the well-known HRTF. As described in previous chapters, HRTF is an individual characteristic of the listener that represents how the listener's head, neck, body, and especially the pinnae, contribute to modify the incoming sound (Moller et al., 1995). This frequency-dependent set of functions are used to spatialize binaural sounds in a virtual environment and create the illusion of sounds coming from specific locations. It is widely accepted that the choice of the HRTF is important for localisation accuracy and realism regarding the spatial perception of sounds and that non-individual HRTFs will allow listeners a good performance localizing sounds. In addition, previous works demonstrate that some attentional processes, such as the cocktail party effect (presented in Section 1.1.3), use HRTF cues to support focusing auditory attention in a specific direction. It is known that speech intelligibility is improved in a Cocktail Party situation when target and maskers are spatially separated. However, the impact of the choice of the HRTFs within this situation remains unclear. To the best of our knowledge, there are no studies describing how the use of different non-individualized HRTFs affects the speech-in-noise performance for a specific listener.

This chapter presents a study in a VR-based Cocktail Party scenario where 22 participants have been exposed with a frontal speech target and two lateral maskers, spatialised using a set of eight non-individual HRTFs. In this condition, the Speech Reception Threshold (SRT) was measured in several repeated sessions for each HRTF. Results show an impact of the HRTF on the SRT, since 82% of the participants presented significant differences between the different HRTFs. Furthermore, the best and worst HRTFs were different across participants, indicating that that impact can be different for different participants. These findings could be very relevant to several research areas, related with spatial hearing and speech intelligibility. Suggesting that, in a virtual scenario that involves binaural 3D audio and speech-in-noise performances, the choice of the HRTF for each listener should be carefully considered.

5.2 State of the art

In addition to human sound localization abilities, one of the potential benefits of binaural spatial hearing is the improvement of speech-in-noise perception, allowing spatial separation of speech and noise sources by the listener in a Cocktail Party situation. The cocktail party effect is a name usually given to the phenomenon by which most people are able to focus on one voice (considered target) and discriminate against all other sounds (maskers or interferers), also known as "spatial unmasking". Although the cocktail party effect was originally described by Cherry (1953) as the ability to

"recognize what one person is saying when others are speaking at the same time", it has been extensively studied with multiple types of masking sounds (Adelbert W Bronkhorst, 2000; Culling et al., 2004; Hawley et al., 2004; Jones & Litovsky, 2011). Within this chapter we are referring to it as a phenomenon of selective attention which allows humans to focus on a *single speech sound source* when this is competing with *multiple masking noise sources*. Several studies have described how HRTF cues help focusing auditory attention in a specific direction in a cocktail party situation. As mentioned, in this experiment we study the effect using noise maskers, but different works can be found using also speech maskers (Douglas S Brungart, 2001), (Freyman et al., 1999), (Hawley et al., 2004). Culling et al. (2004) and Hawley et al. (2004) demonstrated in different studies that the effect of different spatial configurations were smaller in magnitude for speech-shaped-noise than for speech interferers, although with a similar pattern.

A. W. Bronkhorst & Plomp (1988), presented a study about the contribution of two cues to the spatial unmasking: ITD and head shadow. In this work, they define and quantify the benefit of spatial separation between masker and target to enhance binaural speech intelligibility. To study the effect of these cues, the Speech Reception Threshold (SRT) for speech in noise was used as a subjective method. This work was carried out in a virtual environment with a single steady-state noise masker. In (A. W. Bronkhorst & Plomp, 1992), they repeated the study extending it to a more realistic situation using multi-maskers and concluding that speech intelligibility in a situation of Cocktail Party is determined by many factors among which are the binaural cues (ITD and ILD) and the envelope fluctuations of the maskers.

Later on, different perceptual studies have been carried out based on Bronkhorst and Plomp work (Koehnke & Besing, 1996; Peissig & Kollmeier, 1997). Hawley et al. (2004) studied the importance of four main effect presented in a cocktail party situation: the Spatial Release from Masking (SRM), the properties of the interfering sound, the differences in fundamental frequencies when voices are used as maskers, and the informational masking. Regarding noise interferers, we will focus here on the first effect, the SRM. They measured SRT using one, two and three interferers, all in the horizontal plane, and one target in frontal position. The HRIR used was from a manikin from the AUDIS database (Jens Blauert et al., 1998). According to their study, SRM is contributed by two independent components: (1) monaural advantages, based on best-ear listening (Edmonds & Culling, 2006), which is related with the Interaural Level Difference (ILD), as the target to interferer ratio is better in one of the ears. And (2) binaural advantages, caused by Interaural Time Differences (ITD). An extension of this work was carried out by Culling et al. (2004), who studied the individual role of ILD and ITD in the Speech Reception Threshold (SRT) for multiple independent interferes in common or distributed locations. Culling et al. conducted a series of experiments to clarify the contribution of these cues using HRTFs manipulated (HRIR also obtained

from a manikin from the AUDIS database (Jens Blauert et al., 1998)) to extract one or the other cue. Results revealed that, in case of a spatial separation of the target speech and the interferers, it is more advantageous for the speech intelligibility to have both ITD and ILD cues. SRTs in ILD-only conditions were lower only when interferers were in just one hemisphere. There was no spatial advantage when maskers where in different hemispheres. This suggests that in this case, with ILD-only, the subjects were using the best ear effect (sound level reduction at the sound source opposite ear, due to the head acoustic shadow) instead of the binaural cue. If ITD is included in the HRTF (ITD-only or both ITD and ILD), the SRT will improve when markers (all in the same hemisphere or different) were separated from the target regardless the effect of the individual ear.

Many of these experiments are also related with another set of studies, which have the main goal of finding a mathematical formula to model the effect of the binaural hearing on the speech intelligibility. Adelbert W. Bronkhorst (2015) presented a review on the Cocktail Party problem and presented a model of binaural speech perception. This formula was firstly presented in by Jones & Litovsky (2011) and arose from several sets of published data (A. W. Bronkhorst & Plomp, 1992; Peissig & Kollmeier, 1997). The model is based on an expression to predict the Binaural Intelligibility Level Difference (BILD) for the target located in the frontal position and any configuration of noise maskers, given the number of maskers and their azimuth. Also, Culling et al. (2004) found that Bronkhorst's model predicts their results when ITD and ILD are combined. Moreover, they proposed a formula to estimate the SRT (Culling et al., 2005) in anechoic configurations with multiple interferers at different azimuths, considering the cross correlation between both left and right impulse responses. Based on that formula, Lavandier & Culling (2010) developed a model to predict SRT with multiple interferers spatially separated, including the effect of room acoustics in estimations of speech intelligibility. To create reverberant noise, they sum all the interferers and convolve the signal with the BRIR, same with the target. Then, they processed the signals using two paths. The first one calculates the advantage caused by the binaural unmasking, predicting the BMLD using the Equalization-Cancellation theory and the formula from Culling et al. (2004). The second path predicts the effects of the better-ear listening, calculating the SNR as the target-to-interferer ratio at each frequency. Both paths integrate the signals across frequency using the Speech Intelligibility Index (SII) weighting method (ANSI, 1997). This model was later optimized and tested with multiple, spatially distributed interferers by Jelfs et al. (2011). Jelfs' model manage each interfere signal separately and operates directly upon BRIR, instead of summing the noise samples at the beginning of the chain and convolve the result with the BRIR. The revised model also introduces an improvement of the Lavandier & Culling (2010) model using gammatone filters in the second path of the model implementation, where effect of best ear is estimated. Jones & Litovsky (2011) also presented a model for estimating

the SRT in a cocktail party situation. This revised model allows the use of multiple speech maskers and use as input the Binaural Impulse Response. The model predicts the spatial benefit as a combination of two components: the BILD (from Adelbert W. Bronkhorst (2015) model) and the signal-to-noise ratio at the ears.

These previous works demonstrate that the cocktail party effect is enhanced when binaural sound is used. All of them suggested that the mechanisms that arise thanks to the best ear and binaural unmasking are largely sufficient to explain performance in a Cocktail Party situation. In fact, all the presented models are based in these two components. However, they do not consider the role of individual differences. They are able to predict an average performance, which depends on spatial configuration of sources, and some of them room acoustics, but do not take into account how listeners can leverage the individual characteristics of the HRTF in the Cocktail Party problem, and more specifically, the impact of the spectral cues of each HRTF.

5.3 Summary and hypothesis

The importance of the use of the individual HRTF is widely known. In addition, and regarding spatial localization performances, several works have been carried out to study the effect of using non-individual HRTFs. However, the impact of non-individualized HRTFs on speech in noise ratio has not been studied yet, since, as far as we know, all the studies use a generic HRTF measured in a manikin to study these effects. Furthermore, considering the individual nature of HRTFs and its relevance with speech intelligibility within a Cocktail Party situation, the following question arises: could we assess the fit of an HRTF to a specific subject by observing the performances in a VR-based Cocktail Party task?

The main goal of this work is to use the 3DTI Toolkit-BS presented in previous chapter to evaluate, with real subjects, the impact of different non-individual HRTFs (measured from another individual) on the Speech Reception Threshold (SRT) within a Cocktail party situation. The study presents the following hypothesis:

- H1: There is a significant effect of the HRTF (used to create virtual spatialized sound) on speech recognition within a Cocktail Party scenario. That is to say, for a given subject, different HRTFs provide different performances in terms of speech recognition in an environment where masking noise comes from a different direction than speech.
- H2: The influence of a specific HRTF on speech recognition in Cocktail Party scenario is different for different subjects. That is to say, different subjects will have different best-matching HRTF from the same database according

with their performance in a Cocktail party situation. In this way, there are non-individualized HRTFs that are universally better or worse than others are when evaluated on specific task.

5.4 Material and methods

5.4.1 Pilot experiment

A pilot experiment was carried out in preparation for the final and larger experiment. With this pilot study we wanted to test the experiment procedure, decide some details of the experimental design and identify potential problems. Two subjects participated in the pilot experiment, one male and one female.

The pilot experiment design was similar to that of the final experiment. It consisted in a virtual Cocktail Party scenario where the SRT was estimated for 8 different non-individual HRTFs, in order to evaluate the impact of the HRTF on the SRT. One target speech source was placed in front of the listener (azimuth = 0°), and two maskers were placed in different positions according to three different Masker Configurations (MC), shown in Figure 62, all of them on the horizontal plane.

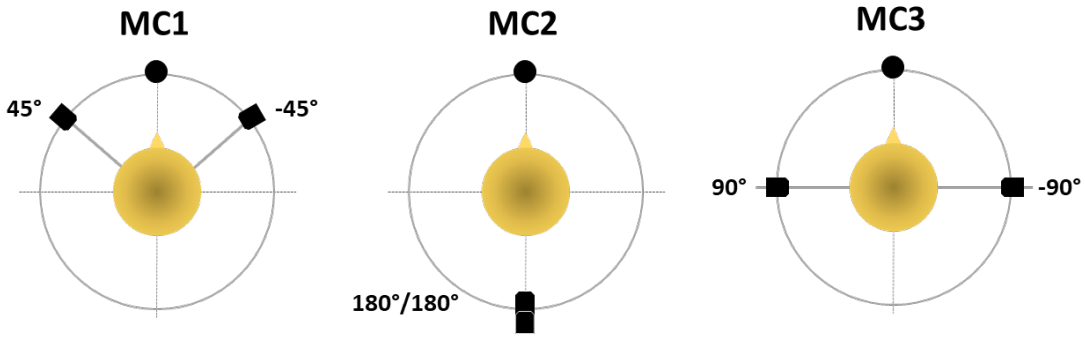


Figure 63. Configuration of the three virtual scenarios of the pilot experiment.

Each participant carried out 10 sessions. In each session the SRT value was estimated for each HRTF and virtual scenario configuration. Each participant was analysed as an independent experiment. A two-way ANOVA was carried out with the HRTF condition and the MC as the independent variables. For participant 1 we found that the MC had a significant effect ($F_{2,18}$=10,212; p=0,001) while the effect of the HRTF was not significant ($F_{8,72}$=1,256; p=0,278). For participant 2 both factors had significant effect, MC with $F_{2,18}$=7,458; p=0,004 and HRTF with $F_{8,72}$=3,502; p=0,002.

Regarding the different Masker Configurations, the spatial separation between target and maskers of the MC2 produces the larger spatial advantage and the analysis of the

results presented significant differences in the effect of the HRTF on the SRT (participant 1 with $F_{8,72}$=4,27; p<0,001 and participant 2 with $F_{8,64}$=3,304; p=0,003). We can think that this is due to the attenuation presented by the HRTFs when the source is behind the listener, and especially in cases where the HRTF have been measured in a listener with prominent or protruding ears (N. Gupta et al., 2002). This can cause larger attenuations of the maskers than the target and helps in the speech intelligibility, regardless the binaural and spectral cues of the HRTF. Regarding the MC1, the analysis of the results did not show a significant influence of the HRTF on the SRT (participant 1 with $F_{8,72}$=0,585; p=0,787 and participant 2 with $F_{8,64}$=1,532; p=0,164). Again, this evidence suggested that the orientation of the ears, which are at a slight angle to the face, has an influence on the speech recognition for this MC, since, in this case, the maskers are in ±45° azimuth and they will be amplified more than the target at 0° azimuth. Finally, MC3 data analysis presented significant differences in the impact of the HRTF on the SRT for both participants (participant 1 with $F_{8,72}$=4,878; p<0,001 and participant 2 with $F_{8,64}$=2,992; p=0,007). Due to those results we concluded that the final experiment could skip MC1 and MC2 and use only MC3. In this way, we would also reduce the time for each session which would allow us to carry out a larger number of sessions per participant.

In addition, from the pilot experiment results, we obtained the variance of the estimated SRT for each participant. Using the GPower statistical analysis tool (Mayr et al., 2007), and both estimated variance, we concluded that 20 sessions would be needed for each participant for the final experiment.

5.4.2 Participants and ethics

Twenty-three subjects were recruited to participate in the final experiment. All of them were students and researchers of the School of Telecommunication Engineering in the University of Malaga. 22 participants finished the experiment (16 males and 6 females), 17 of them with ages between 18 and 29, and 5 of them with ages between 30 and 50. This number of participants was chosen based on a previous study where a similar HRTF set was analysed (Parseihian & Katz, 2012). All participants were Spanish native speakers with normal hearing. In gratitude, they received an USB stick at the end of the study.

The study was approved by the Ethical Committee for Research in Málaga (*Comité de Ética de la Investigación Provincial de Málaga*) and participants gave written informed consent. Appendix A shows the original documents (in Spanish) of the application for the ethics committee, the approval, the participants informed consent form and the demographic questionnaire.

5.4.3 Stimuli

The target sound consisted of one word that came from a position in front of the listener. Before playing the target word, a sentence saying "*por favor, escriba la palabra*" (*please, type the word*) was always played back in the same position as the target, in order to help to focus the attention on the target source direction. The target stimulus was taken from a set of 221 two-syllable Spanish words spoken by a female voice, obtained from a list used for logo-audiometry studies (de Cárdenas & Marrero Aguiar, 1994). These words presented small redundancy, phonetic and syllabic structure balance with Spanish language, similar difficulty and similar familiarity. Figure 63 shows the long-term average spectra of the target and maskers' signal, computed using the IoSR Matlab Toolbox (IoSR, n.d.). The spectra were calculated using the average power spectral density (PSD) obtained from a series of overlapping FFTs (Hann-windowed) of 4096 samples. The average PSD was then Gaussian-smoothed to 1/3-octave resolution.

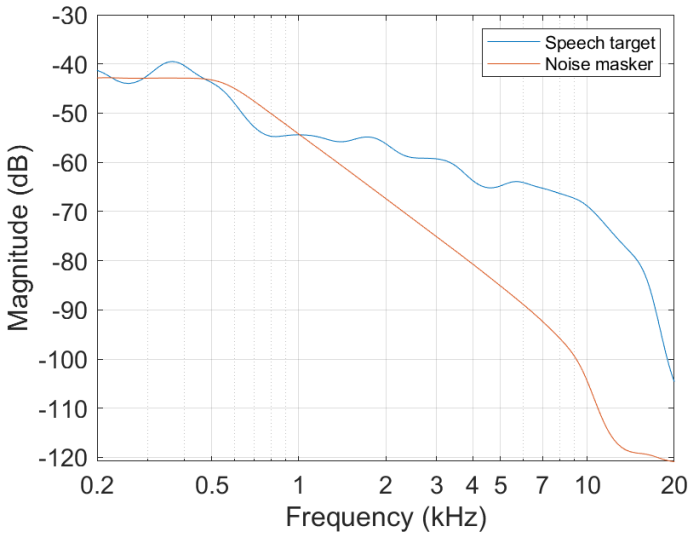


Figure 64. Long-term average spectra of target and maskers' signals. Normalised at 1 kHz.

The interferers consisted of two uncorrelated masker sources. The power of maskers was fixed at 58 dB (SPL) in each ear before being filtered by the HRTF, which ensure a comfortable level while the power of targets was varying during the experiment. Masker stimulus was a coloured noise with the same spectral density as the target words. They were included in the same database (de Cárdenas & Marrero Aguiar, 1994), and obtained from 20 tracks with identical spectral characteristics, but uncorrelated.

5.4.4 Virtual scenario and HRTF dataset

The cocktail party scenario was composed of one target and two maskers. A schematic showing the distribution can be found in Figure 64. The target sound was virtually placed in front of the listener (azimuth = 0°, elevation = 0°). The two maskers were located at right and left sides of the listener (azimuth = ±90°, elevation = 0°).

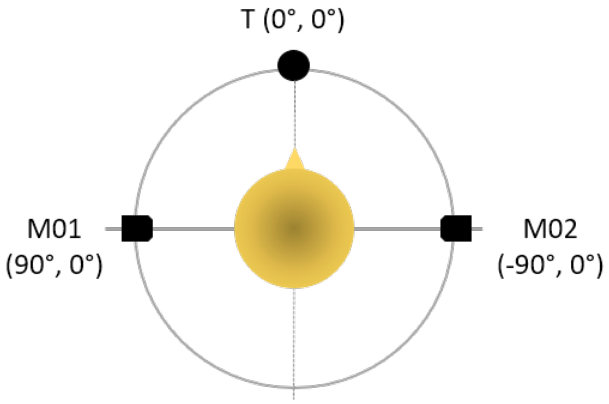


Figure 65. Virtual scenario configuration. Listener is located in the middle. T is the position of the target source, M01 the masker source placed on the left and M02 the masker source placed on the right.

The 3DTI Toolkit-BS (Cuevas-Rodríguez et al., 2019), described in detail in previous chapters, was used to simulate the acoustic virtual scenario. Sound source spatialisation was purely anechoic. A set of eight different non-individual HRTFs were used in the study (7 plus 1).

The set of 7 HRTFs, named as $HRTF_{1-7}$, were taken from the LISTEN database (Warusfel, 2003). The idea was to use the HRTFs that resulted from the Katz & Parseihian (2012) study, as the most representative of the 51 HRTFs of the database. However, and due to an error during the preparation of the experiment, one of the selected HRTF didn't belong to the set presented by Katz, causing that the IRC1002 was taken instead of IRC1032. This corresponded to $HRTF_5$ and it was kept in the fifth position in all the study. The correspondence between the HRTF conditions and the HRTF name in the LISTEN database is presented below:

- $HRTF_1$: IRC1008
- $HRTF_2$: IRC1013
- $HRTF_3$: IRC1022
- $HRTF_4$: IRC1031
- $HRTF_5$: IRC1002
- $HRTF_6$: IRC1048
- $HRTF_7$: IRC1053

Finally, the eighth HRTF was a synthetic snowman-head HRTF used as a control condition and denoted as $HRTF_C$. The synthetic HRTF consists of a spherical-head model that did not provide the effects of the pinnae and only contained the two binaural cues, ITD and ILD. ITD was modelled as a time delay function, using the Woodworth's formula (Woodworth et al., 1954). This formula defines the difference in the arrival time of a wave sound as $r\,(\theta + sin\,\theta)/C$, where r is head radius and was set to 8,75 cm, θ is the azimuth of the source, and C is the speed of sound. ILD was built as a simple one-pole one-zero model, based on the analytical model obtained by Rayleigh & Lodge (1904). The $HRTF_C$ was normalized in order to have the same power as the power average of the LISTEN HRTF set in the front position azimuth = 0°, elevation = 0°.

A numerical analysis of the HRTFs used in the experiment has also been carried out. Figure 65 shows the magnitude of the HRTFs used in the experiment, for the target and masker positions. Looking at these positions, it is possible to see how each HRTF presents a noticeably different spectrum below 10 kHz, where both target and maskers have most of their signal energy (Figure 63). This suggests that they could impact differently in terms of measured speech intelligibility using the chosen experimental configuration. This has been accounted for in the data analysis, found in Section 5.5.

Figure 66 shows the ITD and ILD values for the HRTFs used in the experiment, for both the target and masker positions. ITDs were calculated using a modified threshold method similar to the one presented in (Katz & Noisternig, 2014), where a comparison of the left and right signals was carried out using a threshold detector in order to identify the first arrival time of the incident sound. A threshold of 5% of the maximum amplitude in each HRIR was chosen to detect the onset, visually checking that all HRIR were aligned when the initial silence up to the threshold was removed. ILDs were calculated using the magnitude difference between left and right signals, then averaged by 30 uniformly spaced frequency bands between 1.5KHz and 20kHz on an ERB scale (B. Moore & Glasberg, 1983).

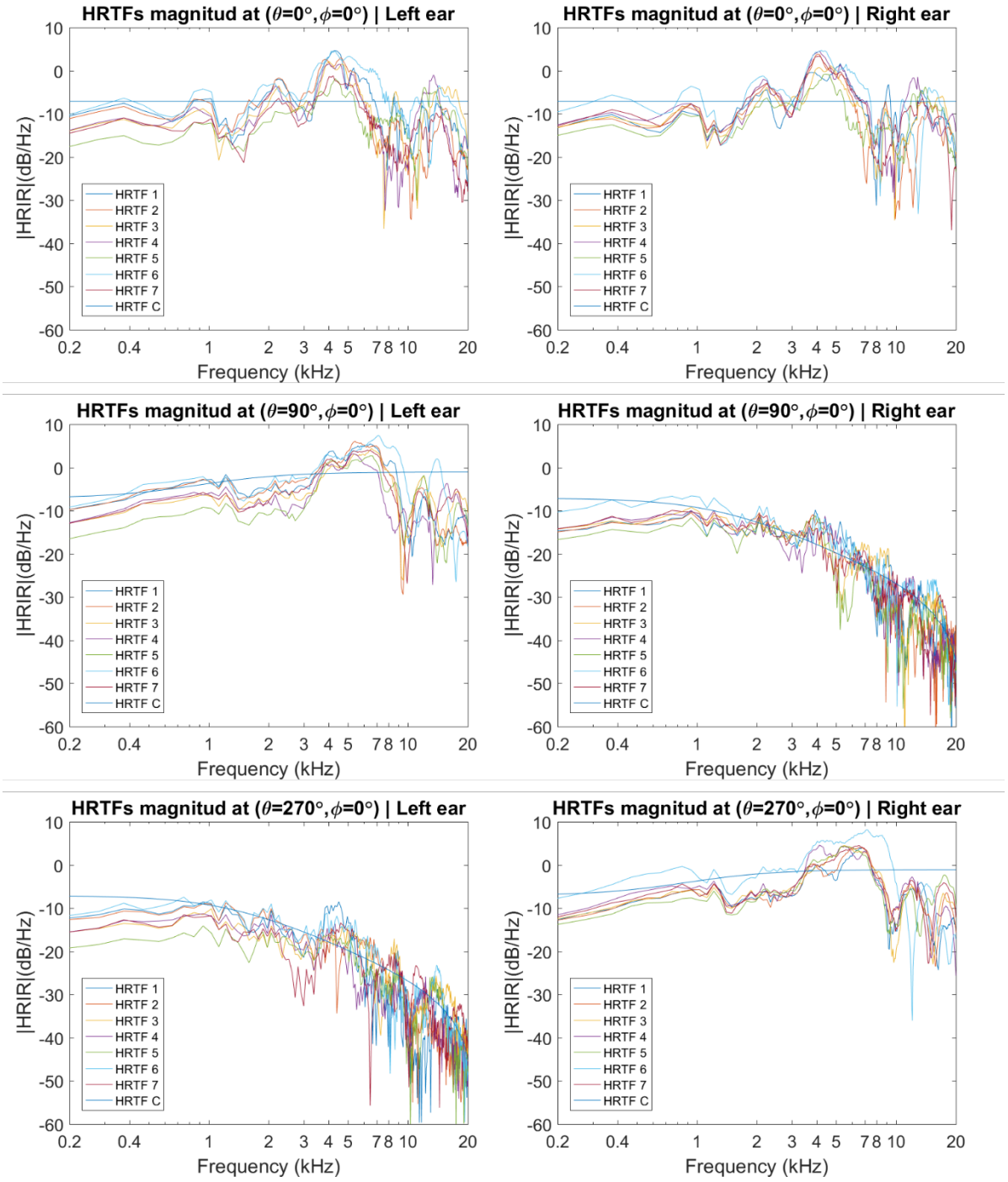


Figure 66. Power Spectral Density of the HRTFs used in the study, for the target position ($\theta = 0°, \phi = 0°$) and masker positions ($\theta = 90°, \phi = 0°$) and ($\theta = 270°, \phi = 0°$) and left and right ear.

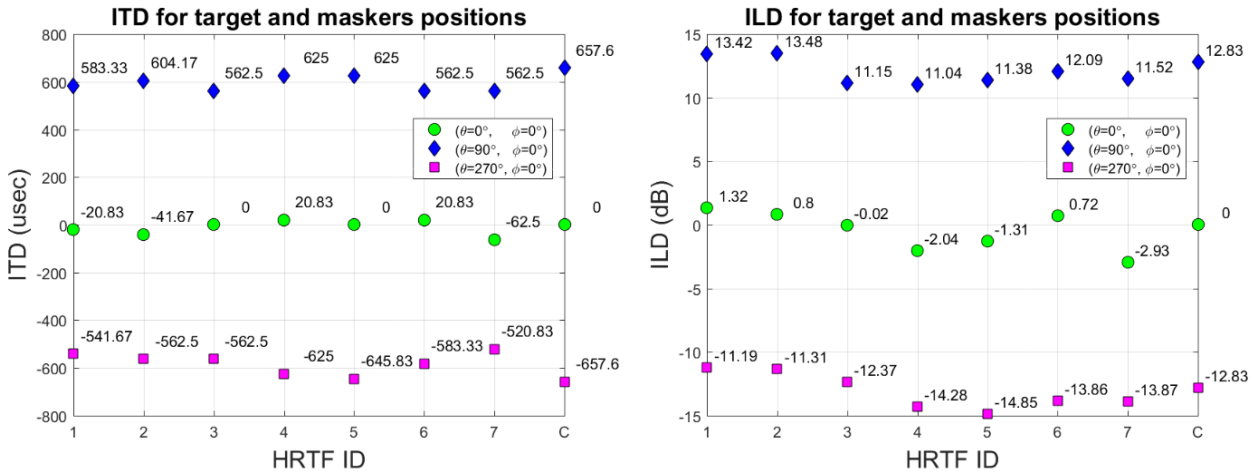


Figure 67. ITD (left) and ILD (right) of the HRTFs used in the study, for the target position ($\theta = 0°, \phi = 0°$) and masker positions ($\theta = 90°, \phi = 0°$) and ($\theta = 270°, \phi = 0°$).

5.4.5 Apparatus

Each participant was seated in front of a monitor with a keyboard and a mouse. A software application was developed specifically for this study. The application automatically sequenced the whole procedure of the experiment, without any intervention of the experimenter. In addition, it included the 3DTI Toolkit-BS library to render spatial sounds and used openFrameworks[27] to create the user interface. A MOTU 896 mk3 audio interface was used to reproduce the sound, connected to the computer using an ASIO driver.

To listen the sound, participants had to wear a pair of headphones SONY MDR-7506. Previous studies have shown that the transfer function between headphones and eardrums (HpTF) can play a role in terms of externalization and overall naturalness of the binaural rendering (Durlach et al., 1992; Masiero & Fels, 2011). Nevertheless, strong evidence has not been found to support that HpTF can improve spatial hearing abilities, such as localisation accuracy (Engel et al., 2019; Schonstein et al., 2008). Furthermore, it has to be noted that HpTFs are not direction dependent, therefore do not vary depending on the position of the source and should not have an influence on HRTF-specific effects, which are the objects of this study. Finally, considering the fact that within this study we explicitly did not want to carry out any personalisation of the rendering and playback systems, and in line with other published research (e.g. (Andreopoulou & Katz, 2016b)), no HpTF was measured and used in this study. In order to ensure consistency within each session and avoid potential spectral alterations

[27] https://openframeworks.cc/ (retrieved January, 2022)

due to repeated donning of the headphones, participants were instructed to wear the headphones at the beginning of each session and not to remove them until the end. To our knowledge, they all complied with this requirement.

The application allowed the participant to type each word, using a keyboard. The software automatically recorded the entire activity of the participant: the word used as target, the typed word, whether it was correct or not, the time elapsed between the word was presented and the participant press the first key and the power level in which the target was played.

5.4.6 Procedure

The main goal of this study is to figure out if there is an impact of the choice on the HRTF on the speech intelligibility. Since this impact can be subject-dependent, it was important to analyse results for each participant separately. Therefore, it was necessary to carry out a large number of repetitions per participant. From the pilot experiment we concluded that each participant had to carry out a total of 20 sessions.

Participants were received the first day and they were informed about the purpose of the experiment and asked to give their written consent. The study consisted on 20 sessions. Participants could choose the days and times they came to the laboratory. They were allowed to carry out a maximum of three sessions per day, keeping a break between sessions of at least 10 minutes. Among the 440 conducted sessions (22 participants and 20 sessions each), the average duration was 11 minutes, with a standard deviation of 2.63 minutes. The procedure of each session is described as follows.

Participants sat at a table with a desktop computer, including a keyboard and a mouse and wore a pair of headphones. The acoustically simulated virtual environment, thought the headphones, was described in section *5.4.4 Virtual scenario and HRTF data.* Participants were instructed to hear the target word and type it. They were told that the target was played in front of them and the maskers at left and right sides. They were instructed to face straight ahead and focus their attention on the target, trying to ignore the surrounding noise. No information about distance was given.

One session was composed of eight blocks. One block corresponds to one HRTF condition ($HRTF_{1-7} + HRTF_C$), ordered randomly in each session. Each block was composed of a set of trials and it had as output one SRT. In this way, at the end of the session we got 8 SRTs, one for each HRTF condition. Participants had to indicate when they wanted to start a block pressing a star button. They were encouraged to rest if they needed between blocks.

Within a block, the SRT was measured using an up-down procedure, which requires repeated presentations of several stimuli in different trials (Levitt, 1971). The structure of the block procedure is described below, and an example is shown in Figure 67.

1. The participant pressed a button "Ready" to start the block. See Figure 68a.
2. The trial started, and maskers and target were reproduced (structure of a trial play backs is explained further below).
3. The participant typed the target word, according to what he/she had heard. See Figure 68b.
4. If the word was correctly identified, the level of the next target decreased 2dB, if not it increased 2dB. The level of the maskers remained always the same.
5. The participant repeated step 2 to 4 until four up-down reversal occurred. We considered a reversal when the target level was increased in previous trial and then it decreases or vice-versa, see examples in Figure 67. At this point, we calculated the Speech to Noise Ratio (SNR), considered as the SNR_i. If after the reversal, the participant correctly identified four targets or fail four targets in a row, the reversal was discarded and the counter i started in 0 again, as it was considered that the failure was for a different reason than not understanding the target word.
6. Once the participant had four reversals (i=4), the block finished.
7. At the end of the block, the SRT value, corresponding to the HRTF condition used in this block, was calculated as the mean of the SNR values of the four reversals (mean of $SNR_{1\text{-}4}$).

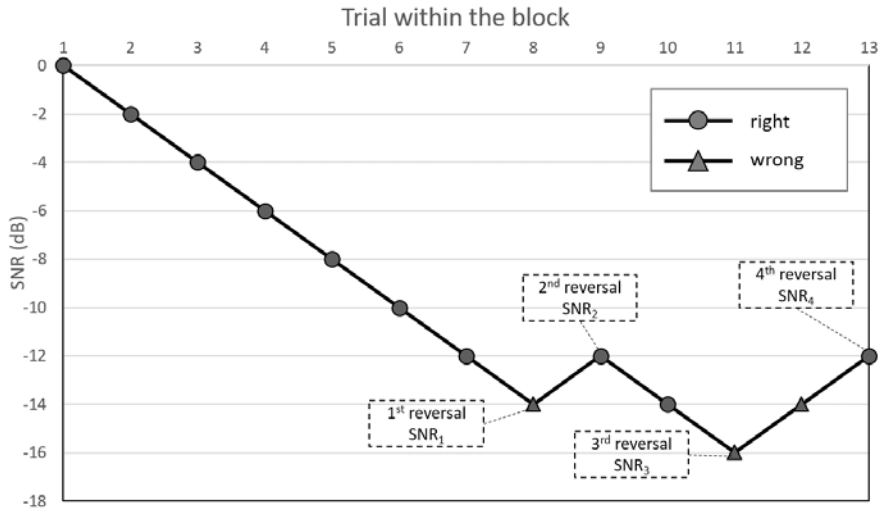


Figure 68. Example of a block procedure

(a) Init screen at the beginning of the block. (b) Trial screen to insert the target word

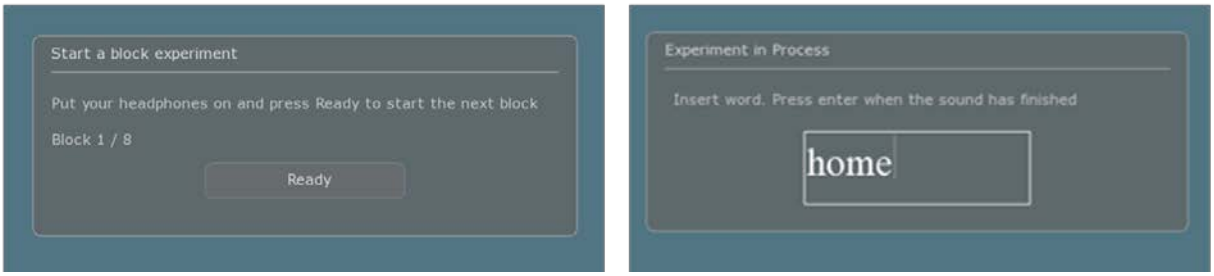


Figure 69. Screenshot of Step 1 and 3 of the block procedure.

In each trial, the participant had to listen to the target and type the word using the computer keyboard. They were instructed to guess in case they could not identify the word, or the word-input space empty if they had no clue. It was considered that a word was correct when it matched the target or when there was a spelling error of one single letter. Both target and maskers were randomly selected in each trial, ensuring that both maskers were uncorrelated. The number of trials within a block depended on the participant's performance, as we will see below.

For the first trial of the block, the target was played 6dB quieter than the sum of the maskers. Here, the total noise level was considered 3dB higher than the noise level of each masker, as maskers were uncorrelated. Therefore, the target was played 3dB quieter than each of the maskers was. In this study, we defined the Speech to Noise Ratio (SNR), and therefore the SRT (Speech Reception Threshold), as the ratio between Speech and one Masker. In this way, we had an initial SNR of -3dB.

The structure of each trial is presented in Figure 69 and described below.

- First, they listened to the prompt, virtually located in the same position as the target.
- After a short silence, randomly selected with a uniform distribution between 500 and 700 milliseconds, the maskers started.
- A few milliseconds later, also randomly selected with a uniform distribution between 200 ms and 800 ms, the target word started.
- The maskers stopped 600 ms after the target finished.

Participants could type the word since the beginning of the trial, but the response could not be confirmed by pressing the "enter" key until the end of the target utterance. Then, the trial finished, and next trial started automatically. No feedback about whether the typed word was correct or not was given.

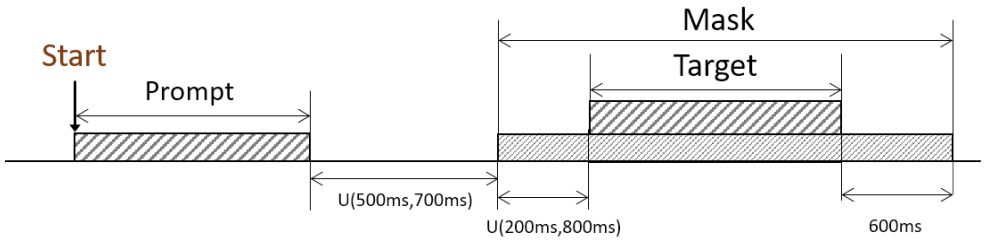


Figure 70. Audio sequence within a Trial

5.5 Results and analysis

A total of 3520 SRTs were measured at the end of the experiment (22 participants x 20 sessions x 8 HRTF conditions). The unprocessed data are referred to here as Raw SRT and will be analyse in Section 5.5.1 . In addition, we compensated the Raw SRT values to remove the overall HRTF-specific benefit, which is to be considered as common for every subject. This compensation was carried out following two approaches. The first one used the masker-target power ration to perform the compensation and the results are shown in Section 5.5.2. The second one uses the SRM from the model of Jelfs et al. (2011), and the results are shown in Section 5.5.3.

5.5.1 Raw data

For all the collected data an individual analysis was carried out for each participant as an independent study (Section 5.5.1.1). In addition, to explore the data regardless of the individual subject differences, an overall analysis was carried out pooling all participants together (Section 5.5.1.2).

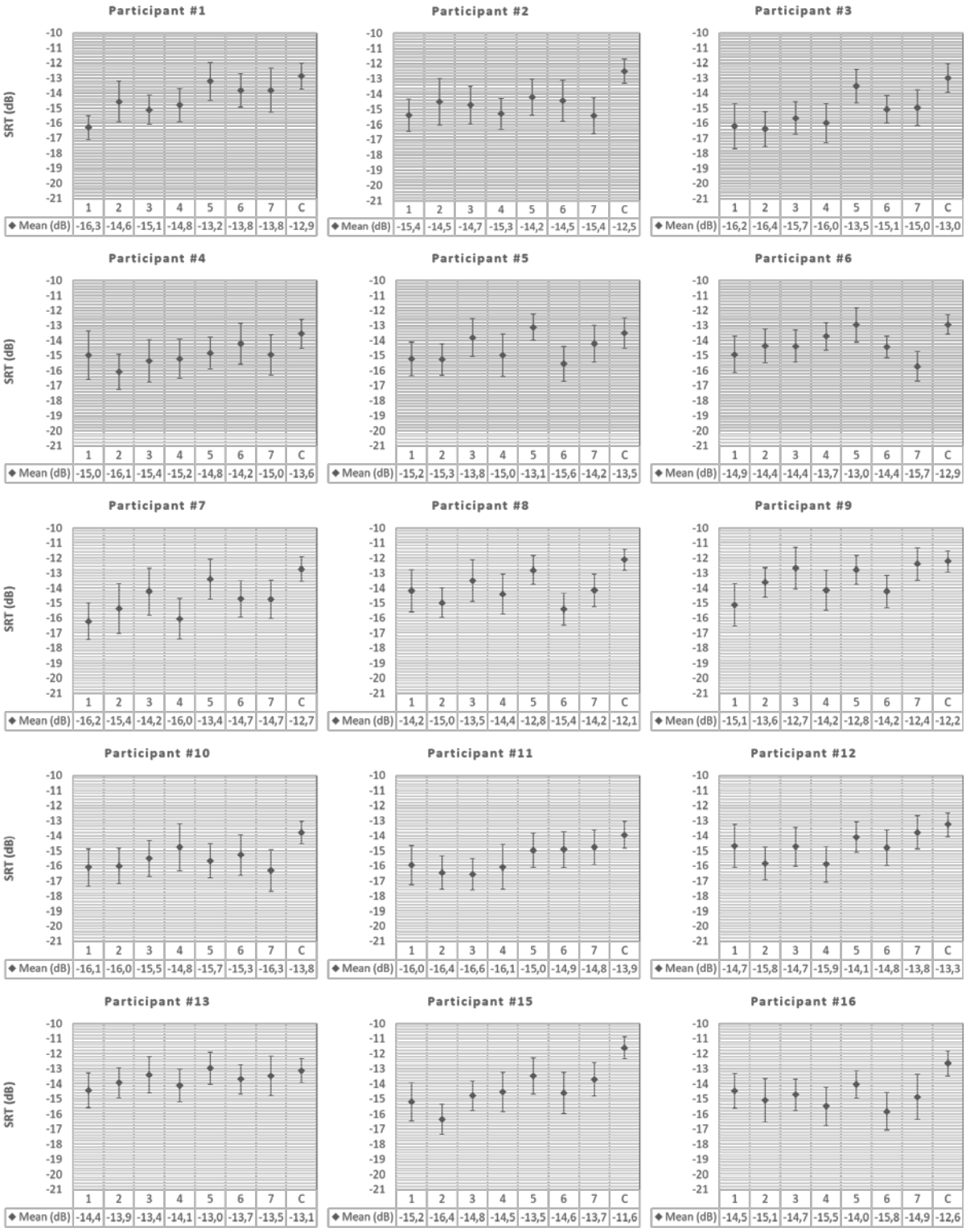
5.5.1.1 Individual analysis

The results of each participant were analysed independently, as this study deals with individual characteristics of the listeners and also tries to the impact of the HRTF choice on a specific subject.

- **Collected data**

Data collected was classified per participant as eight SRT values per session, one for each HRTF condition, with 20 sessions per participant, we had a total of 160 SRT values per participant. The mean SRT for each HRTF condition across sessions is denoted as

$\overline{SRT}_{H_j}^{P_i}$, where P_i indices the participant, with $i \in [1, 22]$ and H_j the HRTF condition, with $j \in [1, 7]$ *and* C (see Section 5.4.4). The mean and the 95% Confidence Interval (CI) are shown in Figure 70, where the Horizontal axis shows the HRTF Condition (j) and the $\overline{SRT}_{H_j}^{P_i}$ in decibels. Vertical axis indicates the SRT value in decibels. The title of each graph shows the participant ID(i).



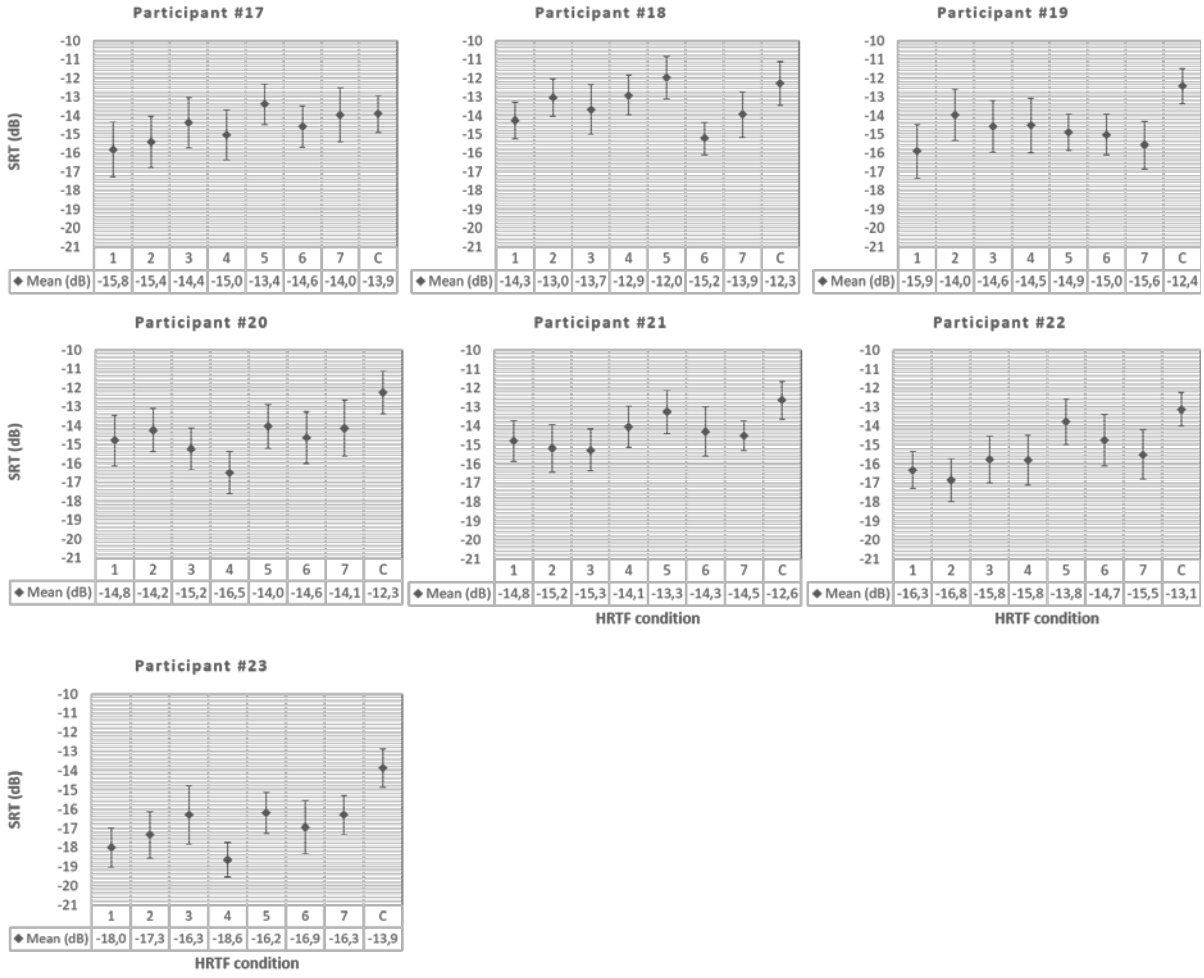


Figure 71. $\overline{SRT}_{H_j}^{P_i}$ and 95% CI of the SRT for each HRTF condition and participant.

The largest the $\overline{SRT}_{H_j}^{P_i}$ value, the worse the speech intelligibility performance of the listener. In this way, for determining the *best HRTF* we took the one with the smallest $\overline{SRT}_{H_j}^{P_i}$, and for the *worst HRTF*, the one with the largest $\overline{SRT}_{H_j}^{P_i}$. The *best HRTF* would be considered as the HRTF that best matches the listener.

Figure 70 provides evidence for which HRTFs have smaller or larger impact on the SRT for each participant. If we look at some specific cases, for example participant #1, we can select as the *best HRTF* the $HRTF_1$, with an $\overline{SRT}$ outside the 95% CIs of the other HRTF conditions. For participant #2, $\overline{SRT}_{H_7}^{P_2}$ is the lowest value, so $HRTF_7$ can be considered as the *best HRTF* for P_2. However, in this case, the effect of this HRTF is smaller, since the $\overline{SRT}_{H_7}^{P_2}$ is within the confidence interval of many other HRTF conditions, so we cannot confidently say that $HRTF_7$ is the best matching HRTF for P_2. However, this way of classifying the HRTF is still of practical use as a methodology to select the best matching HRTF for a specific listener and will be discussed in the Conclusions.

The ***best*** and the *worst HRTF*(excluding the control condition, $HRTF_c$) for each participant are shown in Table 8, as well as the SRT that corresponds with those HRTF ($\overline{SRT}^{P_i}_{H_{best}}$and $\overline{SRT}^{P_i}_{H_{worst}}$). In addition, the last column shows the difference between $\overline{SRT}^{P_i}_{H_{best}}$ and $\overline{SRT}^{P_i}_{H_{worst}}$. Those differences are between -3.25 dB ($\overline{SRT}^{P_{18}}\ diff$) and -1.2 dB ($\overline{SRT}^{P_2}\ diff$). These numbers are comparable with the ranges found in previous studies when looking at BMLD and at the impact of interaural differences on SRM (e.g. Culling et al. (2004)).

Table 8. $\overline{SRT}^{P_i}_{H_j}$ values for the best and worst measured HRTF.

	***best HRTF*$_{1-7}$**		***worst HRTF*$_{1-7}$**		
Participant ID (i)	$\overline{SRT}^{P_i}_{H_{best}}$ **(dB)**	**HRTF condition (j)**	$\overline{SRT}^{P_i}_{H_{worst}}$ **(dB)**	**HRTF condition (j)**	$\overline{SRT}^{P_i}\ diff$ **(dB)**
#1	-16.27	1	-13.23	5	-3.05
#2	-15.40	7	-14.20	5	-1.20
#3	-16.38	2	-13.53	5	-2.85
#4	-16.05	2	-14.20	6	-1.85
#5	-15.55	6	-13.13	5	-2.43
#6	-15.70	7	-12.95	5	-2.75
#7	-16.20	1	-13.40	5	-2.80
#8	-15.40	6	-12.80	5	-2.60
#9	-15.13	1	-12.38	7	-2.75
#10	-16.30	7	-14.75	4	-1.55
#11	-16.55	3	-14.75	7	-1.80
#12	-15.88	4	-13.75	7	-2.13
#13	-14.40	1	-12.95	5	-1.45
#15	-16.35	2	-13.48	5	-2.88
#16	-15.80	6	-14.03	5	-1.78
#17	-15.78	1	-13.38	5	-2.40
#18	-15.20	6	-11.95	5	-3.25
#19	-15.88	1	-13.95	2	-1.93
#20	-16.48	4	-14.03	5	-2.45
#21	-15.25	3	-13.25	5	-2.00
#22	-16.82	2	-13.78	5	-3.05
#23	-18.63	4	-16.18	5	-2.45

From Table 8 column 3 and 5, we obtained Table 9, which shows for which percentage of participants each HRTF conditions is the best and the worst. The *best HRTF* is different across participants, being the $HRTF_1$ the most common one but just

for the 26% of the participants (6 out of 23). However, the *worst HRTF* has turned out to be the same one for the majority, the $HRTF_5$, with a 70% of the participants (16 out of 23). This would reject Hypothesis 2, where it is stated that there is not universally better or worse HRTF for all participants. We will continue investigating this issue in the overall statistical analysis.

Table 9. Percentage of participants where each HRTF conditions is the best or the worst

H_j	% of participants where $H_j$is the best	% of participants where $H_j$is the worst
$HRTF_1$	26%	0%
$HRTF_2$	17%	4%
$HRTF_3$	9%	0%
$HRTF_4$	13%	4%
$HRTF_5$	0%	70%
$HRTF_6$	17%	4%
$HRTF_7$	13%	14%

- **Statistical analysis**

Considering the SRT as the dependent variable and the HRTF condition as the independent one, a one-way ANOVA was performed for each participant. Results are shown in Table 10. The first column indicates the participant ID, and columns 2 and 3 the ANOVA results when all conditions are included. Columns 4 and 5 are the results excluding $HRTF_C$ from the analysis.

If the eight conditions were included in the analysis, 18 out of 22 participants presented a statistical difference on the impact of the HRTF on the SRT. If the $HRTF_C$ was excluded, 9 out of 22 participants presented significant differences in the SRT achieved using different HRTFs. Those p-values are a good evidence to think that there is an impact of the choice of the HRTF on the SRT.

Then, to know which pairs of the HRTF conditions were significantly different from each other, a post-hoc simple pairwise comparison was carried out using the Fisher's Least Significant Difference (LSD) test. The test was made per participant; however, all participants together are shown in Table 11. Only participants that had significant differences ($p< 0.05$) for a specific HRTF pair are included in the table. HRTF conditions (H_j) are specified in the header and left column of Table 11.

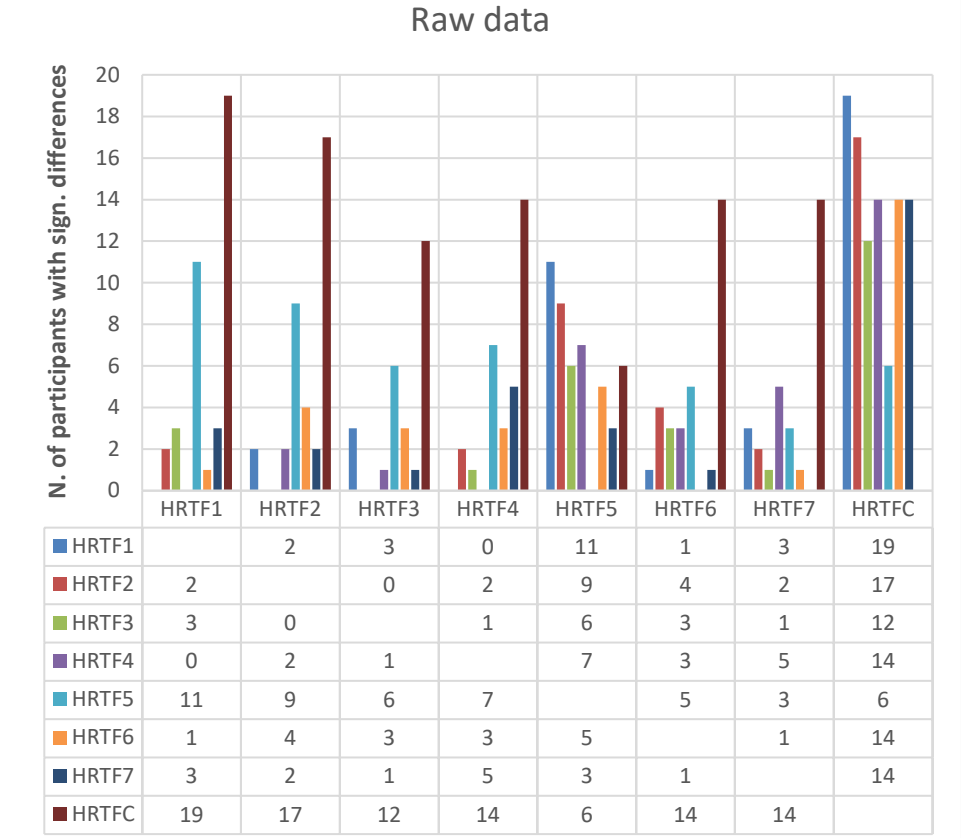
Table 10. ANOVA results

	$\mathbf{HRTF_{1\text{-}7} + HRTF_C}$		$\mathbf{HRTF_{1\text{-}7}}$	
Participant ID	$\mathbf{F_{7,152}}$	**p-value**	$\mathbf{F_{6,133}}$	**p-value**
#1	3.62	0.001**	2.88	0.011*
#2	2.45	0.020*	0.623	0.711
#3	4.40	< 0.001***	2.62	0.019*
#4	1.31	0.247	0.69	0.653
#5	2.44	0.021*	2.25	0.041*
#6	3.49	0.001**	2.71	0.016*
#7	3.20	0.003**	1.99	0.071
#8	3.53	0.001**	2.04	0.063
#9	3.57	0.001**	3.07	0.007**
#10	1.72	0.107	0.66	0.676
#11	2.42	0.022*	1.53	0.172
#12	2.51	0.018*	1.75	0.113
#13	0.79	0.591	0.71	0.640
#15	5.85	< 0.001***	2.53	0.023*
#16	2.61	0.014*	0.91	0.485
#17	1.57	0.146	1.56	0.163
#18	3.65	0.001**	3.55	0.002**
#19	2.80	0.008**	0.99	0.429
#20	3.56	0.001**	1.78	0.106
#21	2.67	0.012*	1.50	0.182
#22	4.39	< 0.001***	2.64	0.018*
#23	5.92	< 0.001***	2.47	0.026*

Table 11. Post-hoc simple pairwise LSD comparison for individual analysis. Underlined IDs indicate those participants whose comparison analysis was significant after Bonferroni corrections.

H_j	$HRTF_2$	$HRTF_3$	$HRTF_4$	$HRTF_5$	$HRTF_6$	$HRTF_7$	$HRTF_C$
$HRTF_1$	#01 #19	#07 #09 #23		**#01** **#03** #05 #06 #07 #09 #15 #17 #18 #22 #23	#01	#01 **#09** #23	**#01** **#02** **#03** #05 #06 **#07** #08 **#09** #10 #11 **#15** #16 #17 #18 **#19** #20 #21 **#22** **#23**
$HRTF_2$			#15 #20	**#03** #05 #07 #08 #12 **#15** #17 #21 **#22**	#04 #15 #18 #22	#12 **#15**	#01 #02 **#03** #04 #05 #06 #07 **#08** #10 #11 #12 **#15** #16 #20 **#21** **#22** **#23**
$HRTF_3$			#23	#01 #03 #06 #18 #21 #22	#05 #08 #09	#11	#01 #02 #03 #06 #11 **#15** #16 #19 **#20** **#21** #22 #23
$HRTF_4$				#03 #05 #07 #12 #20 #22 #23	#18 #20 #23	#06 #09 #12 #20 #23	#01 **#02** **#03** **#07** #08 #09 #11 #12 **#15** **#16** #19 **#20** #22 **#23**
$HRTF_5$					#05 #06 #08 #16 **#18**	**#06** #18 #22	#02 #10 #15 #19 #20 #23
$HRTF_6$						#09	#02 #03 #05 #06 #07 **#08** #09 **#15** **#16** **#18** #19 #20 #21 **#23**
$HRTF_7$							**#02** #03 **#06** #07 #08 #10 #15 #16 #18 **#19** #20 #21 #22 #23

In addition, Figure 71 shows the number of participants with significant differences in each pair-wise comparison ($p< 0.05$). The table within the figure indicates the number of participants with significant differences between the HRTF condition indicated in the header and one in the leftmost column. The graph indicates the number of participants with significant differences between the HRTF condition indicated in the horizontal axis and the one corresponding with the colour in the legend.



	HRTF1	HRTF2	HRTF3	HRTF4	HRTF5	HRTF6	HRTF7	HRTFC
HRTF1		2	3	0	11	1	3	19
HRTF2	2		0	2	9	4	2	17
HRTF3	3	0		1	6	3	1	12
HRTF4	0	2	1		7	3	5	14
HRTF5	11	9	6	7		5	3	6
HRTF6	1	4	3	3	5		1	14
HRTF7	3	2	1	5	3	1		14
HRTFC	19	17	12	14	6	14	14	

Figure 72. Post-hoc pairwise comparisons using LSD of the Raw data. Vertical axis indicates the number of participants with significant differences in the pair-wise comparison between the HRTF ID indicated by the colour and the horizontal axis. In addition, this information is also shown in the table below the graph.

When using LSD test, no mathematical correction was made for multiple comparisons, as recommended by some authors (Rothman, 1990; Saville, 2014). They argue that corrections should be done when interpreting the results, but not in the calculations. Therefore, we had to consider the probability of false-positives with our chosen threshold for significance, alpha = 0.05. In our case, there were 28 pair comparisons per participant, with 22 participants, there was a total of 616 comparisons. With such a large number of comparisons, and considering that 5% of them were false positives, we would have approximately one false positive in average per pair comparisons. However, we can see in Table 11 a good distribution of participants along the table, where most pairwise comparison have more than two participants with significance. In addition, Table 11 shows underlined those comparisons with p-value less than 0.05 after Bonferroni corrections. Participant #13 was found as the only one not having any pair-comparisons with significant differences.

In this table we can see that there are more differences than there would be by pure chance, so it seems that different HRTF would have a different impact on the SRT, but

we cannot confirm specific cases. However, what is more evident is that $HRTF_5$ and $HRTF_C$ are significantly different for many participants. We carried out an overall analysis to further explore this issue.

5.5.1.2 Overall analysis

In order to study the impact of HRTF regardless of the individual differences, an overall analysis was performed considering all participants together. To carry out this analysis we used the eight $\overline{SRT}_{H_j}^{P_i}$ calculated for each participant.

- **Collected data**

First, we calculated the mean of the $\overline{SRT}_{H_j}^{P_i}$ across all participants, denoted as $\overline{SRT}_{H_j}$. The $\overline{SRT}_{H_j}$ are shown in Figure 72, together with the 95% CIs, for each HRTF condition (H_j, with $j \in [1,7]\ and\ C$).

For this experiment it is secondary to check that the SRT values obtained are in line with what is expected from this kind of experiments, since to corroborate our hypotheses we only need to compare our different conditions with each other. However, it should be noted that the $\overline{SRT}_{H_j}$ values obtained show a good correspondence findings from previous studies carried out in similar conditions (e.g. A. W. Bronkhorst & Plomp (1988)).

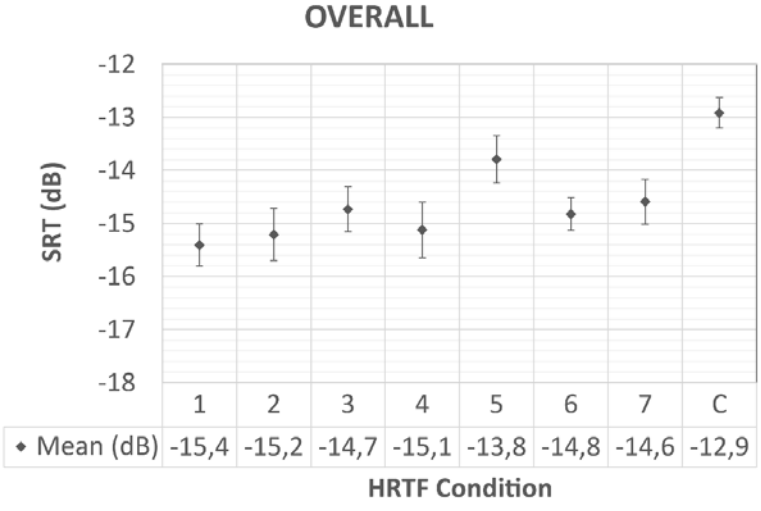


Figure 73. Mean and 95% CI of the SRT for each HRTF condition for the overall study. Horizontal axis shows the HRTF Condition and SRT mean value, in decibels, for each condition. Vertical axis indicates the SRT value in decibels.

As can be seen in the Figure 72, $HRTF_5$ and $HRTF_C$ are the ones that are most different from the others, being the $\overline{SRT}_{H_5}$ and $\overline{SRT}_{H_C}$ outside the CIs of all other

conditions. Those two HRTFs have the largest $\overline{SRT}_{H_j}$values, which means that they are the *worst HRTFs* overall.

- **Statistical analysis**

As for the individual analysis, a one-way ANOVA was carried out with the overall data, using the $\overline{SRT}_{H_j}^{P_i}$. This analysis shows a significant impact of HRTF on SRT when the $HRTF_C$ was included ($F(7,168) = 16.7861$, $p<0.001$), and also when it was removed from the data set ($F(6,147) = 6.3972$, $p<0.001$).

Then, a post-hoc pairwise comparison was carried out using both LSD and Bonferroni adjustments. Results using LSD comparisons are shown in Table 12. This pairwise comparison confirms what it was shown in Figure 72, where $HRTF_5$ and $HRTF_C$ presented significant differences with all other conditions and between them. In addition, this comparison shows significant differences in other pairs, but, considering that no corrections have been applied we cannot claim any specific case. Table 13 shows the pair-wise comparison with the Bonferroni correction, which confirms that even after the corrections, the $HRTF_5$ and $HRTF_C$ have significative differences with all other conditions. Regarding the $HRTF_C$, these results were expected, since this control condition consists in a simplified model of HRTF which includes only interaural differences, but not spectral cues. Nevertheless, the fact that the $HRTF_5$ is very different to the rest was unexpected and, as mentioned before, it made us reconsider Hypothesis2.

Table 12. Post-hoc simple pairwise comparison using LSD for overall analysis

	$HRTF_2$	**$HRTF_3$**	**$HRTF_4$**	**$HRTF_5$**	**$HRTF_6$**	**$HRTF_7$**	**$HRTF_C$**
$HRTF_1$	0.492	0.020*	0.322	< 0.001***	0.045*	0.005**	< 0.001***
$HRTF_2$		0.101	0.761	< 0.001***	0.181	0.032*	< 0.001***
$HRTF_3$			0.180	0.001**	0.758	0.613	< 0.001***
$HRTF_4$				< 0.001***	0.301	0.065	< 0.001***
$HRTF_5$					< 0.001***	0.006**	0.002**
$HRTF_6$						0.416	<0.001***
$HRTF_7$							< 0.001***

Table 13. Post-hoc simple pairwise comparison using Bonferroni for overall analysis

	$HRTF_2$	**$HRTF_3$**	**$HRTF_4$**	**$HRTF_5$**	**$HRTF_6$**	**$HRTF_7$**	**$HRTF_C$**
$HRTF_1$	1	0.577	1	< 0.001***	1	0.141	< 0.001***
$HRTF_2$		1	1	< 0.001***	1	0.912	< 0.001***
$HRTF_3$			1	0.035*	1	1	< 0.001***
$HRTF_4$				< 0.001***	1	1	< 0.001***
$HRTF_5$					0.012*	0.170	0.075
$HRTF_6$						1	< 0.001***
$HRTF_7$							< 0.001***

5.5.1.3 Discussion

When looking at the individual statistical analysis, the choice of HRTF seems to have a significant impact on the SRT scores for the majority of participants, 82% of them. If we look at the pairwise comparison, we can find differences between the different HRTFs, confirming the first of our initial hypotheses (H1). However, all previous results highlighted that $HRTF_5$ and $HRTF_C$ are the conditions that present significantly lower performances than all the other HRTFs. These two cases seem to confirm that the second of our initial hypotheses (H2) should be rejected, since, according to our results, the $HRTF_5$ is usually the worst regarding the measured HRTFs. However, this is not the same as for the best HRTF, which is pretty much spread out over all the conditions.

The $HRTF_C$ consists in a simplified model of HRTF, derived from a snowman spherical model which includes only interaural differences, but not spectral cues. These results seem to indicate that the HRTF spectral cues have a significant impact on SRT in Cocktail Party conditions and that an HRTF without those cues causes a worse performance in the speech intelligibility tasks. Considering $HRTF_5$, although all measured HRTFs come from the same database, $HRTF_5$ is the only one which was not included in the sub-set obtained by Katz and colleagues (Katz & Parseihian, 2012), which was used to select our sample (see Material and Methods section). This is definitely something remarkable, but after studying the spectral characteristics of the $HRTF_5$ nothing special was found comparing with the rest of HRTFs, so it does not seem a convincing justification.

In addition to these mentioned reasons, the fact that the $HRTF_5$ and $HRTF_C$ are worse than others for the overall sample of participants can be due to differences in the power ratio between sides (where maskers were places) and front (where the target was placed). For instance, in the case of a specific HRTF having an attenuating filter for the speech band at ($\phi = 0°,\ \theta = 0°$), the target would be specially attenuated yielding a higher SRT, regardless of other individual differences in the spectral cues of that HRTF. Those differences in the power ratio between different positions can be caused by the size, shape and orientation of the ear pinna, which can provoke an additional "shadowing" or amplification of some of the sounds. For example, one HRTF measured in a listener with protruding ears can be helpful in discriminating sounds placed in frontal versus lateral positions (N. Gupta et al., 2002). On the other hand, previous work looking at binaural loudness, it is argued that sound from the side can be perceived as being louder than sound coming from the other positions (Sivonen & Ellermeier, 2011). Lokki & Pätynen (2011) demonstrated that sources located in lateral positions are perceived as louder than the ones located in frontal positions, due to amplifications/attenuations caused by the specific shape of the human head. In order to investigate if this is happening in our experiment, the next section studies the

amplification/attenuation profile of each HRTF condition in order to adjust the SRT value and analyse again the results.

5.5.2 Data compensated by masker-target power ratio

To calculate the global gain of each $HRTF(\phi,\theta)_{ear}$, where ϕ and θ indicate the sound source position, and *ear* the left or the right ear, they were firstly filtered using an A-weighting filter and secondly integrated to calculate the total energy of each impulse responses for both ears (denoted here as W). A-weighting is a commonly used filter which quantifies each frequency as in the auditory system, in order to mimic the effect of the human hearing. A-weighting filter coefficients are defined in IEC 61672-1 standard (Iec, 2002). Then, the total energy was calculated as the root mean square (rms) of the filtered HRTF:

$$W_{HRTF(\phi,\theta)_{ear}} = rms\,(A_{filtered} HRTF(\phi,\theta)_{ear}) \tag{5.1}$$

Once the total energy of each HRTF was calculated in each ear, we obtained W_M and W_T as the energy of masker and target directions respectively, for a given HRTF, with the following equations:

$$W_M(dB) = 10\log_{10}\frac{W_{HRTF(90,0)_L} + W_{HRTF(-90,0)_L} + W_{HRTF(90,0)_R} + W_{HRTF(-90,0)_R}}{2} \tag{5.2}$$

$$W_T(dB) = 10\log_{10}(W_{HRTF(0,0)_L} + W_{HRTF(0,0)_R}) \tag{5.3}$$

Remember that the position of the target was ($\phi = 0°,\ \theta = 0°$) and maskers where at ($\phi = 90°,\ \theta = 0°$) and ($\phi = -90°,\ \theta = 0°$). The L and R subindices indicate the left and right ear respectively. The energy corresponding to the maskers was divided by 2 because we considered SRT as the ratio between target and **one** masker, as explained in Material and Methods section. Finally, a Masker to Target Ratio (MTR), in decibels, was calculated for each HRTF as the difference between W_M and W_T. The results are shown in Table 14.

$$MTR(dB) = W_M - W_T \tag{5.4}$$

Table 14. Masker-target ratio for each HRTF condition

H_j	$HRTF_1$	$HRTF_2$	$HRTF_3$	$HRTF_4$	$HRTF_5$	$HRTF_6$	$HRTF_7$	$HRTF_C$
MTR_{H_j} (dB)	-0,40	-0,82	-1,17	0,40	-3,20	-1,50	-1,69	-2,76

Notice that larger compensations were needed for $HRTF_5$ and $HRTF_C$, which were the ones that, in the overall analysis, presented significant differences with the rest of the conditions. This tells us that, on the one hand, we have used a control HRTF that is harming the target with respect to the masker, and this should be corrected. And, on the other hand, that $HRTF_5$ could be generating a bad performance caused by an attenuation of the target, regardless the spectral cues of that HRTF.

These ratios were used to adjust the obtained raw SRT for each condition and participant, following the next formula, where P_i is the participant, with $i \in [1, 22]$ and H_j the HRTF condition, with $j \in [1, 7]$ *and* C

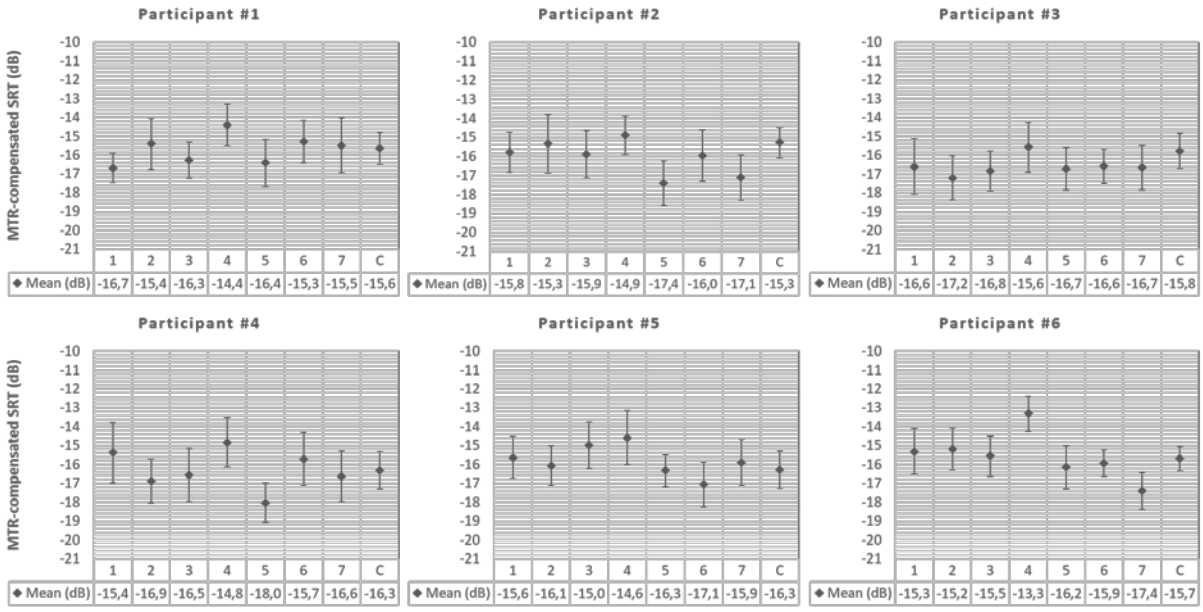
$$MTR\_compensated\ SRT_{H_j}^{P_i}\ (dB) = raw\ SRT_{H_j}^{P_i}\ (dB) + MTR_{H_j}\ (dB) \tag{5.5}$$

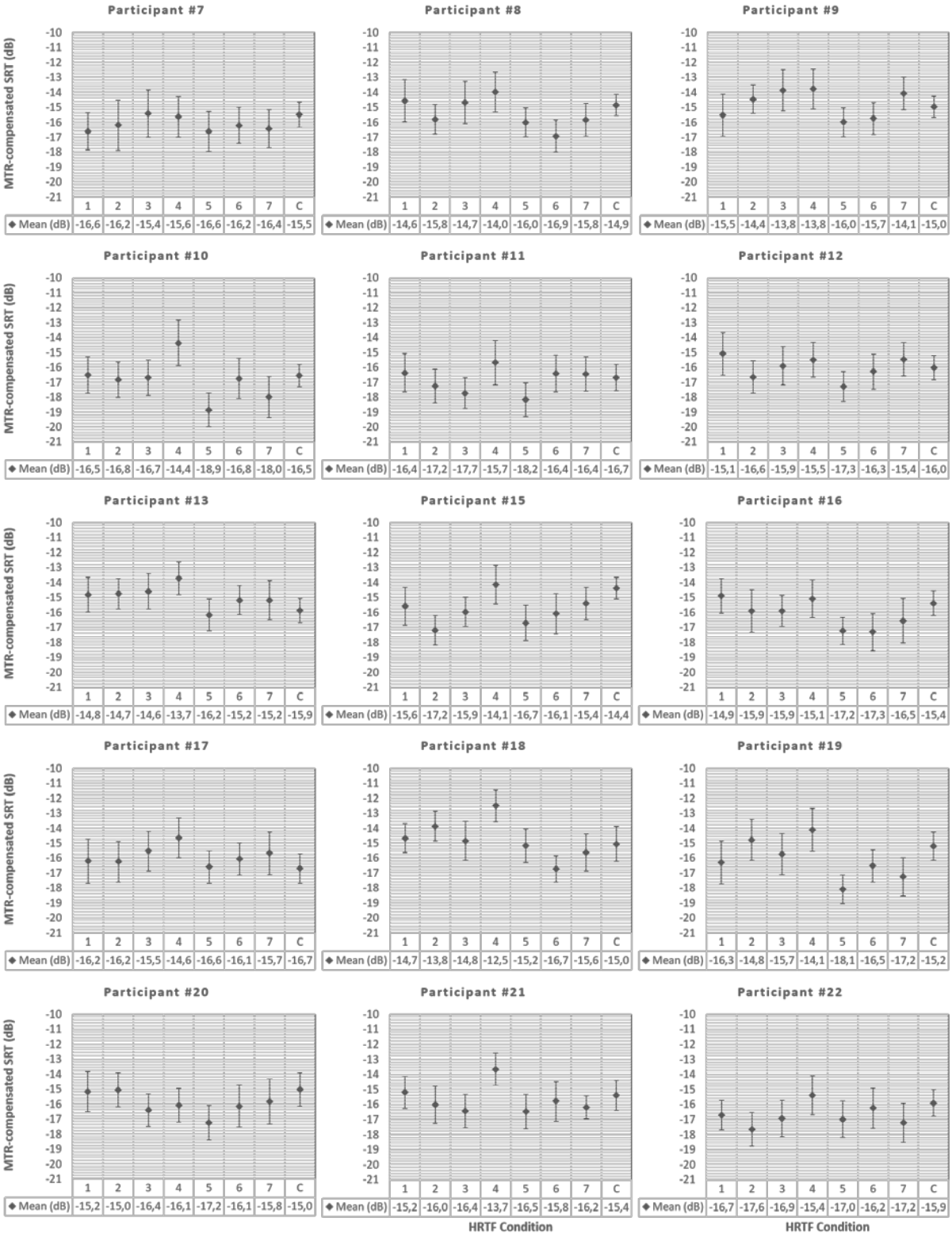
Once the MTR-compensated SRT was obtained, a new analysis was carried out, using the same methods as for the raw SRT data.

5.5.2.1 Individual analysis

- **Collected data**

To see how this compensation affects to the results, we started with an individual analysis to each participant. The MTR-compensated SRT means across sessions ($\overline{SRT^*}_{H_j}^{P_i}$) and 95% CIs are shown in Figure 73. Horizontal axis shows the HRTF Condition and $\overline{SRT}^S$value, in decibels, for each condition. Vertical axis indicates the SRT value in decibels. The title of each graph indicates the participant ID.





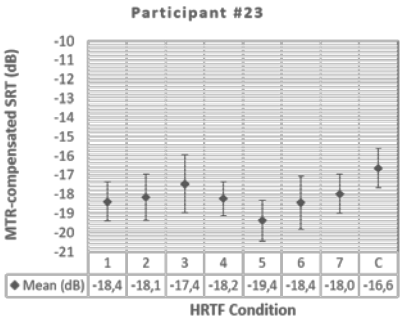


Figure 74. Mean and 95% CI of the MRT-compensated SRT for each HRTF.

The 95% CI remains the same since the compensation affects in the same way to the all the SRT values for a specific HRTF condition. However, the $\overline{SRT^*}^{P_i}_{H_j}$ is now different for each H_j and P_i. Figure 74 shows an example, comparing SRT means before and after the compensation for participant #1.

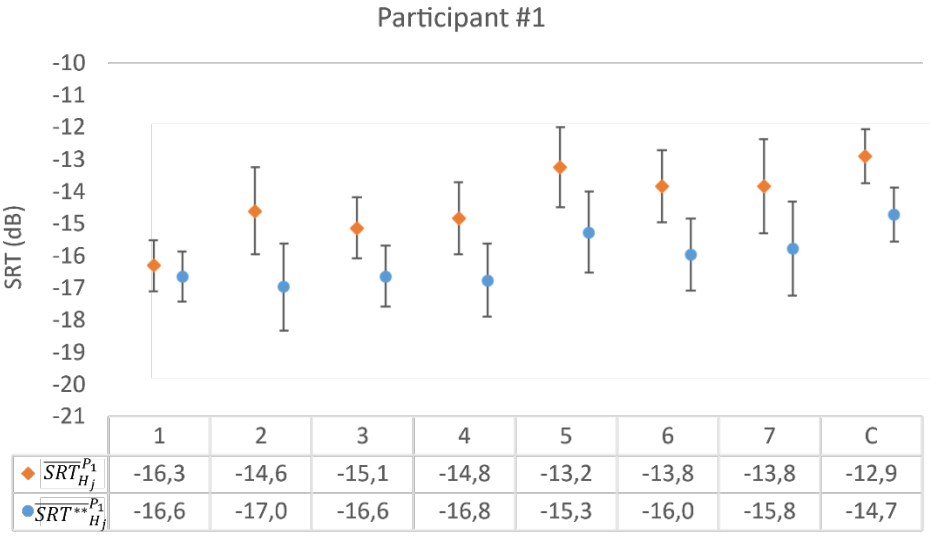


Figure 75. Mean and 95% CI of the raw (diamonds) and MRT-compensated (circles) SRT for each HRTF condition.

As mentioned, $\overline{SRT^*}^{P_1}_{H_5}$ and $\overline{SRT^*}^{P_1}_{H_C}$ have been the most affected after the compensation, followed by the $\overline{SRT^*}^{P_1}_{H_6}$ and $\overline{SRT^*}^{P_1}_{H_7}$. In addition, and this is something that can be seen for many participants, the $\overline{SRT^*}^{P_i}_{H_4}$ has been separated from all other values, which is expected when looking at the data used for the adjustments (MTR_{H_j}), since all of them are negative values, which make the SRT better, while 4 applies a positive value which makes it worse. With these data we see that the compensation caused by the MTR of each HRTF greatly affects the $\overline{SRT^*}^{P_i}_{H_j}$ for each condition. Below we see how the best and worst HRTF obtained in this case have been altered for each participant.

The best and worst measured HRTF after the MTR compensation have also changed and they are shown in Table 15 for each participant. The rightmost column shows the SRT difference in dB between the *best* and the *worst HRTF*. Being the smallest difference $1.2dB$ for participant #7 and the largest difference $4.5\,dB$ for participant #10. This range is consistent with the one obtained for the raw data.

Table 15. MTR-compensated SRT values for the best and worst measured HRTF

	Best $HRTF_{1\text{-}7}$		**Worst $HRTF_{1\text{-}7}$**		
ID	**SRT (dB)**	**HRTF condition**	**SRT (dB)**	**HRTF condition**	**SRT difference (dB)**
#1	-16,67	1	-14,40	4	-2,27
#2	-17,40	5	-14,88	4	-2,52
#3	-17,19	2	-15,57	4	-1,62
#4	-18,02	5	-14,82	4	-3,20
#5	-17,05	6	-14,57	4	-2,48
#6	-17,39	7	-13,30	4	-4,09
#7	-16,60	1	-15,40	3	-1,20
#8	-16,90	6	-13,97	4	-2,93
#9	-15,97	5	-13,75	4	-2,22
#10	-18,85	5	-14,35	4	-4,50
#11	-18,15	5	-15,68	4	-2,47
#12	-17,27	5	-15,07	1	-2,20
#13	-16,15	5	-13,70	4	-2,45
#15	-17,17	2	-14,12	4	-3,04
#16	-17,30	6	-14,85	1	-2,45
#17	-16,57	5	-14,63	4	-1,95
#18	-16,70	6	-12,50	4	-4,20
#19	-18,07	5	-14,10	4	-3,97
#20	-17,22	5	-15,04	2	-2,18
#21	-16,45	5	-13,65	4	-2,80
#22	-17,64	2	-15,38	4	-2,27
#23	-19,37	5	-17,45	3	-1,93

From columns 3 and 5 in Table 15, we obtained Table 16, which shows for which percentage of participants HRTF condition is the best and which is the worst. After the compensation, the *best HRTF* is less distributed than before, being now $HRTF_5$ the best HRTF for half of the participants (11 out of 22). On the other hand, $HRTF_4$ is now the *worst HRTF* for the majority of participants, 74% of the them (16 out of 22).

Table 16. Percentage of participants where each HRTF conditions is the best or the worst

H_j	% of participants where $H_j$is the best	% of participants where $H_j$is the worst
$HRTF_1$	9%	14%
$HRTF_2$	13%	4%
$HRTF_3$	0%	14%
$HRTF_4$	0%	74%
$HRTF_5$	52%	0%
$HRTF_6$	17%	0%
$HRTF_7$	4%	0%

- **Statistical analysis**

A one-way ANOVA was carried out, this time with the MTR-compensated SRT as the dependent variable and the HRTF condition as the independent one. The results are shown in Table 17. The first column indicates the participant ID, columns 2 and 3 indicate the ANOVA results when all conditions are included ($HRTF_{1-7} + HRTF_C$), and columns 4 and 5 indicates the results when the control condition, $HRTF_C$ has been removed in the analysis.

Table 17. ANOVA results for compensated data using the MTR

	$HRTF_{1-7} + HRTF_C$		**$HRTF_{1-7}$**	
Participant ID	**$F_{7,152}$**	**p-value**	**$F_{6,133}$**	**p-value**
#1	1.61	0.137	1.76	0.112
#2	2.13	0.044*	2.08	0.060
#3	0.85	0.549	0.67	0.676
#4	2.27	0.032*	2.49	0.026*
#5	1.80	0.091	1.93	0.080
#6	5.07	< 0.001***	5.47	< 0.001***
#7	0.53	0.808	0.44	0.848
#8	2.72	0.011*	2.81	0.013*
#9	2.54	0.017*	2.76	0.015*
#10	4.23	< 0.001***	4.49	< 0.001***
#11	1.82	0.087	1.99	0.071
#12	1.50	0.172	1.63	0.145
#13	1.92	0.071	1.69	0.129
#15	3.30	0.003**	2.65	0.018*
#16	2.35	0.027*	2.35	0.034*
#17	1.04	0.407	0.92	0.481
#18	4.83	< 0.001***	5.70	< 0.001***

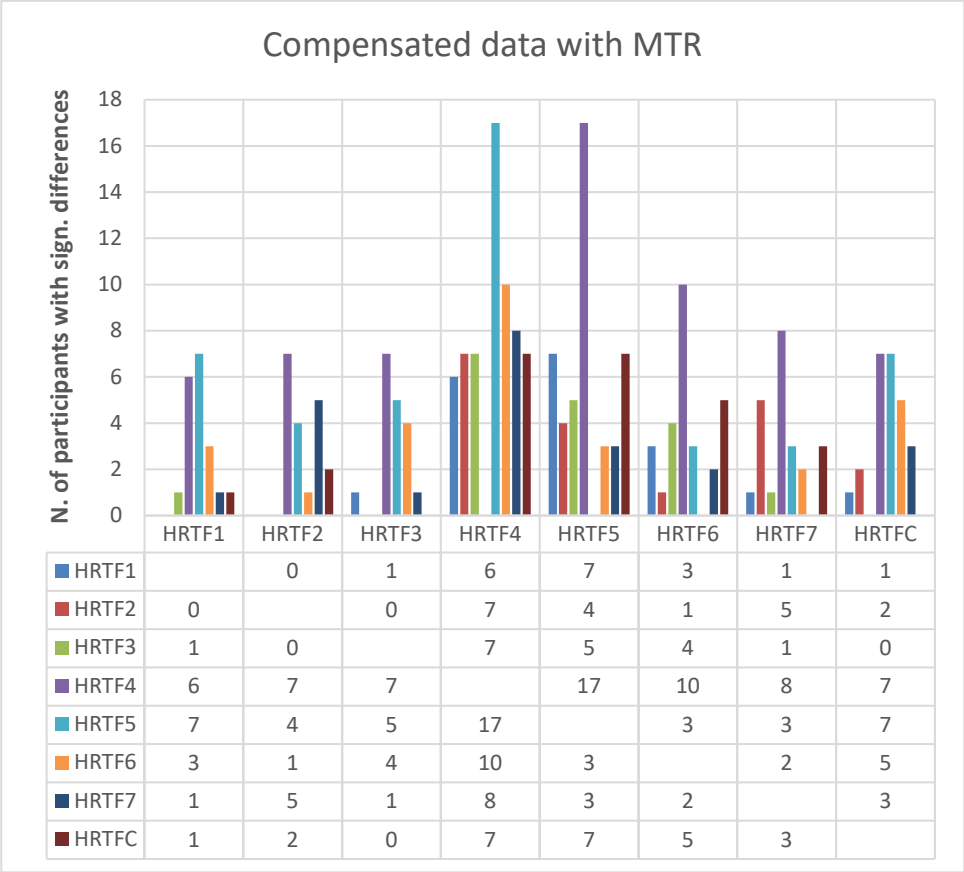
#19	4.26	< 0.001***	4.39	< 0.001***
#20	1.48	0.177	1.36	0.237
#21	2.71	0.011*	3.04	0.008**
#22	1.52	0.166	1.40	0.218
#23	1.83	0.086	0.94	0.468

Using the MTR-compensated SRT data, if the eight conditions ($HRTF_{1-7} + HRTF_C$) are included in the analysis, half of the participants (11 out of 22) present statistical differences. If the $HRTF_C$ is excluded, 10 out of 22 participants present significant differences in the SRT achieved using different HRTFs. In the same way as the previous analysis of the raw data, to know which pairs of the HRTF conditions are significantly different from each other, a post-hoc simple pairwise comparison was carried out using the LSD test and show Table 18. Participants #3 and #7 do not have any pair-comparisons with significant differences. In addition, the table shows underlined participants who present significant differences after Bonferroni corrections.

Table 18. Post-hoc simple pairwise comparison for individual analysis of the compensated SRT data using MTR

HRTF condition	$HRTF_2$	$HRTF_3$	$HRTF_4$	$HRTF_5$	$HRTF_6$	$HRTF_7$	$HRTF_C$
$HRTF_1$		#09	#01 #06 #09 #10 #18 #19	#04 #10 #11 #12 #16 #19 #20	#08 #16 #18	#06	#23
$HRTF_2$			#04 #06 #08 #10 **#15** #21 #22	#02 #10 **#19** #20	**#18**	#02 #06 #15 #18 #19	**#15** #22
$HRTF_3$			#01 #06 #10 #11 #15 #18 **#21**	#09 #10 #13 #19 #23	#05 #08 #09 #18	#06	
$HRTF_4$				#01 #02 **#04** #05 **#06** #08 #09 **#10** #11 #12 #13 #15 #16 #17 **#18** **#19** **#21**	#05 **#06** **#08** #09 #10 #15 #16 **#18** #19 #21	#02 **#06** #08 **#10** **#18** **#19** **#21** #22	#05 **#06** #10 #13 #17 **#18** #21
$HRTF_5$					#04 #10 #11	#09 #11 #12	#02 #10 #15 #16 **#19** #20 **#23**
$HRTF_6$						#06 #09	#08 #15 #16 #18 #23
$HRTF_7$							#02 #06 #19

In addition, Figure 75 shows the number of participants with significant differences in each pair-wise comparison ($p < 0.05$). The table within the figure indicates the number of participants with significant differences between the HRTF condition indicated in the header and one in the leftmost column. The graph indicates the number of participants with significant differences between the HRTF condition indicated in the horizontal axis and the one corresponding with the colour in the legend.



	HRTF1	HRTF2	HRTF3	HRTF4	HRTF5	HRTF6	HRTF7	HRTFC
HRTF1		0	1	6	7	3	1	1
HRTF2	0		0	7	4	1	5	2
HRTF3	1	0		7	5	4	1	0
HRTF4	6	7	7		17	10	8	7
HRTF5	7	4	5	17		3	3	7
HRTF6	3	1	4	10	3		2	5
HRTF7	1	5	1	8	3	2		3
HRTFC	1	2	0	7	7	5	3	

Figure 76. Post-hoc pairwise comparisons using LSD of the Raw data. Vertical axis indicates the number of participants with significant differences in the pair-wise comparison between the HRTF ID indicated by the colour and the horizontal axis. In addition, this information is also shown in the table below the graph

As in the raw data analysis, for the MTR-compensated data analysis there were 28 pair comparisons per participant. With 22 participants, this gives a total of 616 comparisons, and we would have approximately one false positive in average per pair comparisons. Considering that Table 18 shows a good distribution of participants along the table, having more than expected by chance, it seems that there are differences between the different HRTFs, although we cannot draw a conclusion for specific cases. However, it should be noted that $HRTF_4$ and $HRTF_5$ are the ones that, after the

compensation, are more different from the others. To go deeper into this issue, we repeat the overall analysis with the MTR-compensated data.

5.5.2.2 Overall analysis

- **Collected data**

We repeat the overall analysis with the MTR-compensated SRT data. In this case, the $\overline{SRT^*}_{H_j}$ and 95 %CIs are shown in Figure 76. The horizontal axis shows the HRTF Condition and the SRT mean values, in decibels, for each condition. The vertical axis indicates the SRT value in decibels.

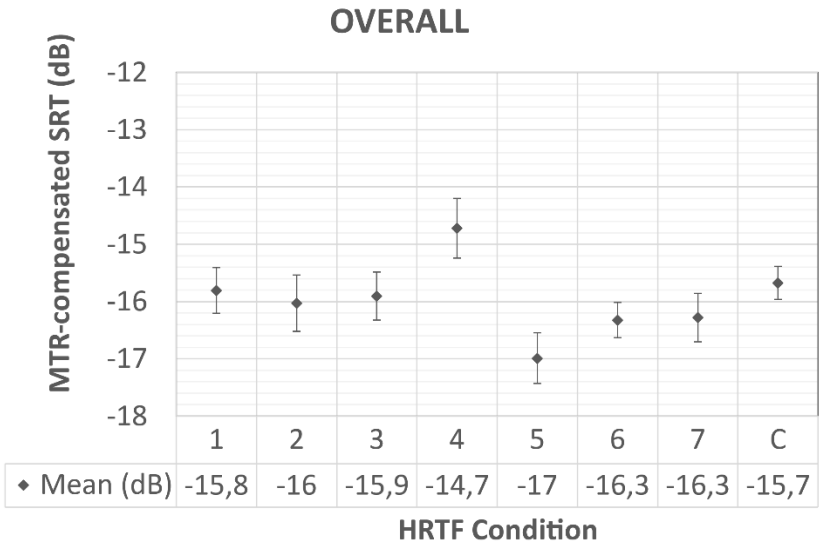


Figure 77. $\overline{SRT^*}_{H_j}$ and 95% CI of the MTR-compensated SRT for each HRTF condition (overall study).

We can see how the $HRTF_5$ seems to be the *best HRTF* when all the data is analysed together and with a $\overline{SRT^*}_{H_5}$ very different to the others. In the same way, $\overline{SRT^*}_{H_4}$ stands out from the rest, resulting the *worst HRTF* in overall.

- **Statistical analysis**

A one-way ANOVA was carried out, showing a significant impact of HRTF on SRT when the $HRTF_C$ was included ($F(7, 168) = 10.1834, < 0.001$ ***), and also when it was removed from the data set ($F(6, 147) = 10.6762, < 0.001$ ***).

Then, a post-hoc pairwise comparison was carried out using both LSD and Bonferroni adjustments, with the idea of clearly report all statistical tests conducted during the

analysis and also apply some corrections to help in the interpretation. The results using LSD comparisons are shown in Table 19. This pairwise comparison confirms what it was shown in Figure 76, where $HRTF_4$ and $HRTF_5$ presented significant differences with all other conditions. In addition, this comparison shows significant differences between the $HRTF_C$ and $HRTF_6$ and $HRTF_7$ but, considering that any corrections have been applied, some of these differences may happen by chance. If we apply Bonferroni correction, Table 20 results confirms that even after the corrections, the $HRTF_4$ and $HRTF_5$ have significative differences with all other conditions. In this case, $HRTF_5$ seems to be the *best HRTF* for the majority of the participants, and $HRTF_4$ the *worst*.

Table 19. Post-hoc simple pairwise comparison using LSD for overall analysis for MTR-compensated SRT data

	$HRTF_2$	**$HRTF_3$**	**$HRTF_4$**	**$HRTF_5$**	**$HRTF_6$**	**$HRTF_7$**	**$HRTF_C$**
$HRTF_1$	0.440	0.732	< 0.001***	< 0.001***	0.074	0.102	0.646
$HRTF_2$		0.667	< 0.001***	0.001**	0.306	0.384	0.219
$HRTF_3$			< 0.001***	< 0.001***	0.147	0.194	0.424
$HRTF_4$				< 0.001***	< 0.001***	< 0.001***	0.001**
$HRTF_5$					0.021*	0.014*	< 0.001***
$HRTF_6$						0.878	0.025*
$HRTF_7$							0.037*

Table 20. Post-hoc simple pairwise comparison using Bonferroni for overall analysis for MTR-compensated SRT data

	$HRTF_2$	**$HRTF_3$**	**$HRTF_4$**	**$HRTF_5$**	**$HRTF_6$**	**$HRTF_7$**	**$HRTF_C$**
$HRTF_1$	1	1	0.006**	0.002**	1	1	1
$HRTF_2$		1	< 0.001***	0.028*	1	1	1
$HRTF_3$			0.002**	0.006**	1	1	1
$HRTF_4$				< 0.001***	< 0.001***	< 0.001***	0.031*
$HRTF_5$					0.601	0.401	< 0.001***
$HRTF_6$						1	0.705
$HRTF_7$							1

5.5.2.3 Discussion

With this compensation, we intended to eliminate the advantage or disadvantage of a HRTF conditions for a particular target and masker configuration, which is inherent in the HRTFs and therefore not dependent on individual differences of the participants.

The post-compensation analyses results suggest that, when the effect of MTR is removed, the impact of the various HRTFs on SRT is still significant for many

participants (50% of the participants) and different participants behave differently for the different HRTFs in a very spread manner, as shown in the pair-wise comparison.

However, this compensation has altered the SRT values for some specific cases, as the $HRTF_4$ and $HRTF_5$. $HRTF_5$ now seems to be the *best HRTF* for most participants, although before compensation it was considered as the *worst HRTF* in overall. On the other hand, the $HRTF_4$ has now become the *worst* for many of the participants, and the $HRTF_C$, despite being a control condition that has no spectral cues, is not so different from the rest.

This compensation has shown that the results can be greatly altered by the gains offered by HRTFs and consequently it is something very important to consider. However, this compensation may have some limitations, since it is based on a simplistic binaural loudness summation assumption, and only considers power ratios. There are other characteristics of HRTFs that can be altering the results, for example, Bronkhorst et al (A. W. Bronkhorst & Plomp, 1988) demonstrated that the gain due to ILD relies on the ears presented with the most favourable signal-to-noise ratios, which is known as the best-ear benefit.

The auditory models presented in the state of the art offers prediction regarding the binaural advantage for speech when it is being interfering by noises. More specifically, Jelfs' model (Jelfs et al., 2011) allows us to know the benefit provided by the better-ear and binaural unmasking for each HRTF. Therefore, the next step of our study was to analyse the data by adjusting again the raw values, but this time using the results of the model to achieve a more accurate compensation.

5.5.3 Data compensated by SRM from Jelfs auditory model

Jelfs et al. presented a model (Jelfs et al., 2011) to get the Spatial Release from Masking (SMR) in a Cocktail Party situation, predicting the increase in the target speech intelligibility when target and maskers are spatially separated. The model is based on the one presented by Lavandier & Culling (2010) and divides the SRM into two contributions: the binaural advantage due to binaural unmasking (Binaural Masking Level Difference - BMLD) and the benefit of the better-ear listening. This speech perception model is included in the Matlab Auditory Toolbox (*The Auditory Modeling Toolbox*, n.d.) as *JELFS2011 - Predicted binaural advantage for speech in reverberant conditions.*

This model combines noise and reverberation, using BRIRs. In this way, target and maskers are represented in the model by the impulse responses. Since our study was carried out in anechoic conditions, we have used as model input parameters the HRIR for the target and masker positions. The model offers three different outputs, the total benefit of the spatial release from masking (SRM) in dB, the weighted_SNR in dB, which is the component of SRM due to better-ear listening and the weighted_BMLD in dB, which is the component of SMR due to binaural unmasking. The total benefit is the sum of the weighted_SNR and the weighted_BMLD. The values of these three parameters for each HRTF condition are shown in Table 21.

Table 21. Spatial Release from Masking (SRM) in dB from the Jelfs model (Jelfs et al., 2011)

HRTF condition	**$HRTF_1$**	**$HRTF_2$**	**$HRTF_3$**	**$HRTF_4$**	**$HRTF_5$**	**$HRTF_6$**	**$HRTF_7$**	***$HRTF_C$***
SRM_{H_j} (dB)	2,56	2,42	2,02	2,22	1,27	2,11	2,21	1,29
$weighted\_SNR_{H_j}$(dB)	0,00	-0,11	-0,53	-0,35	-1,32	-0,47	-0,37	-1,82
$weighted\_BMLD_{H_j}$(dB)	2,56	2,53	2,56	2,57	2,59	2,58	2,58	3,11

The idea was to use the total benefit (SRM_{H_j}) to do a new compensation of the obtained raw SRT. But, in this case, the SRM_{H_j} is calculated with respected to one target and two maskers. Since our SRT was defined as the signal to noise ratio between one target and one masker, we reduced 3dB the SRM value to compensate the SRT and to be able to compare the two different compensation approaches used. The SRM for **one** masker is shown in Table 22.

Table 22. Spatial Release from Masking (SRM) in dB from the Jelfs model (Jelfs et al., 2011) considering the relation between **one** target and **one** masker.

HRTF condition	$HRTF_1$	$HRTF_2$	$HRTF_3$	$HRTF_4$	$HRTF_5$	$HRTF_6$	$HRTF_7$	$HRTF_C$
$\boldsymbol{SMR^1_{H_j}}$	-0,44	-0,58	-0,98	-0,78	-1,73	-0,89	-0,79	-1,71

These ratios were used to adjust the obtained raw SRT for each condition and participant, following the next formula, where P_i is the participant, with $i \in [1, 22]$ and H_j the HRTF condition, with $j \in [1, 7]$ *and* C

$$SRT\_compensated\ SRT^{P_i}_{H_j}\ (dB) = raw\ SRT^{P_i}_{H_j}\ (dB) + SRM_{H_j}\ (dB) \qquad (5.6)$$

Figure 77 shows a comparison between the two used compensation factors. Light grey shows the ratios used to compensate the raw data in previous section, the ($MTR_{H_j}\ (dB)$). Dark grey shows the new ratios calculated in this section, using Jelfs' model ($SMR^1_{H_j}\ (dB)$). Both ratios follow the same trend but note that the larger

differences occur for conditions $HRTF_4$ and $HRTF_5$. This makes us think that the compensation based on the power ratio was not misguided, but there was something that was missing and made the $HRTF_4$ and $HRTF_5$ very different from the rest. We will now analyse how the data have been modified and results have changed with this new compensation.

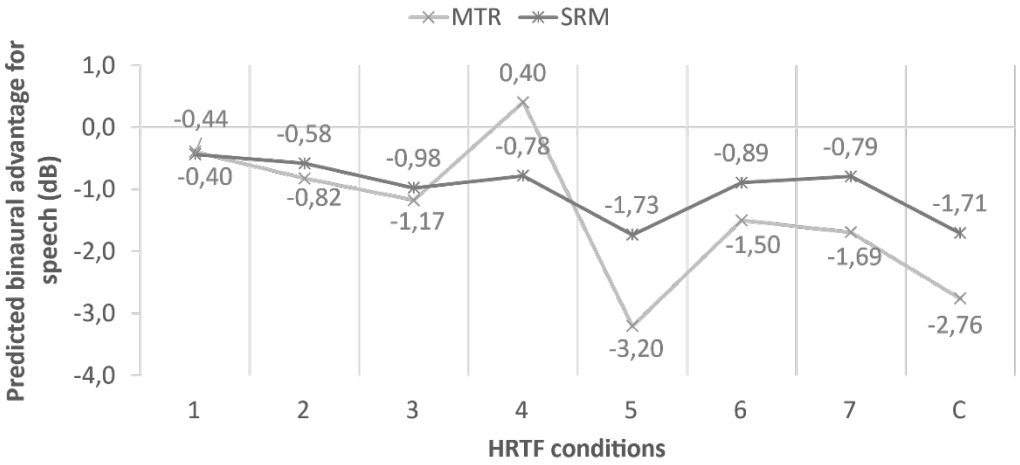


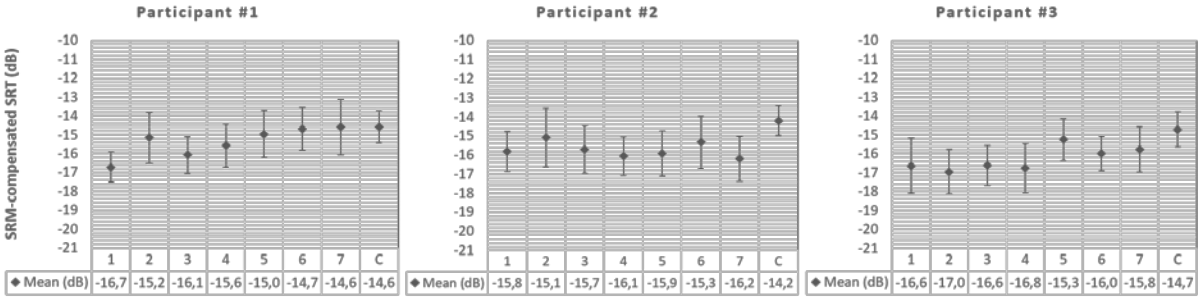
Figure 78. Factors used to compensate SRT values

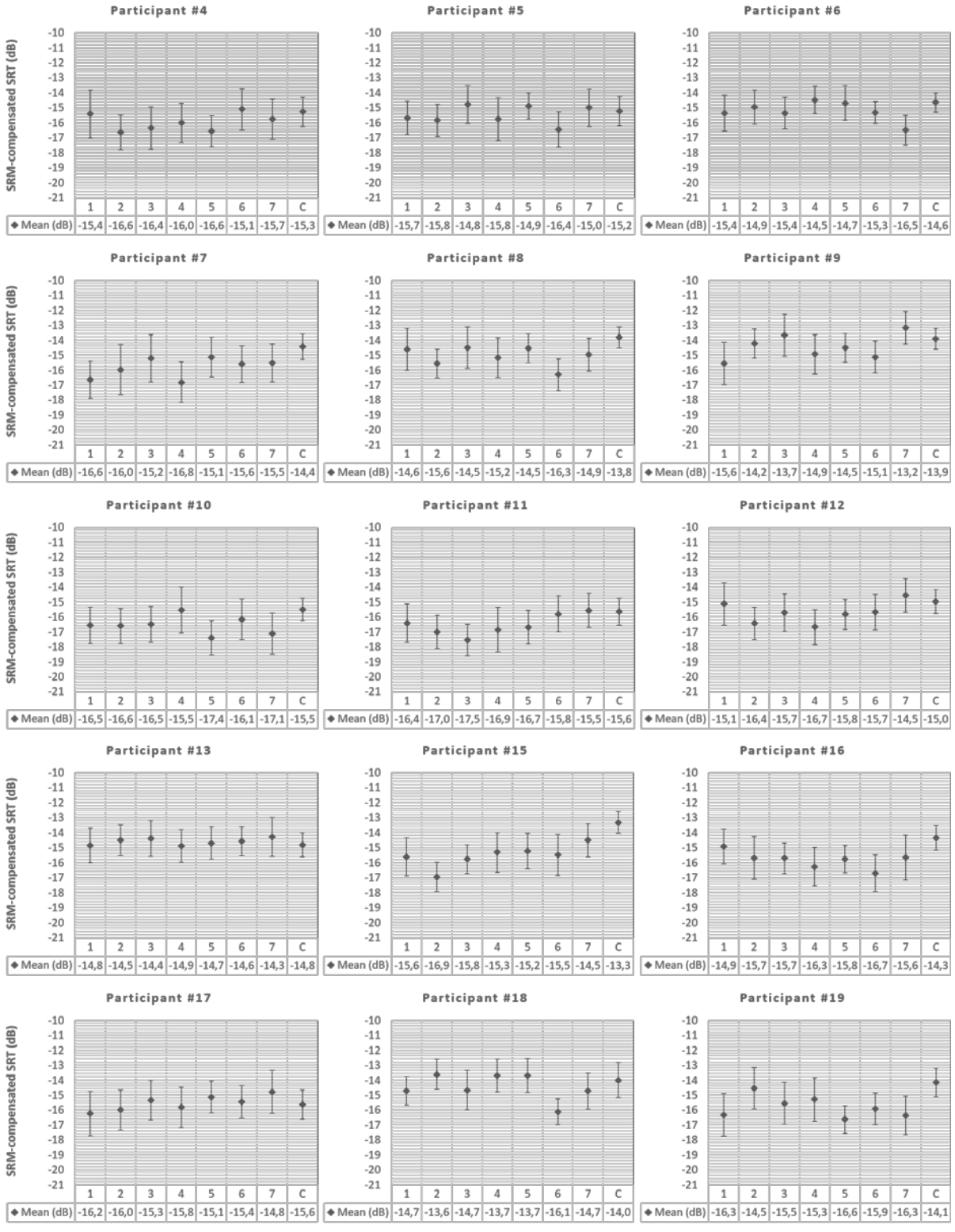
5.5.3.1 Individual analysis

The same analysis of the previous sections has been carried out here using this $SRT\_compensated\ SRT_{H_j}^{P_i}$ data.

- **Collected data**

Again, an individual analysis of the data was carried out. The SMR-compensated SRT mean for each HRTF condition ($\overline{SRT^{**}}_{H_j}^{P_i}$) is shown in Figure 78 as well as the 95% Confidence Interval (CI). The horizontal axis shows the HRTF condition and $\overline{SRT^{**}}_{H_j}^{P_i}$ in decibels. The vertical axis indicates the SMR-compensated SRT value in decibels. The title of each graph indicates the participant ID.





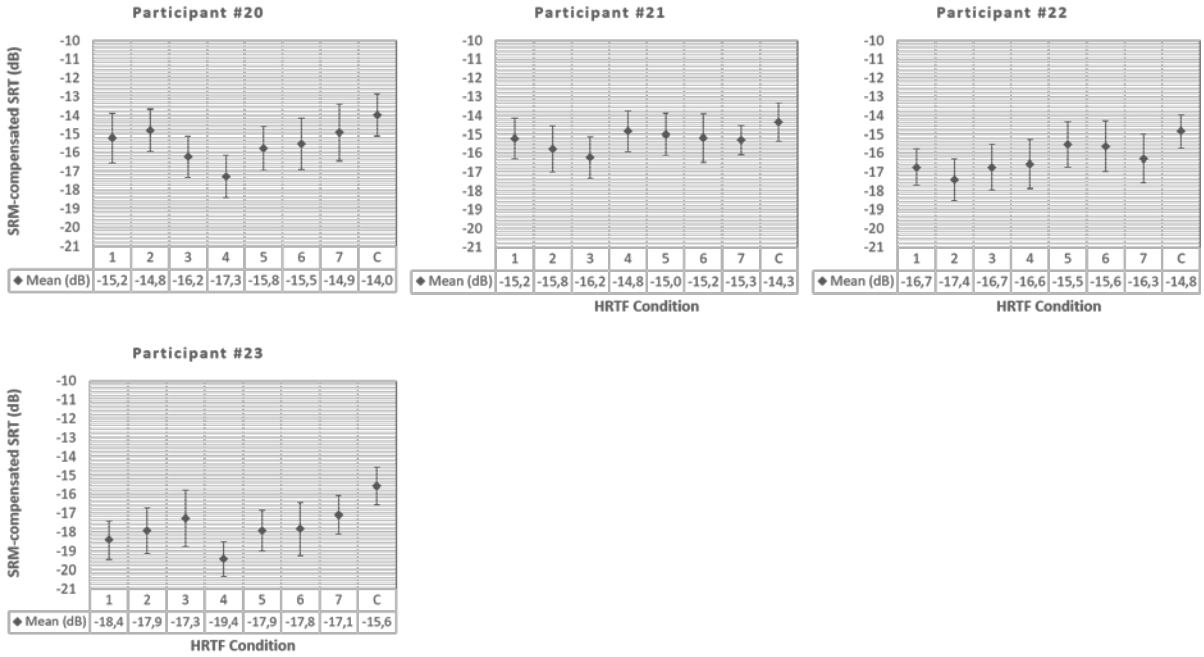


Figure 79. $\overline{SRT^{**}}_{H_j}^{P_i}$ and 95% CI of the SMR-compensated SRT for each HRTF condition

The 95% CI remains the same since the compensation affects in the same way to all the SRT values for a specific HRTF condition. However, the $\overline{SRT^{**}}_{H_j}^{P_i}$ is now different for each H_j and P_i. This compensation reduces the differences among the $\overline{SRT^{**}}_{H_j}^{P_i}$, as can be seen if we compare specific cases, as participant #1 in Figure 79. In this case, after the compensation, the $HRTF_1$ is not significantly different to the others anymore, since its SRT mean is within the rest of the CIs.

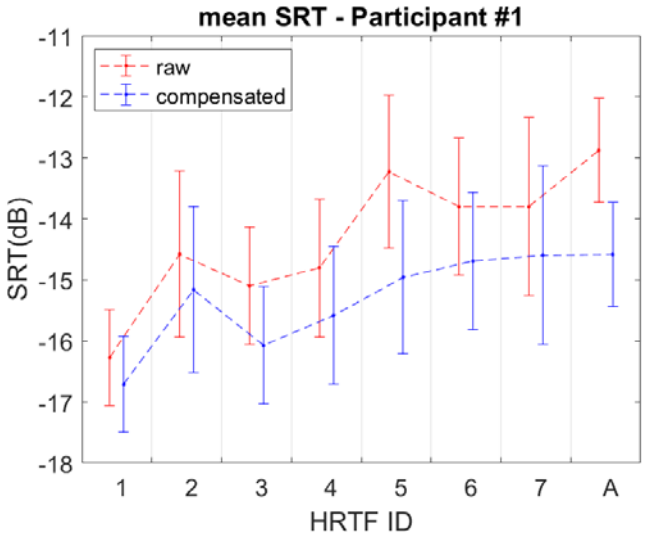


Figure 80. Mean and 95% CI of the raw (diamonds on the left) and SRM-compensated (circles on the right) SRT for each HRTF condition.

The *best* and *worst* measured HRTF (excluding the control condition, $HRTF_c$) after the compensation are shown in Table 23, together with the $\overline{SRT^{**}}^{P_i}_{H_j}$ value. In addition, last column shows the $\overline{SRT^{**}}^{P_i}$ *difference* in dB between the *best* and the *worst HRTF* condition. Being the smallest difference in absolute terms 0.61 dB ($\overline{SRT}^{P_{13}}$ *diff*) and the largest difference 2.48 dB ($\overline{SRT}^{P_{18}}$ *diff*), numbers that are still comparable with the ranges found in previous studies when looking at BMLD and at the impact of interaural differences on SRM (e.g. Culling et al. (2004))

Table 23. SRM-compensated SRT values for the best and worst measured HRTF

	Best HRTF$_{1\text{-}7}$		**Worst HRTF$_{1\text{-}7}$**		
Participant ID (i)	$\overline{SRT^{**}}^{P_i}_{H_{best}}$ **(dB)**	**HRTF condition (j)**	$\overline{SRT^{**}}^{P_i}_{H_{worst}}$ **(dB)**	**HRTF condition (j)**	$\overline{SRT^{**}}^{P_i}$ ***diff*** **(dB)**
#1	-13,71	1	-11,59	7	-2,12
#2	-13,19	7	-12,08	2	-1,11
#3	-13,96	2	-12,26	5	-1,7
#4	-13,63	2	-12,09	6	-1,54
#5	-13,44	6	-11,78	3	-1,66
#6	-13,49	7	-11,48	4	-2,01
#7	-13,80	4	-12,13	5	-1,67
#8	-13,29	6	-11,48	3	-1,81
#9	-12,56	1	-10,17	7	-2,39
#10	-14,38	5	-12,53	4	-1,85
#11	-14,53	3	-12,54	7	-1,99
#12	-13,65	4	-11,54	7	-2,11
#13	-11,88	4	-11,27	7	-0,61
#15	-13,93	2	-11,49	7	-2,44
#16	-13,69	6	-11,89	1	-1,81
#17	-13,21	1	-11,77	7	-1,44
#18	-13,09	6	-10,61	2	-2,48
#19	-13,61	5	-11,53	2	-2,07
#20	-14,25	4	-11,81	2	-2,45
#21	-13,23	3	-11,83	4	-1,4
#22	-14,41	2	-12,51	5	-1,9
#23	-16,40	4	-14,07	7	-2,34

The *best* and *worst HRTF* are now widely distributed along all participants, as can be seen in Table 24, which shows for which percentage of participants each HRTF conditions is the best or the worst. This time, contrary to previous analysis, none of the

HRTF are better or worse for more than 50% of participants. Here, the *best HRTF* is the $HRTF_4$ for just five participants followed by the $HRTF_2$ and $HRTF_6$ for four participants each. The *worst HRTF* is the $HRTF_7$ for eight participants, followed by the $HRTF_2$ which is the worst for four participants. The fact that the $HRTF_2$ is best for some participants and $worst$ for others is a good evidence that HRTF impacts on SRT differently for different users, as we argued in Hypothesis 2.

Table 24. Percentage of participants where each HRTF conditions is the best or the worst

H_j	% of participants where $H_j$is the best	% of participants where $H_j$is the worst
$HRTF_1$	13%	4%
$HRTF_2$	18%	18%
$HRTF_3$	9%	9%
$HRTF_4$	22%	13%
$HRTF_5$	9%	13%
$HRTF_6$	18%	4%
$HRTF_7$	9%	35%

- **Statistical Analysis**

A final one-way ANOVA was carried out with the SRM-compensated SRT as the dependent variable and the HRTF as the independent one. Table 25 shows the results for the statistical analysis. The first column indicates the participant ID, columns 2 and 3 indicate the ANOVA when all conditions are included ($HRTF_{1-7} + HRTF_C$), and columns 4 and 5 indicate the ANOVA when only the measured HRTFs conditions ($HRTF_{1-7}$) are included.

If the eight conditions ($HRTF_{1-7} + HRTF_C$) are included in the analysis, 5 out of 22 present statistical differences between different HRTF conditions. When the $HRTF_C$ is excluded, just one participant presents significant differences. As we saw in the graphs showing the $\overline{SRT^{**}}^{P_i}_{H_j}$, the values have come closer together, and therefore it is now more difficult to find significant differences between the different HRTFs.

To study which HRTFs are more different to each other, a post-hoc simple pairwise comparison using the LSD was again carried out and results are shown Table 26. In addition, the table shows underlined the participants who present significant differences after Bonferroni corrections. Participants #4, #13, #17 do not have significant differences for any HRTF pair.

Table 25. ANOVA results for compensated data using the SRM from Jelfs model

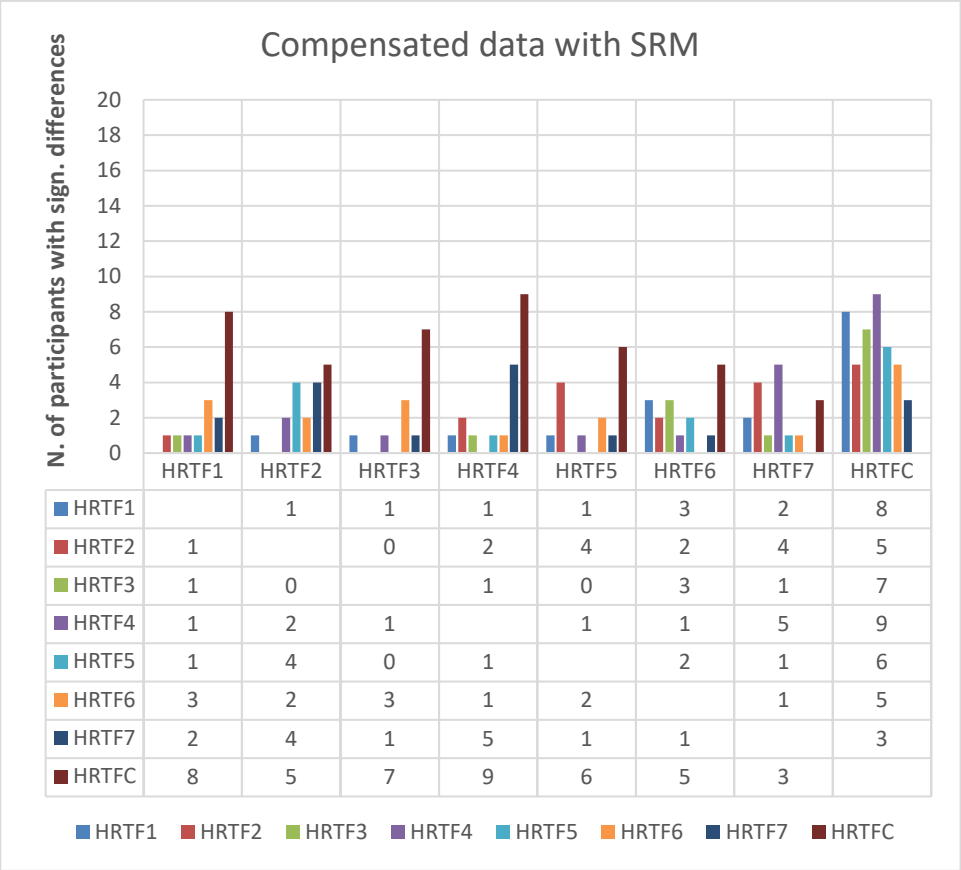
	$HRTF_{1\text{-}7}$ + $HRTF_C$		**$HRTF_{1\text{-}7}$**	
ID	**$F_{7,152}$**	**p-value**	**$F_{6,133}$**	**p-value**
#1	1.78	0.095	1.68	0.130
#2	1.16	0.331	0.39	0.882
#3	1.84	0.084	1.06	0.389
#4	0.83	0.565	0.75	0.610
#5	0.97	0.453	1.06	0.387
#6	1.60	0.138	1.54	0.169
#7	1.36	0.224	0.89	0.505
#8	1.70	0.113	1.18	0.321
#9	2.10	0.047*	2.16	0.051
#10	1.13	0.345	0.84	0.540
#11	1.42	0.200	1.25	0.285
#12	1.51	0.169	1.42	0.212
#13	0.17	0.991	0.16	0.986
#15	3.27	0.003**	1.51	0.181
#16	1.46	0.186	0.80	0.573
#17	0.53	0.809	0.59	0.741
#18	2.23	0.034*	2.55	0.023*
#19	1.96	0.064	1.23	0.295
#20	2.47	0.020*	1.77	0.109
#21	1.05	0.399	0.72	0.632
#22	1.91	0.071	1.16	0.330
#23	3.54	0.001**	1.66	0.135

In addition, Figure 80 shows the number of participants with significant differences in each pair-wise comparison ($p < 0.05$). The table within the figure indicates the number of participants with significant differences between the HRTF condition indicated in the header and one in the leftmost column. The graph indicates the number of participants with significant differences between the HRTF condition indicated in the horizontal axis and the one corresponding with the colour in the legend.

Table 26. Post-hoc simple pairwise comparison for individual analysis of the SRM-compensated SRT data.

HRTF condition	$HRTF_2$	$HRTF_3$	$HRTF_4$	$HRTF_5$	$HRTF_6$	$HRTF_7$	$HRTF_C$
$HRTF_1$	#19	#09	#20	#01	#01 #08 #16	#01 #09	#01 #03 #07 #09 #15 #19 #22 **#23**
$HRTF_2$			#15 #20	#03 #15 #19 #22	#18 #22	#06 #12 #15 #19	#03 #08 **#15** #22 #23
$HRTF_3$			#23		#05 #08 #11	#11	#03 #11 #15 #20 #21 #22 #23
$HRTF_4$				#10	#18	#06 #09 #12 #20 #23	#02 #03 #07 #12 #15 #16 **#20** #22 **#23**
$HRTF_5$					#08 #18	#06	#02 #10 #15 #19 #20 #23
$HRTF_6$						#09	#08 #15 #16 #18 #23
$HRTF_7$							#02 #06 #19

As in previous analysis, using LSD makes that some significant comparisons can appear by chance. In this case, having a total of 616 comparisons (28 comparisons x 22 participants), we can consider that we would have approximately one false positive per pair comparisons. Considering just the measured HRTFs, looking at the table we have obtained 37 pairs with significant differences, so there are more than the ones expected by pure chance (the probability tells us that it would be expected that 23 pairs are significant by chance, 22 participants x 21 pairs x 0.05). This is an evidence of the HRTF influence on the SRT but does not allow us to argue specific cases. However, the $HRTF_C$ presents significant differences with all other conditions, even after Bonferroni corrections. This reinforces the fact that the $HRTF_C$ carries out worse performances than the measured HRTFs, even after the compensation.



	HRTF1	HRTF2	HRTF3	HRTF4	HRTF5	HRTF6	HRTF7	HRTFC
HRTF1		1	1	1	1	3	2	8
HRTF2	1		0	2	4	2	4	5
HRTF3	1	0		1	0	3	1	7
HRTF4	1	2	1		1	1	5	9
HRTF5	1	4	0	1		2	1	6
HRTF6	3	2	3	1	2		1	5
HRTF7	2	4	1	5	1	1		3
HRTFC	8	5	7	9	6	5	3	

Figure 81. Post-hoc pairwise comparisons using LSD of the Compensated data with SRM. Vertical axis indicates the number of participants with significant differences in the pair-wise comparison between the HRTF ID indicated by the colour and the horizontal axis. In addition, this information is also shown in the table below the graph.

5.5.3.2 Overall analysis

- **Collected data**

We repeat the overall analysis this time using the $\overline{SRT^{**}}_{H_j}^{P_i}$ data and calculating the mean ($\overline{SRT^{**}}_{H_j}$) and 95 %CIs. Results are shown in Figure 81. The horizontal axis shows the HRTF Condition and the $\overline{SRT^{**}}_{H_j}$ value, in decibels, for each condition. The vertical axis indicates the SRT value in decibels. Even after the compensation, these SRT mean values agree well with the ones presented by A. W. Bronkhorst & Plomp (1988).

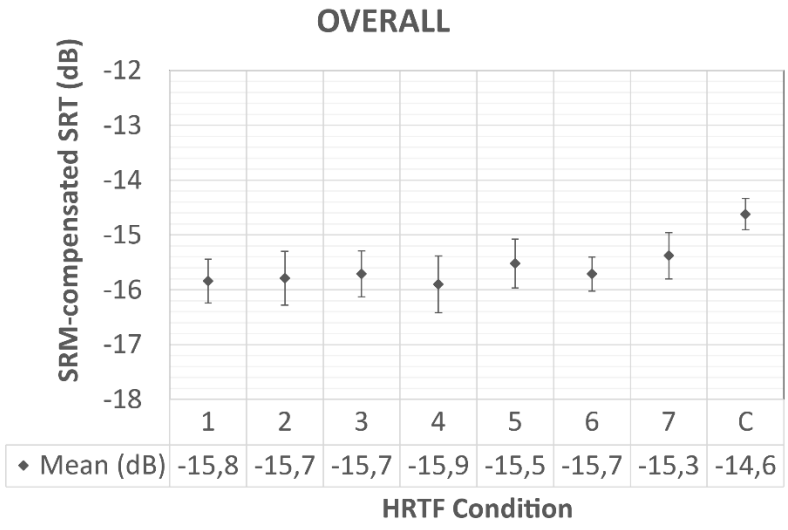


Figure 82. $\overline{SRT^{**}}_{H_j}$ and 95% CI of the SRT for each HRTF condition for the overall study with the compensated SRT using SRM compensation.

In this case, when we analyse all the participants together, we can see how all the $\overline{SRT^{**}}_{H_{1-7}}$ have become more similar for the measured HRTFs, now the range of [-15.9, -15.3]. In addition, $\overline{SRT^{**}}_{H_C}$ had also come close to the other means, but still outside the other CIs, so it seems that, overall, the $HRTF_C$ is the *worst HRTF*. But within the HRTF measures we do not see clearly that there is a better or worse for all participants in general.

- **Statistical analysis**

A one-way ANOVA was carried out, showing a significant impact of HRTF on SRT when the $HRTF_C$ was included (F (7,168) = 4.1892, p<0.001) but not when it was removed from the data set (F (6,147) = 0.76083, 0.602). The post-hoc comparison using LSD and Bonferroni corrections are shown in is shown in Table 27 and Table 28 respectively.

Table 27. Post-hoc simple pairwise comparison using LSD for overall analysis for MTR-compensated SRT data

	$HRTF_2$	**$HRTF_3$**	**$HRTF_4$**	**$HRTF_5$**	**$HRTF_6$**	**$HRTF_7$**	**$HRTF_C$**
$HRTF_1$	0.859	0.649	0.839	0.267	0.657	0.110	< 0.001***
$HRTF_2$		0.781	0.703	0.351	0.791	0.155	< 0.001***
$HRTF_3$			0.510	0.511	0.990	0.251	< 0.001***
$HRTF_4$				0.189	0.518	0.072	< 0.001***
$HRTF_5$					0.504	0.623	0.002**
$HRTF_6$						0.246	< 0.001***
$HRTF_7$							0.009**

Table 28. Post-hoc simple pairwise comparison using Bonferroni for overall analysis for MTR compensated SRT data

	$HRTF_2$	$HRTF_3$	$HRTF_4$	$HRTF_5$	$HRTF_6$	$HRTF_7$	$HRTF_C$
$HRTF_1$	1	1	1	1	1	1	< 0.001***
$HRTF_2$		1	1	1	1	1	0.002**
$HRTF_3$			1	1	1	1	0.006**
$HRTF_4$				1	1	1	< 0.001***
$HRTF_5$					1	1	0.057
$HRTF_6$						1	0.006**
$HRTF_7$							0.253

The post-hoc analysis corroborates the above mentioned in the Collected Data subsection. In both, before and after Bonferroni correction, none of the HRTFs show significant differences with others except the $HRTF_C$. In addition, this allows us to say that, after this SRT adjustment, there is not a universally better or worse measured HRTF for all participants, corroborating Hypothesis 2.

5.5.3.3 Discussion

The Jelfs model was used to estimate the HRTF-specific benefit, and adjustments have been carried out on the SRTs for every HRTF. Larger compensations were needed for $HRTF_5$ and $HRTF_C$ (see Table 21). Looking specifically at this result, it is evident that the model predicted very well the observed data, with a correlation coefficient of 0.9547 (p=0.0008) when comparing across measured HRTF conditions. These correlation coefficients are in line with the ones obtained by Jelfs et al. (2011), where they compare their predictions with results from previous studies. The current study can be considered as a further validation of the Jelfs model, extending its use (and validity) to the comparison of SRT outcomes between different HRTFs (while previous comparisons focused mainly on different acoustic environments and source/receiver configurations).

Applying this second compensation allows us to compare the performance of each participant with respect to the individual differences in the spectral cues of each HRTF conditions, since we have extracted the benefit provided by the better-ear and BMLD. After eliminating all effects contemplated in the models that would be universal, there are still 5 participants out of 20 who continue to show differences between HRTFs. In addition, the pairwise comparison, considering the measured HRTFs, shows that there are significant differences between some HRTF pairs, suggesting that those participants have indeed found an effect on these HRTFs, that can be attributed to individual characteristics, which is compatible with what we proposed in the Hypothesis 1. Regarding the $HRTF_C$, which is a synthetic HRTF used as a control condition, after the

compensation it has more significant differences in the pairwise comparison and presents the worst SRT performance, which is a good evidence to think that the spectral cues of the HRTF give advantages in the Cocktail Party situation, since the $HRTF_C$ (obtained with a spherical-head model with no pinnae) does not contain those cues.

The fact that the *best* and the *worst HRTF* is not the same for all the participants after the compensation, and that the results of the overall analysis show no significant differences when the $HRTF_C$ was excluded, it is a good evidence to affirm that there is not an universally *best* or *worst HRTF* for all participants, following what we suggested in the Hypothesis 2 of our study.

5.5.4 Further analysis of some specific cases

One further element outlined by the observation of the collected data is the high variance of the various SRT values across participants, HRTF conditions and sessions. Considering the same HRTF, participants had variations of up to 10-12dB SRT between different sessions. This is evident, for example, when looking at the data from some participants. Figure 82 shows the boxplot for some specific participants, participants #15 and #23 present significant differences in the ANOVA and participants #4 and #17 do not. Participants without significant differences show high variance in the SRT data distribution for all the HRTFs, like participant #17. However, for example, participant #23 shows a very low variance for his/her *best HRTF* ($HRTF_4$).

In addition, the overall variability of the SRTs along the sessions for each individual is also evident, as well as the relative ranking between HRTFs for each session, as it can be seen in Figure 83. The figure shows the HRTF ranking for each session for four specific participants, where participants #15 and #23 present significant differences in the ANOVA and participants #4 and #17 do not. For participants #15 and #23, we can identify clearly a better and worse HRTF and a certain level of repeatability can be found between the ranking of HRTFs in the different sessions. The *best HRTF* is in the first positions of the ranking for many sessions (bottom of the chart), while the *worst HRTF* is in the last positions (top of the chart). $HRTF_C$ can also be found in the last position of the ranking for many participants along many sessions. For participants #4 and #17 the variability is much higher between the different sessions, resulting in the differences between HRTFs being not significant. These results seem to be in line with the ones from previous research looking at the repeatability of qualitative HRTF rating (Andreopoulou & Katz, 2016a), where only a certain number of participants, which were categorised as "expert assessors", were able to rate a certain number of HRTFs repeatable across different sessions. Given that no clear pattern of evolution is perceived throughout the sessions, the selection of the *best* or *worst HRTF* could be clarified as increasing the number of sessions.

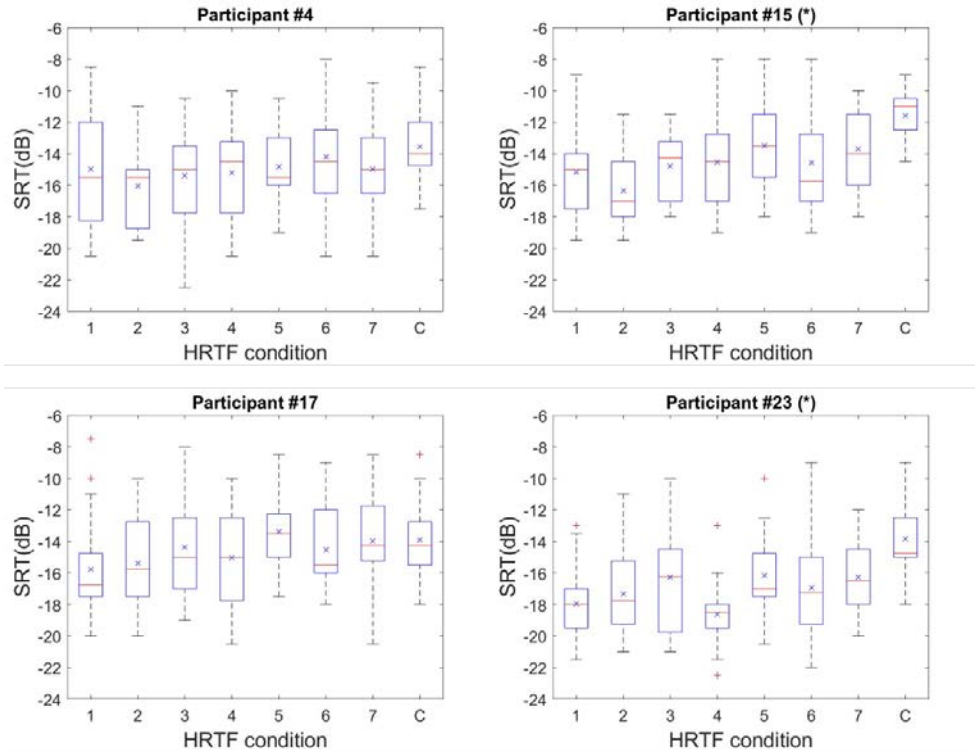


Figure 83. Boxplot of the collected SRT data for four specific participants. On each box, the central horizontal mark indicates the median, the cross mark the mean and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, which are plotted individually using the '+' symbol. Starts in the title of each graph (*) indicate participants with significant differences in the ANOVA.

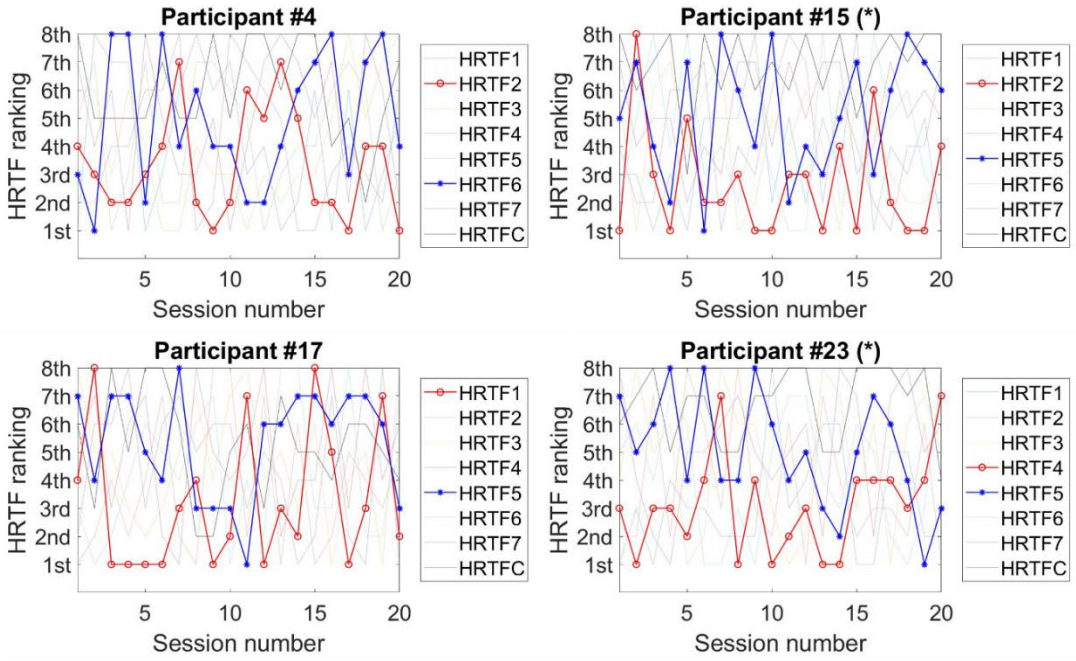


Figure 84. HRTF ranking along all the session for four specific participants. Stars behind the participant ID indicates that this participant presents significant differences in the ANOVA.

The red line with circle markers indicates the *best HRTF* and blue line with star markers indicates the *worst HRTF*. $HRTF_C$ is shown in grey colour and a bit thicker than the rest of HRTFs.

5.5.5 Learning effect

An analysis of the participant's performance improvements across sessions was carried out. Considering the large duration of this study, where each participant went through 20 different sessions, with an average of 11 minutes per session, an study of the relationship between the session number and the SRT values have been done.

A linear regression model was calculated to predict the SRT as a function of the session number and to study the learnability of the study. We collected a total of 3520 SRTs (22 participants x 20 sessions x 8 HRTF conditions) in the whole study. All of these SRT have been plotted in Figure 84 by session. The effect of the session number was found to be significant ($F(1, 3518) = 103.813), p < 0.001$ ***), showing a slightly decreasing level of SRT across sessions, with a slope of the regression line of 0.084. However, only 2.87% of SRT variance was explained by the session number (R^2 of 0.0287). Therefore, we can consider that, even though it is significant, the effect of learning is minimal.

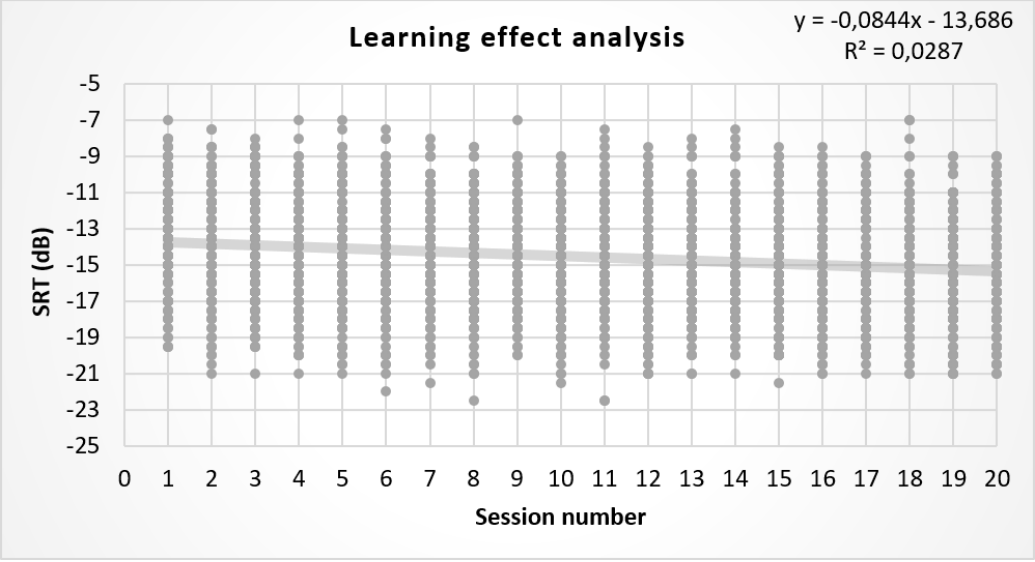


Figure 85. Dispersion graph and lineal regression analysis for every collected SRT data. On the top-right corner the regression line expression and the R^2 value are shown.

To discard any participant having an especially important learning and to see if the behaviour throughout the sessions is consistent for all participants, the same analysis was performed within each participant, calculating an independent linear regression

model for each. The coefficient of determination (R squared) and the significance of the models are shown in Table 29. The effect of the session number on the SRT was found to be significant for 12 out of 22 participants but, again, the percentage of SRT variance explained by the model was always small, going from a minimum of 0.1% for participants #6 and #21 to a maximum of 11.1% for participant #18.

Table 29. Learning effect analysis for each participant

ID	R^2	F(7,158)	p-value
#1	0.061	10.29	0.001**
#2	0.058	9.87	0.002**
#3	0.003	0.48	0.486
#4	0.038	6.38	0.012*
#5	0.046	7.63	0.006**
#6	0.001	0.19	0.662
#7	0.019	3.08	0.081
#8	0.044	7.28	0.007**
#9	0.018	2.99	0.085
#10	0.076	13.01	<0.001***
#11	0.070	12.01	<0.001***
#12	0.006	0.95	0.329
#13	0.052	8.75	0.003**
#15	0.011	1.78	0.182
#16	0.064	10.86	0.001**
#17	0.018	3.01	0.084
#18	0.111	19.73	<0.001***
#19	0.021	3.46	0.064
#20	0.039	6.54	0.011*
#21	0.001	0.28	0.595
#22	0.032	5.23	0.023*
#23	0.014	2.37	0.125

It is known that repeated exposure to a given task can result in a certain amount of improvement due to procedural and stimulus learning, and more specifically, contribute to improvements in performance on ITD discrimination (Ortiz & Wright, 2009). This has been observed also in speech-related auditory tasks (Fu & Galvin, 2003). Considering that feedback is an essential mechanism for both procedural and perceptual learning (Ortiz & Wright, 2009), it should be noted that no feedback was given to the participant during the experiment, so they didn't know if they correctly identified the target words

in each trial, during the whole experiment. In addition, it should be considered that participants did not carry out any training at the beginning of the experiment. For all these reasons mentioned above, it is not surprising that a small effect of learning was found in the SRT data, which could be attributed to improvements in the participants understanding of the task, procedural learning or improvements in their ability to focus attention.

5.6 Conclusions

This chapter presents an original study where the impact of non-individualized HRTFs on speech intelligibility within a Cocktail Party situation was investigated. This experiment is an example of use of the 3DTI-Toolkit which has helped us to verify the good performance of the Toolkit. The library presented in previous chapters has allowed us to design a psychoacoustic experiment in a virtually rendered scenario. Thanks to the 3DTI-Toolkit feature that enables to load different HRTFs, we have carried out the comparison of multiple HRTFs using the same sound source configuration and scenario in real-time.

Regarding the experiment, the findings reveal that there is an impact of the HRTF choice on understanding speech against noise. This impact is larger and more statistically significant if we analyse the raw data. From this analysis we corroborated the initial Hypothesis 1, since in the pairwise comparison we found many participants with significant differences between the HRTFs, which is a good evidence to affirm that different HRTFs provide different performances for speech intelligibility.

The overall analysis of the raw data revealed that there was an HRTF (the $HRTF_5$), which turns out to be the worst for most participants. This fact made us think that some HRTF were specially affecting the sound sources with the chosen spatial configuration (target presented frontally and maskers from both sides). Thus, we focused out attention on the spectral cues of each HRTF, trying to avoid the possible effect of the attenuation/amplification of some HRTFs. We compensated the SRT values with the SMR provided by the Jelfs model. After the compensation, the results showed different performances for the different HRTF conditions, which suggest that spectral cues of the HRTF have an effect on how a listener can discriminate speech in noise conditions. It should be noted that these differences were not as significant as with the raw data. However, the overall analysis confirmed Hypothesis 2, since, in this case and after adjusting the SRT, eliminating the part that comes from a common benefit to all participants, none of the HRTFs was universally better or worse for many participants.

The only condition that presents significant differences in the overall study before and after the compensation was the $HRTF_C$, which also presented the worst performance. This was expected since this HRTF does not presents spectral cues. This is consistent with many previous works where the benefit of binaural hearing in a cocktail party situation has been demonstrated (Hawley et al., 2004; Meesawat & Hammershøi, 2008).

One of the ideas that emerged at the beginning of this study was: can we use the designed procedure to create a mechanism to select, among several non-individualized HRTFs, the best matching HRTF for a specific listener? We have seen that there is an influence on the choice of the HRTF on the SRT. However, the results obtained in this study are not sufficient to validate the selection method, as we have not clearly obtained a *best HRTF* for each of the participants. To see if this is a good method to select the best matching HRTF, a study where the individual HRTF of each participant is included should be done. However, since the influence of the HRTF on the SRT has been demonstrated, it would be worth trying to repeat the experiment but including the individual HRTFs, although it is a very complicated procedure and requires a complex equipment, as it was seen in the state of the art in Chapter 2. In addition, further work can include the study of different databases of HRTFs, a larger number of participants and shorter procedure times, but with more repetitions.

Results of this experiment are very relevant, particularly to the scientific community in the area of binaural audio and spatial hearing, as they remark the importance of the HRTF choice when testing speech intelligibility in binaural rendered virtual environment. In addition, they provide relevant information to the research areas related to mathematical modelling speech-in-noise performances, considering that we have taken into account the output of existing models, more specifically the one from Jelfs et al. (2011). In addition, having using a control HRTF allows us to affirm that the HRTF spectral cues play an important role in speech intelligibility, which should be included in the models. Many auditory models suggested that the benefit that comes from the best ear and binaural unmasking are sufficient to explain performance in a Cocktail Party situation. However, our study has revealed that this performance can be affected by additional cues and an additional study would be very interesting in order to specify those cues.

Chapter 6

Conclusions

This chapter summarizes the main contributions of this thesis, the future work and the list of publications derived from this work. The structure of this chapter is as follows. Section 6.1. summarizes the main contributions of this thesis, presented throughout the whole document. Section 6.2 describes the work that has been done together with the Audio Experience Design Research Group at Imperial College London[28], with whom we have collaborated closely in many of the developments and studies presented in this thesis. The contributions of this work are currently being used by two research projects, which are described in Section 6.3. Then, Section 6.4 proposes some future work that is planned at the time of writing. Section 6.5 describes some contributions carried out during this thesis, which are closely related, but falls outside the scope of this work. Finally, Section 6.6 lists the publications which support this thesis and other publications of the author.

6.1 Contributions

This PhD thesis presents the development and evaluation of a binaural audio spatialisation library for real-time virtual environments, called the 3DTI Toolkit-BS. The motivation of this work came from the need of a tool that works as a platform to perform psychoacoustical experiments to study virtual auditory perception. This type of experiment requires a set of features, listed below, which are all met by the 3DTI Toolkit-BS and have not been found all together in other tools (as can be seen in Section 2.4, where a comparison with other tools is shown).

[28] https://www.axdesign.co.uk

Modularity. The architecture of the tool has been designed in a modular manner to provide the maximum possible flexibility. The library consists of multiple components, which together enable the generation of a complex 3D audio virtual scene. The detailed structure of the tool as well as all the implemented algorithms within each component are described in detail in Chapter 3. Each component has been implemented separately, which allows an easy substitution of any of them when a different algorithm or rendering technique is required. The fact that the anechoic and reverb paths are simulated with different components presents a great advantage, since it allows to simulate the anechoic path with high fidelity, and the reverb (which is one of the most computationally demanding processes) with techniques that sacrifice fidelity but improve performance. In addition, the object-oriented approach allows for adding multiple sources to a VAS in a very intuitive way.

Some of the previous mentioned components contain algorithms that have been implemented following the state of the art, but all together have not been found in other tools. Some of the most relevant components are listed below and all of them are explained in detail in Chapter 3.

- **HRTF, SOFA files and interpolation**. Binaural audio spatialisation using HRTFs makes 3D audio immersion possible for any listener using just a standard pair of headphones. The 3DTI Toolkit-BS allows to load the individual HRTF of the specific listener in SOFA format, which is a very used standard to store HRTF. Currently, many HRTF databases are available in SOFA format. In addition, the interpolation allows to simulate a directional sound coming from any location in the 3D space, even if this direction is not included in the loaded HRTF. The acoustical parallax, depending of the source distance, is also taken into account, as left and right HRIRs are selected independently, according to the relative angle between each ear and the sound source.
- **ITD customization.** The ITD is managed separately from the HRIR. In this way, the interpolation of the HRIR is carried out using aligned HRIRs, which reduces the comb filter effect that arises when adding impulse responses with similar magnitude but different phases. In addition, the ITD can be calculated as a barycentric interpolation of the ITD of the nearest positions or can be customized, re-computing its value according with the listener head circumference, as explained in Section 3.5.3.
- **Near field source simulation.** The Toolkit can also simulate the ILD adding an extra shadow in the contralateral ear for sources placed in a distance lower than 2 meters to the listener's head. This process is based on the frequency-domain solution for the diffraction of an acoustic wave by a rigid sphere, presented in Section 3.5.4.

- **Spatialized reverb simulation.** The 3DTI Toolkit-BS employs a virtual Ambisonic approach (presented in Section 3.6) based on a first order Ambisonic encoding, then decoded into a set of virtual speakers, and, finally, a convolution of those speakers as sources with the BRIRs of the specific position of the speakers. The Ambisonic approach allows to keep the location-dependent characteristics of the sources (but with low resolution, as it employs a first-order codification) and reduces the number of convolutions to the number of speakers. Another further improvement implemented in the 3DTI Toolkit, reduces the number of convolutions to the number of Ambisonic channels, which for FOA is four. This approach, together with an efficient convolution, implemented using the UPOLS algorithm presented in Section 3.4.1, allows the Toolkit to compute large reverberating scenes, with unlimited number sources in dynamic situations. In addition, this method allows for a certain amount of flexibility since the complexity of the rendering can be reduced or increased by modifying the Ambisonic order and the number of speaker used for the decoding.

Open source. Releasing a software as open-source allows clarity and reproducibility. It opens the door to new collaborations or external contributions to the tool, which will imply further enhancements and extensions. This feature, together with the modularity, makes the tool scalable, allowing, for example, the addition of more complex algorithms when the capability of the processor is increased. The 3DTI Toolkit is continuously maintained, with an active repository (https://github.com/3DTune-In/3dti_AudioToolkit) that also allows for issue reporting and tracking.

Multiplatform. The library has been developed in C++ and it is not implemented on top of any constrained hardware requirements, such as the presence of specific DSP technology for audio processing. This implies that it can be compiled on multiple platforms. It has been tested on Windows, MacOS and Linux. This feature, together with the fact that the library is open-source, allows other research groups to implement and modify the library to suit their needs.

Low latency performance. The latency of the system can be controlled by the modification of the frame size and frame rate. This allows the scene to be configured for having an audio rendering with no noticeable latency. For example, if we select a frame size of 256 samples, the library can anechoically spatialise 30 sources with an update rate of 5.8 milliseconds (see Section 4.6 where the real-time performance of the Toolkit is shown).

Good behaviour in dynamic situations. The 3DTI Toolkit-BS has been developed with special focus on having a good performance for scenarios with moving

sources and listener, where the renderer can dynamically update filter coefficients with real-time streaming of data with smooth transitions. This has been evaluated, estimating that the distortion produced by the library in dynamic situations is very low and without audible artefacts (see Section 4.5 of Chapter 4, where an objective evaluation shows the reduction of the artefacts).

Thanks to all these features, the 3DTI Toolkit-BS is a flexible, efficient and robust tool that can be used in multiple scenarios, as we review below.

6.1.1 The 3DTI Toolkit as a tool to perform psychoacoustical virtual experiments.

Already in 2000, Blauert et al. presented a tool for psychoacoustic research where they reflected on the importance of developing a platform that would allow the creation of complex auditory virtual environments with a good level of presence, and the generation of physiologically adequate signals to be delivered to the listeners. Since then, multiple tools for 3D audio generation have appeared, as discussed in Section 2.4, but none of them has been established as a "base tool" to be used in psychoacoustics labs. One of the main objectives of the 3DTI Toolkit-BS is to fill this gap and become a tool that serves as a base to support psychoacoustical experiments. In this way, the library has been developed in accordance with a set of requirements such as high degree of control, accuracy, realistic virtual 3D audio simulations, easy use and availability. These requirements are fulfilled thanks to all the features mentioned above, together with all the algorithms presented in the Chapter 3.

To validate this idea and to demonstrate and qualitatively test the performance of the 3DTI Toolkit-BS, this PhD thesis has presented a study about the impact of non-individualized HRTFs on speech intelligently. In this study, the Toolkit has been used as the audio rendering engine for the virtual psychoacoustic experiment. The experiment has been described in Chapter 4 and published in (Maria Cuevas-Rodriguez et al., 2021). In this study we have discovered some new aspects of the relationship between HRTFs and speech intelligibility and we have found evidence of the influence of HRTF in the spatial release from masking, more specifically in the horizontal plane. Currently we are using the 3DTI Toolkit-BS to perform a new experiment about the same topic, but this time using individual HRTFs and a source configuration in the vertical median plane.

An audio research group from Imperial College carried out a perceptual study where the 3DTI Toolkit-BS was used as a rendering tool to create binaural spatialized audio (Engel et al., 2021). They performed an objective and subjective evaluation of two different types of reverb simulation algorithms based on Ambisonics. First, what they

called an “hybrid Ambisonics”, where they used an Ambisonics approach to simulate the reverb and second, the virtual Ambisonic approach already implemented in the Toolkit (described in Section 3.6). From the first approach they found that the perceived quality of the sound ceased to improve beyond the third order of the Ambisonics, which is a lower threshold than the ones found by previous studies. They suggested that this can be obtained thanks to the fact that the 3DTI Toolkit-BS allows to process the direct sound in a different path using a different approach, in this case an HRTF convolution, which is not the case in other studies. In addition, the second approach is shown to produce a reverb simulation with comparable perceived quality to Ambisonics renderings.

The 3DTI Toolkit-BS can be used also as a generator of binaural 3D audio. Thanks to its control at such a low level, it allows the creation of spatialized audio signals that can be used for non-real-time scenarios. Among these scenarios, we can mention: scenarios with static sources where the listener does not interact, experiments that can be carried out via web, artificial perception systems (such as robots that are trained and tested to operate in a real environment), and input stimuli to aid the design and validation of auditory models.

In addition, it is worth mentioning at this point that the library also includes a hearing aid and a hearing loss simulators, which have not being described in this work because they are out of the scope of this PhD. People who use hearing aid devices are not able to wear a standard pair of headphones. The hearing aid simulator included in the 3DTI Toolkit allows these people to remove their hearing aid, wear a pair of headphones and compensate for their hearing loss using the virtual hearing aid simulator. This would allow people with hearing loss to enjoy the experience of binaural spatialized sound and perform multiple psychoacoustic virtual experiments, which can help, for example, to explore more deeply the different configurations of their hearing aids or to study how listeners with hearing impairments react to different acoustic scenarios. In addition, as mentioned, the 3DTI Toolkit includes a hearing loss simulation. Like the hearing aid simulator, this software component is easily added at the end of the binaural audio spatialisation process and would allow for psychoacoustics experiments that will help to learn more about the relationship between hearing loss and spatialised audio. In addition, this allows for development of applications aimed at enabling individuals with no hearing impairment to understand how hearing loss can compromise everyday activities, and how a hearing aid can improve this situation.

6.1.2 The 3DTI Toolkit as a tool to integrate 3D audio in Virtual Reality applications.

The 3DTI Toolkit is also a very powerful tool to carry out experiments in the field of VR, where the virtual scenario involves 3D audio. Immersive scenes where the relationship between spatial audio and video is analysed or studies regarding how the illusion of plausibility or presence change when spatial sound is included in the scene are some examples where the 3DTI Toolkit can be used to synthetize binaural audio.

Lerner et al. (2021) presented an experiment to study the limits of the peri-personal space (the close space surrounding our body) representation using the relationship between tactile processing and location of a sound in the 3D space, with a VR setup where the 3DTI Toolkit-BS was used to generate the audio stimuli. In this case, they pre-recorded the binaural sounds using the 3DTI Toolkit-BS and then imported the audio files to a virtual scene (created with Unity 3D game engine) where subjects had to perform an audio-tactile interaction task using the HTC Vive System.

Since Unity is a widely used tool for creating VR scenarios, the 3DTI Toolkit-BS has been integrated in a Unity package (presented in Section 3.7). This package allows to create immersive scenarios where the 3D audio can be rendered in real time. A pilot experiment, presented in a Spanish journal (Reyes-Lecuona, Márquez-Moncada, et al., 2021), makes use of the 3DTI Toolkit-BS, to perform a preliminary study where, within an immersive dynamic scenario, a subject is wearing an Oculus Rift. In this work we studied the influence of 3D audio on the perception of rotation gain in a virtual environment. The results reveal that the perception of the rotation gain is better for the visual modality since it provides more cues to detect such a gain. However, the manipulation of interaural differences in binaural audio may significantly affect such detection.

The 3DTI Toolkit was completed and evaluated successfully in the framework of the 3D Tune-In EU project, where multiple applications were also developed. These applications integrate the 3DTI Toolkit-BS to generate spatialized audio and the hearing loss and hearing aids components described in the previous sections. In this way, the applications, which are listed below, aimed to improve the lives of those affected by hearing loss.

- Musiclarity. An interactive web-based application that aims to improve the experience of listening to music, allowing the 3D spatialisation of each audio track separately. Moreover, it includes a hearing aid simulator, enabling the enjoyment of music for people with hearing loss.

- AudGam Pro. The main goal of this application was to allow users to test and adapt the settings of their hearing aid to different virtual scenarios where 3D audio was included.
- Play&Tune. Application oriented towards elderly hearing-aids users and simulates a series of 3D audio virtual scenes with different conditions. The goal of this application was to enable end-users to calibrate parameters of their hearing aids.
- Darius Adventure. A video game aimed at children without hearing loss, which tried to educate normal-hearing people about hearing impairment through simulated hearing loss.
- Dartanan. This application was specially oriented for children with hearing loss. That includes a series of mini-games, each one related to different settings of the player's hearing aid.

Within the framework of the PLUGGY EU project (Pluggable Social Platform for Heritage Awareness and Participation), a platform was developed which consists of a set of applications that were created to import, edit, process, manage and create binaural audio content within the PLUGGY Social Platform and Curatorial Tools (Comunità et al., 2020). This platform integrated the 3DTI Toolkit-BS to allow users to create and experience realistic 3D interactive virtual soundscapes within a web-based and mobile-based (iOS) platform.

6.2 Collaborations with Imperial College London

As mentioned on several occasions, the work on this thesis has been carried out within the 3D Tune-In project, in which we have closely collaborated with the Audio Experience Design group at Imperial College London led by Lorenzo Picinali. Many of the algorithms integrated in the 3DTI Toolkit have been discussed among all of us.

In addition to numerous face-to-face and online meetings, a 3-month stay at the Imperial College of London was held. In this stay, a pilot experiment was carried out for the study presented in Chapter 5. During this period all the experimental conditions and procedures of the study were agreed upon, and the pilot study was carried out with 10 English speakers to see the feasibility of the language and to have a first approach to the final study.

6.3 Current research projects where the Toolkit is currently included

At the time of writing there are two projects that make use of the 3DTI Toolkit. These projects guarantee the use and maintenance of the tool in the coming years, as well as the introduction of improvements, new experiments, publications and a wider dissemination.

The project SAVLab - Spatial Audio Virtual Laboratory - (PID2019-107854GB-I00) exploits the 3D Binaural Spatialisation Toolkit presented in this thesis to conduct a series of studies in the field of psychoacoustics. These studies look at the influence of spatial audio rendering on spatial sound perception, plausibility of virtual sound sources, speech intelligibility and listening effort. One of the main goals of this project is to make the 3DTI Toolkit a reference tool in psychoacoustics and acoustic Virtual Reality research and promoting its use.

Another project where the library will be integrated is the European project SONICOM (H2020-101017743; 8.06UE/58.8090)[29]. This will explore, map and model how the physical characteristics of spatialised auditory stimuli can influence observable behavioural, physiological, kinematic, and psychophysical reactions of listeners within social interaction scenarios. The project will develop a framework which will include a real time binaural rendering toolbox. As a rendering core, the 3D Tune-In Toolkit will be adapted and integrated. This core will be connected to different modules, such as modules for HRTF personalisation and calculation, room acoustics simulation, headphone equalisation or virtual hearing devices prototyping.

6.4 Future work

This thesis is focused on the design and implementation of a the 3DTI Toolkit Binaural Spatialisation tool, with the main goal of using it as a base for virtual psychoacoustical experiments. We propose some future lines of research related with the Toolkit.

- Regarding new developments:
  - Improving the customization by computing ILD compensation for near-field effects on-line, instead of relying on a pre-computed filter.

[29] https://www.sonicom.eu/ (retrieved January, 2020)

  - Adding multi-listener support, which would allow the use of binaural sound in collaborative virtual environments.
  - Adding delay simulation. Implement the simulation of sounds arriving at the listener delayed in time. It can be a very interesting feature to create 3D audio since is a characteristic of the sound sources that is very presented in our daily life.
  - Improving reverb simulation by implementing other widely used algorithm such as image-based technique or delay networks.
- Further work on assessing the 3DTI Toolkit performances is being planned, both in terms of signal processing and perceptual/subjective attributes (e.g. realism, audibility of processing artefacts, etc.), including comparisons with other existing tools and subjective evaluation, using subjects or auditory models.
- Use the 3DTI Toolkit to continue with the study about the influence of the HRTF in speech intelligibility, focusing on the monoaural cues of the HRTF and using the individual HRTF of the subjects. An abstract of this study has been presented in (Reyes-Lecuona, Cuevas-Rodriguez, et al., 2021).

6.5 Other studies

During the period of this PhD thesis, other parallel works have been carried out. They have not been described in this dissertation, but they are closely related to the subject of the PhD thesis and are described below.

A work titled "*Evaluation of the effect of head-mounted display (HMD) on individualized head-related transfer functions*" was carried out by the PhD candidate during a 6-month internship in Facebook Reality Labs. It is known that any element attached to the listener body, placed on the path between the sound source and the listener's ears, modifies the sound and thus the individual HRTF of the listener. With this study we wanted to know how the HRTF was modified if we measured it while the subject was wearing an HMD. To do so, the HRTF was acoustically measured on 24 human participants and a manikin head, with and without HDM, over 612 different directions. For the measurements we used a system with a rotating arc-shaped array of loudspeakers placed in an acoustically treated chamber. Then, we performed an objective evaluation comparing both HRTFs with and without HMD for each subject, based on the Spectral Difference Error (SDE) and on discrepancies in the ITD. The analysis of the result shown that distortion of the HRTF when looking at the SDE depends on both frequency and the direction of the incident sound, and it was bigger in the contralateral ear. ITD errors were found larger around the front side of the head. In addition, a perceptual evaluation was carried out, where 15 subjects evaluated the effect of the HMD

regarding the timbre and localization quality of the sound. The subjective evaluation validated the objective analysis and showed that the effect of the HMD has perceptible by the subjects. The study demonstrated that the distorted part of the HRTF measured wearing the HMD has to be discarded and generated using different algorithms that calculate the discarded direction using some of the directions that are not affected by the HMD. This study was presented in the International Congress on Acoustics (Maria Cuevas-Rodriguez et al., 2019), where more details about the procedure, the result and an extensive discussion can be found. In addition, this work was patented under the United States Patent called "Compensation for effects of headset on head related transfer functions" (Alon et al., 2020).

Another work, related with the study presented in this Chapter 5, "Study of the impact of non-individual HRTFs on speech intelligibility", was carried out together with the Technical University of Denmark. In this work, in contrast to Chapter 5 which performs a psychoacoustic evaluation to study the influence of different HRTFs on Speech Intelligibility (SI) with real subjects, we used an auditory model to study this influence. We employed a computational binaural speech intelligibility model developed by Jelfs et al. (2011). In this case, we used multiple HRTFs from the publicly available databases LISTEN and CIPIC, at different angles in the horizontal plane. The SI threshold obtained was different for the different conditions, concluding that there is an influence of the HRTF on the speech intelligibility. This was a conclusion that we also drew from the perceptual study presented in Chapter 5. This work was published in a JASA Express Letter (Ahrens et al., 2021). More details about the study are shown in the paper.

6.6 List of publications

6.6.1 Journals

Cuevas-Rodríguez, M., Picinali, L., González-Toledo, D., Garre, C., de la Rubia-Cuestas, E., Molina-Tanco, L., & Reyes-Lecuona, A. (2019). *3D Tune-In Toolkit: An open-source library for real-time binaural spatialisation.* PloS one, 14(3), e0211899. doi: 10.1371/journal.pone.0211899.

Cuevas-Rodriguez, M., Gonzalez-Toledo, D., Reyes-Lecuona, A., & Picinali, L. (2021). *Impact of non-individualised head related transfer functions on speech-in-noise performances within a synthesised virtual environment.* The Journal of the Acoustical Society of America, 149(4), 2573-2586. doi: 10.1121/10.0004220.

Ahrens, A., **Cuevas-Rodriguez, M.**, & Brimijoin, W. O. (2021). *Speech intelligibility with various head-related transfer functions: A computational modelling approach.* JASA Express Letters, 1(3), 034401. doi: 10.1121/10.0003618.

Reyes-Lecuona, A., Moncada, A. M., Bottcher, H. L., González-Toledo, D., **Cuevas-Rodríguez, M.**, & Molina-Tanco, L. (2021). *Audio Binaural y Ganancia de Rotación en Entornos Virtuales.* Revista de la Asociación Interacción Persona Ordenador (AIPO), 2(2), 54-62.

6.6.2 International conferences

Cuevas-Rodriguez, M., Gonzalez-Toledo, D., Rubia-Cuestas, E., Garre, C., Molina-Tanco, L., Reyes-Lecuona, A., ... & Picinali, L. (2017). *An open-source audio renderer for 3D audio with hearing loss and hearing aid simulations.* In AES Convention 142.

Cuevas-Rodríguez, M., González-Toledo, D., De La Rubia-Cuestas, E., Garre, C., Molina-Tanco, L., Reyes-Lecuona, A., ... & Picinali, L. (2018). *The 3D Tune-In Toolkit–3D audio spatialiser, hearing loss and hearing aid simulations.* In 2018 IEEE 4th VR workshop on sonic interactions for virtual environments (SIVE) (pp. 1-3). IEEE. doi: 10.1109/SIVE.2018.8577076.

Picinali, L., **Cuevas-Rodriguez, M.**, Gonzalez-Toledo, D. & Reyes-Lecuona, A. (2019). *Speech-in-noise performances in virtual cocktail party using different non-individual Head Related Transfer Functions.* In 23rd International Congress on Acoustic, 2019, no. 1, pp. 2158–2159.

Cuevas-Rodríguez, M., Lou Alon, D., Clapp, S. W. and Robinson, P. W., (2019) *Evaluation of the effect of head-mounted display on individualized head-related transfer functions.* In Proceeding of the 23rd International Congress on Acoustic, *ICA2019*, pp. 2635–2642.

6.6.3 Patent

Lou Alon, D., **Cuevas-Rodríguez, M.**, Mehra, R., and Robinson, P. W. (2021) *Compensating for effects of headset on head related transfer functions.* United States Patent: 10798515. https://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-

bool.html&r=1&f=G&l=50&co1=AND&d=PTXT&s1=10%2C798%2C515&OS=10%2C798%2C515&RS=10%2C798%2C515 (Retrieved February, 2022).

6.6.4 Demonstrations and workshop in international conferences

The 3DTI Toolkit was presented together with a demonstration, and tested by the attendees, at the following conferences:

4th International Conference on Spatial Audio, September 7th to 10th, 2017, Graz, Austria. Workshop the title: *An open-source C++ library for audio spatialisation and simulation of hearing loss and hearing aids - the 3D Tune-In Toolkit.*

EuroVR conference 2017 in Laval (France), 12th-14th December 2017. Workshop the title: *3D Tune-In Toolkit workshop.*

3rd International Congress of Art, Science and City, November 23 – 24, 2017, Málaga (Spain). Demonstration title: *3D Tune-In Toolkit demonstration.*

6.6.5 Other publications by the candidate not directly related with the PhD topic

Cuevas-Rodriguez, M., Poyade, M., Reyes-Lecuona, A., & Molina-Tanco, L. (2013). *A VRPN server for haptic devices using OpenHaptics 3.0.* In New Trends in Interaction, Virtual Reality and Modeling (pp. 73-82). Springer London.

Rivas-Blanco, I., Bauzano, E., **Cuevas-Rodriguez, M.**, del Saz-Orozco, P., & Munoz, V. F. (2013). *Force-position control for a miniature camera robotic system for single-site surgery.* In Proceeding Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on (pp. 3065-3070). IEEE.

Cuevas-Rodriguez, M., Rivas-Blanco, I., Bauzano, E., Gomez-deGabriel, J., & Muñoz, V. F. (2013). *Incorporación de un sistema de Mini-robots a la cirugía laparoscópica de incisión única.* XXXIV Jornadas de Automática, Terrasa (Barcelona). Con obtención del premio del área de Bioingeniería al mejor trabajo de investigación.

Rivas-Blanco, I., **Cuevas-Rodriguez, M.**, Bauzano, E., Gomez-deGabriel, J., & Muñoz, V. F. (2014). *Single Incision Laparoscopic Surgery Using a Miniature Robotic System.* In XIII Mediterranean Conference on Medical and Biological Engineering and Computing (pp. 105-108). Springer International Publishing.

Rivas-Blanco, I., Estebanez, B., **Cuevas-Rodriguez, M.**, Bauzano, E., & Munoz, V. F. (2014). *Towards a Cognitive Camera Robotic Assistant.* In Biomedical Robotics and Biomechatronics (BioRob), 2014, 5th IEEE RAS/EMBS International Conference. August 12-15, 2014, Anhembi Convention Center in São Paulo, Brazil. IEEE.

Gonzalez-Toledo, D., **Cuevas-Rodriguez, M.**, Molina-Tanco, L., Garre, C. & Reyes-Lecuona, A. (2015). *A Tool for Collaborative Decision Making on Service Information Linked to 3D Geometry of Complex Hierarchical Products.* In Proceedings of the EuroVR Conference 2015, CNR Lecco (Italy).

Cuevas-Rodriguez, M., Gonzalez-Toledo, D., Molina-Tanco, L., & Reyes-Lecuona, A. (2015). *Contributing to VRPN with a new server for haptic devices.* In Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology (pp. 193-193). ACM.

Gonzalez-Toledo, D., **Cuevas-Rodriguez, M.**, and Flores-Holgado, S., (2017). *Collaborative Management of Inspection Results in Power Plant Turbines,* in Dynamics of Long-Life Assets: From Technology Adaptation to Upgrading the Business Model, S. N. Grösser, A. Reyes-Lecuona, and G. Granholm, Eds. Cham: Springer International Publishing, pp. 193–208.

González-Toledo, D., **Cuevas-Rodríguez, M.**, Garre, C., Molina-Tanco, L., & Reyes-Lecuona, A. (2018). *HOM3R: a 3D viewer for complex hierarchical product models.* Journal of Virtual Reality and Broadcasting, 14(3).

González-Toledo, D., **Cuevas-Rodriguez**, **M.**, Molina-Tanco, L. and Reyes-Lecuona, A., (2018). *3D Object Rotation Using Virtual Trackball with Fixed Reference Axis*, in Proceeding of EuroVR, no. 1, pp. 3–5.

Gonzalez-Toledo, D., **Cuevas-Rodriguez, M.**, Molina-Tanco, L., & Reyes-Lecuona, A. (2022). *Still room for improvement in traditional 3D interaction: selecting the fixed axis in the virtual trackball*. The Visual Computer. https://doi.org/10.1007/s00371-021-02394-x

Appendix A

Forms and approval of the ethics committee

This appendix shows all the documents given to the participants of the experiment presented in Chapter 5, as well as the application sent to the ethic committee and the letter of acceptance. All documents are in Spanish as this was the language used with the participants

A.1 Consent form

HOJA DE CONSENTIMIENTO
Título del estudio: ***Individualización de Funciones de Transferencia Relativa a la Cabeza (HRTF) usando el Umbral de Recepción de Voz. (SRT).*** Investigador: Número de Participante:
• Me han explicado el estudio con un lenguaje que comprendo. Me han respondido todas las preguntas que he planteado. Comprendo qué es lo que sucederá durante el experimento y que es lo que se espera de mí. • Me han informado de mi derecho a abandonar el experimento en cualquier momento y de que, si lo hago, no tengo que dar ninguna explicación. • Me han informado de que cualquier cosa que diga en el cuestionario que tengo que completar o cualquier dato que se pueda obtener de mi participación en este experimento será totalmente confidencial; ni mi nombre ni ninguna otra información que pueda identificarme será usada. • Me han explicado que los resultados que se obtengan de este experimento podrán ser publicados de forma agregada junto con los de otros participantes con el

correspondiente tratamiento estadístico, pero ni mi nombre ni ninguna otra información que pudiera identificarme serán publicadas nunca.

Rodee la respuesta con un círculo:

Acepto participar en este experimento	Sí	No

Firma del participante:

NOMBRE (in capital letters)	**FIRMA**	**FECHA DE LA FIRMA** (in DD/MM/YYYY)

Firma del investigador que obtiene el consentimiento:

He discutido el estudio con el/la participante que firma más arriba con términos que él/ella comprende.

Creo que él/ella ha comprendido mis explicaciones y acepta participar en el experimento.

NOMBRE (in capital letters)	**FIRMA**	**FECHA DE LA FIRMA** (in DD/MM/YYYY)

A.2 Information sheet

Individualización de HRTF usando SRT: Hoja de Información del Participante

En este estudio, que se está realizando como parte del proyecto 3D Tune-In, estamos buscando una forma de poder personalizar el proceso de espacialización de audio binaural. Cuando usted percibe en su vida diaria los sonidos que le rodean, su sistema auditivo le permite percibir también de dónde provienen dichos sonidos. Esto es porque el sonido se transforma ligeramente dependiendo de la forma de su cabeza y orejas, y esa transformación es diferente dependiendo de la dirección de la que procede

el sonido. El sonido binaural es una simulación de estos fenómenos para proporcionar esa sensación de espacialización del sonido usando solamente auriculares.

En este experimento vamos a probar diferentes modelos de transformación del sonido para darle la sensación de que está oyendo una serie de palabras como si estuvieran pronunciadas enfrente de usted, al mismo tiempo que proporcionamos un ruido que viene de izquierda y derecha. Le pediremos que haga un esfuerzo por focalizar su atención en las palabras que provienen de delante y nos diga que palabras entiende. Esto nos permitirá, analizando sus respuestas, determinar qué trasformaciones son más adecuadas para proporcionar la sensación de espacialidad que buscamos.

Se le pedirá que complete una serie de actividades:

Primero se le pedirá que complete un breve cuestionario con preguntas sencillas como su edad y si sabe si tiene algún problema auditivo o atencional.

Se le pedirá entonces que complete una serie de pruebas durante diez días. Cada día se llevarán a cabo dos sesiones de 10 minutos, con otros 10 minutos de descanso entre las sesiones, lo que hace un total de media hora por día.

Durante la prueba se le pedirá que oiga una serie de palabras con ruido de fondo. Las palabras sonarán delante de usted mientras que el ruido sonará a los lados. Usted deberá prestar atención a las palabras y teclearlas cuando las entienda. Durante la sesión deberá usar auriculares y los sonidos que oirá nunca superarán los 66 decibelios, sonido equivalente a una conversación en un lugar concurrido.

Participa en este estudio voluntariamente y puede abandonarlo en cualquier momento sin dar ninguna razón. También puede abandonar cualquier sesión sin necesidad de dar ninguna explicación.

Privacidad y confidencialidad de los datos

No anotaremos su nombre en el cuestionario de modo que nadie podrá identificarle como participante y todo lo que conteste en el cuestionario será tratado confidencialmente. Es posible que publiquemos los resultados del estudio, pero nunca se mencionará su nombre.

A.3 Demographic Questionnaire

Título del estudio: ***Individualización de Funciones de Transferencia Relativa a la Cabeza (HRTF) usando el Umbral de Recepción de Voz (SRT).***

Número de participante: ……

1. ¿Cuál es su edad? Por favor, escríbalo aquí: ……..

2. ¿Cuál es su nivel de estudios? …….

3. Marque su género.

   Mujer ☐ Hombre ☐

4. ¿Sabe si padece usted algún tipo de problema de audición? En caso afirmativo, por favor, indíquelo aquí: …….

5. ¿Es el español su lengua materna? ……

6. Si el español no es su lengua materna, ¿Cuánto tiempo lleva usted viviendo inmerso en un ambiente de habla española? ……

7. ¿Sabe si tiene usted algún tipo de problema atencional?

   ……..

A.4 Application for the Ethic Committee

This section shows the original document (in Spanish) submitted to the ethic committee.

Individualización del HRTF mediante SRT

Este informe es una descripción del procedimiento y metodología experimental propuesta para un experimento que pretende determinar que: la capacidad de atención, en un problema de *cocktail party*, medida mediante la estimación del umbral de recepción del habla, SRT (del inglés Speech Reception Threshold), es una técnica apropiada para individualizar las funciones de transferencia de la cabeza (HRTF) para la espacialización binaural. Se presenta a este Comité de Ética para considerar su aprobación.

Objetivos del estudio justificados

Problema de investigación

Los sistemas inmersivos de Realidad Virtual vienen experimentando un constante desarrollo y popularización desde hace varias décadas. Este desarrollo ha tenido como principal protagonista la modalidad visual, sin embargo, la potencia de la estimulación auditiva para crear situaciones inmersivas es también muy alta. En este sentido, la localización tridimensional de las fuentes sonoras juega un papel importante en la capacidad inmersiva de estos sistemas. En este estudio nos centraremos en el conocido como sistema binaural, capaz de generar la sensación de espacialización tridimensional mediante el uso de auriculares.

Dentro de la percepción 3D del sonido, la anatomía del oyente juega un papel muy importante. Cuando un oyente recibe un sonido desde diferentes direcciones, su cabeza, cuello, torso y especialmente la forma de las orejas, modifican ligeramente dicho sonido. Esta modificación es diferente dependiendo de la posición desde donde se emita el sonido. Esa capacidad de modificación del sonido se puede caracterizar como una atenuación que depende de la frecuencia y de la dirección de la que viene el sonido, y es lo que se conoce como función de transferencia de la cabeza (Head Related Transfer Function: HRTF) (V.Ralph Algazi et al., 1997). Por otro lado, la HRTF juega también un papel importante en otros fenómenos diferentes a la simple percepción de localización de un sonido. Por ejemplo, algunos trabajos demuestran que algunos procesos atencionales, como el efecto *cocktail party,* se valen de la HRTF para poder fijar la atención hacia una

determinada dirección de procedencia del sonido. El efecto *cocktail party* es un fenómeno de atención selectiva por el que somos capaces de fijar la atención en una voz que se encuentra mezclada con otras voces de fondo. Ya en el primer trabajo publicado en el que se mencionó este efecto como el *cocktail party problem* (Cherry, 1953b), se informaba de que el efecto es mucho más claro en condiciones binaurales que monoaurales, por lo que podríamos plantear el uso de este proceso atencional para evaluar la adecuación de una HRTF a un determinado oyente.

Dado que la HRTF depende de la forma de la cabeza y orejas, se trata de una característica individual, diferente para cada persona. Por lo tanto, si queremos crear de forma sintética la ilusión de que un sonido proviene de una determinada, tendremos que usar la HRTF de esa persona específica. Pero medir una HRTF es difícil y requiere de un costoso equipamiento. Un procedimiento común es usar una HRTF real medida, no en el mismo oyente que la va a usar, sino en otra persona o bien medida en un maniquí. Cabe aquí, por tanto,la posibilidad de escoger entre varias HRTF reales, proporcionadas por bases de datos públicas, la que mejor se adapte al oyente. Es precisamente este mecanismo de individualización por selección de HRTF en el que se centra este estudio. Se pretende buscar un procedimiento, mediante una prueba experimental, que permita al oyente encontrar la HRTF idónea, es decir, aquella que mejor se ajusta a la suya propia.

Justificación de la investigación (propósito de la prueba)

El principal objetivo de la prueba es la validación de una novedosa técnica desarrollada para la individualización por selección del HRTF. Proponemos una técnica de individualización basada en procesos atencionales. El objetivo de este estudio es utilizar el efecto *cocktail party* y el reconocimiento del habla para la selección de una HRTF de entre un conjunto de candidatas. Estudios previos han demostrado que el efecto *cocktail party* se ve beneficiado por el sonido binaural (Hawley, 2004), y en particular, se ha demostrado la influencia de ciertas características intrínsecas a la HRTF, como la diferencia de tiempo interaural (ITD) o la diferencia de nivel interaural (ILD), aunque hasta donde sabemos, no se ha usado para comparar diferentes HRTF.

Proponemos llevar a cabo el estudio en dos idiomas diferentes. En primer lugar en el Imperial College de Londres con personas de habla inglesa y en segundo lugar, en la Universidad de Málaga para personas de habla hispana.

Hipótesis planteadas

Si la HRTF es una característica idiosincrática y el sistema atencional la utiliza para mejorar el reconocimiento de palabras en un entorno en el que hay ruido de fondo, deberíamos encontrar experimentalmente que hay una influencia del HRTF usada en el efecto *cocktail party y* que el efecto es diferente para diferentes sujetos. Por lo tanto, las hipótesis que planteamos comprobar experimentalmente son:

H1: El efecto *cocktail party* en el reconocimiento de palabras es sensible al HRTF usada para proporcionar sonido binaural. Esto es, diferentes HRFT proporcionaran diferentes capacidades de reconocimiento de las palabras objetivo sobre ruido (máscara) proveniente de otra dirección.

H2: La influencia de la HRTF en el reconocimiento de palabras en las condiciones del problema de la *cocktail party* es diferente para cada sujeto. Esto es, no hay HRTF que sean universalmente mejores para potenciar dicho efecto.

Metodología

La metodología del experimento será la misma para ambas pruebas, que se llevarán a cabo en el Imperial College de Londres y en la Universidad de Málaga.

Estímulos

Para el experimento se utilizarán dos tipos de estímulos. Las palabras que constituirán el estímulo objetivo, y que los participantes tienen que intentar reconocer, y un ruido que se usará como estímulo enmascarador.

Palabras objetivo

Como estímulo objetivo se usarán dos bases de datos de palabras en español e inglés. La primera pertenece a una lista ponderada para discriminación (de Cárdenas & Marrero Aguiar, 1994). Esta base de datos tiene como propósito realizar estudios de logoaudiometría y, por lo tanto, está diseñada para hacer medidas de umbral de recepción de voz (SRT), que es lo que vamos a hacer en nuestro experimento. De esta base de datos se han seleccionado las 222 palabras bisílabas. La segunda base de datos está formada palabras bisílabas en inglés, también ponderadas, de la que se han seleccionado un total de 200 palabras.

La fuente sonora que emite estas palabras se situara virtualmente justo enfrente del participante (azimut = 0°; elevación = 0°). Y la espacialización será puramente anecoica, es decir, no se añade reverberación. Antes de cada palabra se reproducirá

siempre la frase “Escribe la palabra…”, para ayudar a fijar la atención del sujeto en la fuente objetivo.

Máscaras

Como máscara se usarán señales de ruido coloreado que posee la misma densidad espectral de potencia que las palabras objetivo, y que se encuentran incluidos en la misma base de datos.

Se utilizará una pareja de máscaras diferentes e incorreladas pero de iguales características. Por este motivo, la potencia sonora de las dos máscaras juntas será considerada como 3dB por encima de la de cada una de ellas por separado. Las máscaras se situarán virtualmente a derecha e izquierda del participante ((azimut = ±90°, elevación = 0°).

Diseño experimental

Condiciones experimentales

Para contrastar nuestras hipótesis, consideraremos una variable independiente: la función de transferencia de la cabeza (HRTF), con ocho niveles: las siete HRTF de la base de datos LISTEN seleccionadas por Katz, B. F. G., & Parseihian, G. (2012) como más representativas de todo el conjunto de 51 cabezas que se incluyen en dicha base de datos, junto con una HRTF como condición de control. La HRTF control será generada sintéticamente siguiendo el modelo de propagación del sonido sobre una esfera perfecta para simular el ILD simplificado a un filtro de primer orden y un retardo (ITD) calculado también para una esfera perfecta (Brown, 1997).

Así pues, consideraremos en total las siguientes ocho HRTF como condiciones experimentales:

- HRTF 1: Sujeto #1008 de la base de datos LISTEN
- HRTF 2: Sujeto #1013 de la base de datos LISTEN
- HRTF 3: Sujeto #1022 de la base de datos LISTEN
- HRTF 4: Sujeto #1031 de la base de datos LISTEN
- HRTF 5: Sujeto #1032 de la base de datos LISTEN
- HRTF 6: Sujeto #1048 de la base de datos LISTEN
- HRTF 7: Sujeto #1053 de la base de datos LISTEN
- HRTF 8: Sintética. Se modela una cabeza esférica sin orejas.

Procedimiento

Los participantes serán recibidos el primer día y se les explicará el propósito del experimento, se les dejará leer las hojas de información y se les pedirá que den su consentimiento. El experimento consistirá en 20 sesiones, donde el participante completará dos sesiones cada día. A continuación, se describe el procedimiento de cada sesión.

Nuestra variable dependiente es el umbral de recepción de voz, SRT (del inglés Speech Reception Threshold). Para operacionalizar dicha variable desarrollaremos un procedimiento UP/DOWN típico de los procedimientos experimentales de psicoacústica. El procedimiento requerirá de la selección de una condición experimental de forma aleatoria y la presentación repetida de estímulos (palabras) al sujeto bajo dicha condición. Además, se capturará la respuesta del sujeto para cada uno de los estímulos (palabras) presentados. Cada uno de estos segmentos del experimento, en el que obtenemos un SRT para una determinada condición experimental, será denominado como un bloque. La selección SRT con el procedimiento UP-DOWN se repetirá 8 veces (una para cada HRTF), por lo que cada sesión tendrá un total de 8 bloques.

Cada vez que se reproduce un objetivo y se pide al sujeto que reconozca la palabra de destino, se denomina prueba. La estructura de cada uno de estas pruebas es la siguiente:

- Cada prueba comienza con la presentación de la palabra "Escribe la palabra" localizada virtualmente enfrente del participante.
- Un tiempo después, elegido aleatoriamente con una distribución uniforme de entre 500ms y 700ms, comienzan a sonar las máscaras con una potencia antes de ser filtradas por el HRTF de 66 dB_{SPL}.
- Un tiempo después, elegido aleatoriamente con una distribución uniforme de entre 200ms y 800ms, comienza a sonar la palabra objetivo, escogida aleatoriamente de la base de datos. Esta palabra se localiza virtualmente siempre enfrente del participante y tiene un nivel sonoro que va variando a lo largo de la prueba.
- Terminada la palabra, se retiran las máscaras 600ms después del final de la palabra objetivo.
- Durante todo este tiempo, el participante puede escribir mediante el teclado del ordenador la palabra que ha sido presentada, aunque no puede enviarla hasta que no termine de sonar.
- Una vez tecleada la palabra y pulsada la tecla "Intro", el sistema pasa automáticamente a presentar la siguiente prueba del bloque sin proporcionar ninguna información explícita sobre si la palabra tecleada era correcta o no.

En la Figura 1 se muestra un diagrama con la temporización de todos estos estímulos dentro de una prueba.

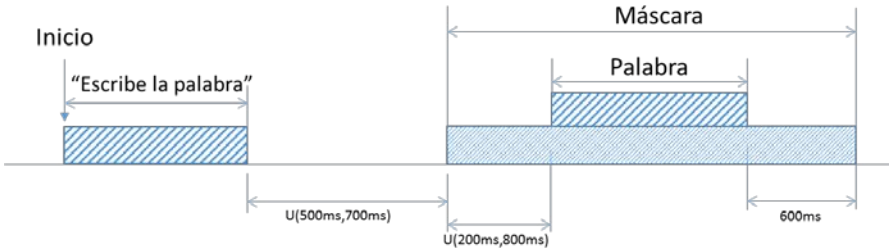


Figura 1. Distribución temporal de los estímulos dentro de una prueba.

En la primera prueba de cada bloque la palabra objetivo se presenta con una potencia de 6 dB inferior a la de la suma de las máscaras (3dB superior a la de cada una de las máscaras por separado). Si el participante teclea correctamente la palabra, en la siguiente prueba el nivel de la palabra objetivo desciende 2dB. Si el participante comete algún error al teclear la palabra, en la siguiente prueba el nivel será incrementado en 2dB. Se considera que una palabra es correcta cuando coincide con el target o cuando se produce un solo error en una de las letras.

El bloque concluye cuando se han producido cuatro inversiones en el sentido del cambio de potencia de la palabra objetivo. En la Figura 2 se muestra un ejemplo de un bloque con datos reales de un experimento piloto llevado a cabo.

Se considerará como umbral de recepción de voz (SRT) la diferencia entre la media aritmética de los niveles de potencia de la palabra objetivo al producirse cada una de estas cuatro inversiones, y el nivel de potencia de las máscaras juntas.

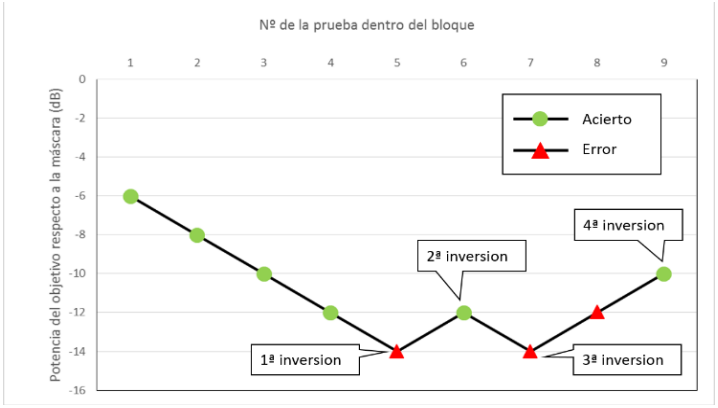


Figura 2. Ejemplo del procedimiento iterativo que permite calcular el SRT. Se señalan las pruebas en las que la respuesta del participante fue correcta (Acierto) y aquellas en las que fue

errónea (Error). En este caso, el SRT calculado sería SRT=(-14dB-12dB-14dB-10dB)/4=-12.5dB

Cada bloque contendrá un número variable de pruebas y arrojará como resultado una muestra del umbral de recepción de voz (SRT). En una sesión completa, el participante recorre una vez todas las condiciones experimentales que se derivan de las ocho HRTF, estando por tanto una sesión compuesta de 8 bloques.

Para desarrollar este procedimiento se utilizará un software desarrollado ex profeso que secuencia de forma automática todo el procedimiento, sin que ningún investigador tenga que intervenir durante el proceso. El programa guarda de forma automática un registro completo de toda la actividad. La información de cada prueba se guardará en un documento Excel, incluyendo: la palabra propuesta como objetivo, la que responde el participante, si acierta o falla, el nivel de potencia en el que se presenta y el tiempo que transcurre desde que se presenta la palabra (se inicia su reproducción) y se pulsa la primera tecla de la respuesta. Además, al final de cada bloque, el sistema registra de forma automática el SRT obtenido y lo ordena al final de forma conveniente para su análisis estadístico.

Análisis que se va a realizar de los datos

Para estudiar los resultados del experimento se realizará un análisis de varianza (ANOVA) de un factor para cada uno de los participantes. Hay que recordar que, dado que la principal variable bajo estudio (la función de transferencia de la cabeza, HRTF) es una característica individual, no se puede mezclar en el mismo ANOVA los resultados de dos participantes. Se debe realizar el estudio longitudinal sobre las medidas repetidas de cada uno de los participantes. Se analizará, por lo tanto, cada uno de los participantes por separado para contrastar la hipótesis de si las HRTF son significativamente diferentes para esa persona.

En el caso de que se confirme la primera hipótesis con este análisis, la mejor y peor HRT se extraerá mediante un test post-hoc. De esta forma, las frecuencias con la que cada HRTF aparecen como las mejores o las peores serán comparadas con las obtenidas por Katz, B. F. G. y Parseihian, G. (2012), los cuales utilizaron la misma base de datos de HRTF. Con el fin de contrastar esto, utilizaremos una prueba de $\chi 2$.

Tamaño de la muestra y duración de las sesiones

Como, en este caso particular, cada sujeto es un experimento completo en sí mismo, el análisis del tamaño de la muestra para cada uno de estos experimentos representa en

realidad el número de sesiones que debemos realizar para cada sujeto. Se ha llevado a cabo un experimento piloto con el mismo diseño experimental que el presentado en esta solicitud. De este experimento se han obtenido las varianzas que se utilizan para definir el tamaño de la muestra. Utilizando estos datos y mediante el uso de la herramienta de análisis estadístico GPower (Mayr, 2007) obtenemos un tamaño de la muestra de 160 para una potencia de la prueba del 95 % con $\alpha < 0.01$. Dado que el número de niveles del factor ANOVA es 8, deberemos realizar un total de 20 repeticiones (sesiones) por sujeto.

En cuanto al número de participantes, se reclutarán el mismo número de participantes que en el estudio de Katz, BFG, y Parseihian, G. (2012), donde se realizó una validación similar de este conjunto de HRTF, pero utilizando una técnica diferente. En este estudio, con 20 sujetos, bastaba tener una buena distribución de la mejor y peor HRTF, teniendo para cada HRTF al menos un sujeto que indicaba que esa era su mejor o peor HRTF.

Por tanto, el experimento planteado en esta solicitud constará de un total de 20 sujetos, donde cada uno de ellos deberá llevar a cabo 20 repeticiones del experimento. En el experimento piloto se observó que cada sujeto tardó aproximadamente 10 minutos por repetición, con un tiempo necesario para cada bloque de aproximadamente 1 minuto. Se propondrá a los participantes en cada visita realizar primero una sesión o repetición, pudiendo parar todo el tiempo que estime necesario al final de cada bloque, ya que la aplicación desarrollada les pedirá explícitamente que pulsen una tecla para continuar con el siguiente bloque.

Al terminar la sesión, se pedirá al participante que deje los auriculares y se le ofrecerá bebida (agua, refrescos o café) para que pueda preguntar todo lo que quiera sobre el desarrollo del experimento y se le explicará una vez que descanse, podrá realizar una segunda repetición ese mismo día, si no se encuentra cansado. En todo caso, el número máximo de sesiones por día será de dos y se procurará un descanso mínimo de 10 minutos entre las dos sesiones.

De esta forma, si el participante realiza 2 sesiones por día, con un descanso de 10 minutos entre sesiones, el tiempo estimado por día será de aproximadamente una media hora. Y como en total al sujeto se le pedirá que complete 20 sesiones, se requeriría que asista a la prueba 10 días.

Criterios de selección de la muestra

Los criterios para seleccionar la muestra en ambas universidades serán los mismos:

- Sujetos con edades comprendidas entre 20 y 50 años, para evitar el sesgo que supone la variación en los umbrales de audición que se producen por la edad.
- Sujetos sin problemas de audición, lo cual se las consultará a la hora de reclutarlos.
- Los sujetos deben hablar el idioma de la prueba con fluidez. Para el experimento en la Universidad de Málaga el idioma de la prueba será español y para el experimento en el Imperial College el idioma de la prueba será inglés.

Procedimiento para reclutar a los participantes

Se fijarán carteles informativos ofreciendo la posibilidad de participar en cada una de las facultades donde se va a llevar a cabo el experimento, con la intención de reclutar estudiantes y trabajadores de la misma. Además, se contactará con la delegación de alumnos que nos podrán servir de enlace para contactar con alumnos interesados.

Como incentivo, para asegurar la adherencia al experimento, a los participantes que completen el estudio, se ofrecerá un informe con sus resultados individualizados, lo que incluirá la determinación de la HRTF que debería usar esa persona en la espacialización binaural para conseguir resultados óptimos, y una copia de dicha HRTF en formato SOFA (Spatially Oriented Format for Acoustics). Además, se les entregará al terminar su participación un paquete con las aplicaciones desarrolladas en el proyecto 3D tune-In en las que podrán usar la HRTF que hayamos obtenido como la óptima para renderizar en 3D audios monourales.

Documentos para los participantes

Hojas de información a los participantes

Esta hoja contiene toda la información relevante al procedimiento del experimento. Se les dará a los participantes antes de realizar el experimento y a la hora de reclutarlos. Ver “Hoja informativa participante” adjunta a la solicitud.

Hojas de consentimiento informado

Esta hoja será firmada por los participantes al inicio del experimento, indicando que acepta participar en el experimento y que ha sido informado del procedimiento de la prueba. Ver “Hoja de consentimiento” adjunta a la solicitud.

Cuestionarios

Esta hoja será completada por los participantes al inicio del experimento, donde responderán a algunas cuestiones a cerca de datos personales relevantes para el estudio, como su edad o si padece algún problema de audición. Ver “Cuestionario” adjunto a la solicitud.

Protección de los datos

Solo se tomarán los nombres de los participantes en los formularios de consentimiento. Estos formularios se mantendrán separados del resto de datos del estudio en un lugar seguro. A los participantes se les asignará un número de identificación que les distinguirá de los demás al procesar los datos del estudio.

Los nombres de los participantes no se registrarán en las hojas de datos. Recopilaremos información sobre la edad y el sexo de los participantes, pero no será posible identificar a los participantes a partir de esta información. La información demográfica de los participantes será reportada colectivamente para describir la población de muestra y no usará identificadores individuales.

Todos los datos se almacenarán en sistemas protegidos por contraseña. Los datos de este estudio se utilizarán para generar resultados del proyecto y publicaciones académicas, pero no será posible identificar a los participantes de ninguna forma a partir de dichas publicaciones.

Cronograma de los experimentos

Los experimentos para los que se solicita la evaluación del Comité de Ética se desarrollarán entre los meses de Octubre y Diciembre en ambas universidades, de manera secuencial.

Equipamiento necesario

Los dos experimentos se llevarán a cabo con el mismo equipamiento. Se utilizará un software desarrollado expresamente para ambos experimentos, que secuencia de forma automática todo el procedimiento, sin que ningún investigador tenga que intervenir durante el proceso. Para reproducir el sonido se utilizará un interfaz de audio MOTU 896 mk3 que se controlará desde el ordenador usando un driver ASIO que sustituye al driver del sistema operativo, lo que le permite control total sobre el nivel de audio

proporcionado. Para presentar los estímulos auditivos se utilizará unos auriculares SONY Modelo MDR-7506.

Memoria Económica

El estudio para el que se solicita evaluación se encuentra incluido en los trabajos del proyecto de investigación 3D Tune-In, financiado por la Unión Europea a través del programa Horizonte2020.

Tabla 1. Memoria económica

Concepto	Coste estimado	Financiación
Personal (dos personas-mes)	8.000 €	Adquiridos con cargo al proyecto 3D Tune-In
Interfaz de audio MOTU 896 mk3 (material inventariable)	1164.94 €	Adquiridos con cargo al proyecto 3D Tune-In
Ordenador portátil (material inventariable)	1.000 €	Adquiridos con cargo al proyecto 3D Tune-In
Tres auriculares SONY Modelo MDR-7506 (material fungible)	116.16 €/unidad	Adquiridos con cargo al proyecto 3D Tune-In
Estancia en Londres para llevar a cabo el experimento en el Imperial College	2900 €	Financiado por el proyecto 3D Tune-In y una ayuda concedida por el Vicerrectorado de Estudios de Posgrado de la Universidad de Málaga

Marco regulatorio

El diseño de este estudio, así como todos los del proyecto 3D Tune-In, se adhiere a la Declaración de Helsinki y al informe sobre ética de las tecnologías de la información y las comunicaciones (2012), del Grupo Europeo de Ética en Ciencia y Nuevas Tecnologías de la Comisión Europea. El proyecto 3D Tune-In también se ajusta a la Carta de Derechos Fundamentales de la Unión Europea y a la Directiva 95/46/EC del Parlamento Europeo y del Consejo, de octubre de 1995, sobre la protección de los individuos respecto al procesado de sus datos personales y sobre el movimiento de dichos datos. En la Descripción del Trabajo del proyecto, que se adjunta a la solicitud, se puede encontrar en la sección 5 (pp. 90-96) una descripción más completa de los principios éticos que rigen el proyecto.

Documentación que se adjunta

Listado de documentos adjuntos a la solicitud:

- Hojas de información a los participantes (versión en español e inglés)
- Hojas de consentimiento informado (versión en español e inglés)
- Cuestionarios (versión en español e inglés)
- Autorización del director de la E.T.S. de ingeniería de Telecomunicación de la Universidad de Málaga
- Autorización del director de la Dyson School of Design Engineering del Imperial College de Londres
- Compromiso del investigador principal del proyecto
- Descripción del Trabajo del proyecto 3D Tune-In

Referencias

Algazi, V. R., Divenyi, P. L., Martinez, V. A., & Duda, R. O. (1997). Subject dependent transfer functions in spatial hearing. In Anon (Ed.), *Proceedings of the 1997 40th Midwest Symposium on Circuits and Systems. Part 1 (of 2)* (Vol. 2, pp. 877–880). Univ of California, Davis, Davis, United States: IEEE.

Brown, C. P., & Duda, R. O. (1998). A Structural Model for Binaural Sound Synthesis. *IEEE Transactions on Speech and Audio Processing*, *6*(5), 476–488.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. https://doi.org/10.1121/1.1907229.

de Cárdenas, M. R., & Marrero Aguiar, V. (1994). *Cuaderno de logoaudiometría. Guía de referencia rápida*. Universidad Nacional de Educación a Distancia, UNED.

Hawley, M. L., Litovsky, R. Y., & Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *Journal of the Acoustical Society of America*, *115*(2), 833–843. https://doi.org/10.1121/1.1639908.

Katz, B. F. G., & Parseihian, G. (2012). Perceptually based head-related transfer function database optimization. *The Journal of the Acoustical Society of America*, *131*(2), EL99-105. https://doi.org/10.1121/1.367264

A.5 Approval of the ethic Committee

Servicio Andaluz de Salud
CONSEJERÍA DE SALUD

Comité de Ética de la Investigación Provincial de Málaga

Dra. Dña. Gloria Luque Fernández, Secretaria del CEI Provincial de Málaga

CERTICA:

Que en la sesión de CEI de fecha: 28/09/2017 ha evaluado la propuesta de D/Dña.: Arcadio Reyes Lecuona, referido al Proyecto de Investigación: "Individualización de la función de transferencia de la cabeza (HRTF) para espacialización de audio binaural usando la estimación del umbral de recepción del habla (SRT) en el paradigma de la Cocktail Party".

Este Comité lo considera ética y metodológicamente correcto.

La composición del CEI en esta sesión es la siguiente:

Dra. Ana Alonso Torres (UGC Neurociencias)
Dr. Migel Angel Berciano Guerrero (UGC Oncología Clínica)
Dra. Encarnación Blanco Reina (Farmacología Clínica)
Dra. Marta Camacho Caro (UGC Ginecología)
Dr. Fermín Mayoral Cleries (UGC Salud Mental)
Dr. José L. Guerrero Orriach (UGC Anestesia y Reanimación)
Dr. Antonio López Téllez (Médico de Familia)
Dra. Leonor Ruiz Sicilia (UGC Salud Mental)
Dr. José Leiva Fernández (Médico Familia)
Dra. Mª Dolores Bautista de Ojeda (UGC Anatomía Patológica)
Dra. Gloria Luque Fernández (Investigación)
Dra. Mª Mercedes Márquez Castilla (Médico Familia)
Dña. Inmaculada Lupiáñez Pérez (Enfermera)
Dra. Cristobalina Mayorga Mayorga (Laboratorio)
Dra. Mª Victoria de la Torre Prados (UGC UMI)
Dr. Benito Soriano Fernández (Médico Familia)
Dr. Antonio Pérez Rielo (UGC UCI)
D. Ramón Porras Sánchez (RRHH-Abogado)
Dra. Mª Angeles Rosado Souvirón (UGC Farmacia)
Dra. Mª Carmen Vela Márquez (Farmacéutica Distrito)

Lo que firmo en Málaga, a 16 de Octubre de 2017

CONSEJERÍA DE SALUD
Servicio Andaluz de Salud
JUNTA DE ANDALUCIA
HOSPITAL REGIONAL UNIVERSITARIO DE MÁLAGA
C E I
PROVINCIAL DE MÁLAGA
C.I.F.: Q-9150013-B

Fdo.: Dra. Gloria Luque Fernández
Secretaria del CEI

Appendix B

Resumen extendido

En este anexo se presenta un resumen extendido en español del trabajo realizado en esta tesis doctoral. En primer lugar, en la Sección B.1 se muestra una breve introducción a los conceptos básicos sobre simulación de audio 3D binaural, haciendo alusión al estado del arte. Además, esta sección presenta el contexto de la tesis, la motivación y los objetivos que se han abordado. A continuación, se describen las dos líneas de investigación en las que se fundamenta este trabajo: diseño, desarrollo y evaluación de un espacializador de audio binaural (Secciones B.2 y B.3) y un estudio que hace uso de dicha herramienta y que tiene como objetivo entender el impacto de las HRTFs no individualizadas en la inteligibilidad del habla (Sección B.4). Finalmente, se detallan las conclusiones en la Sección B.5.

B.1 Introducción

La espacialización binaural se refiere a la habilidad que tiene nuestro sistema auditivo para interpretar todas las características del sonido que llega a nuestros oídos y percibir la localización de las fuentes sonoras en el espacio tridimensional. Estas características del sonido se agrupan en indicios binaurales y monoaurales. Los indicios binaurales están basados en las diferencias entre las señales que llegan a cada oído, y se dividen en dos tipos: diferencias en tiempo (ITD – siglas del término en inglés Interaural Time Difference) y diferencias en nivel (ILD – siglas del término en inglés Interaural Level Difference). Estos indicios fueron inicialmente introducidos por Rayleigh (1907), con su teoría Dúplex, que describe la capacidad del oyente para lateralizar las fuentes sonoras (localizar sonidos procedentes de ubicaciones izquierda-derecha). Los indicios monoaurales están basados en las modificaciones del sonido al entrar en contacto con el

torso, los hombros, la cabeza y el pabellón auditivo del oyente (L. Wightman & Kistler, 1996). Todos estos indicios forman un conjunto de filtros dependientes de la dirección de la que provenga el sonido, conocido como HRTF (siglas del término en inglés Head Related Transfer Function). Una dirección determinada viene caracterizada por una repuesta al impulso o HRIR (siglas del término en inglés Head Related Impulse Response). La HRTF es individual para cada oyente y representa una caracterización completa de los indicios utilizados por este para localizar una fuente en un entorno anecoico, donde el sonido viaja por un camino directo desde la fuente hasta el canal auditivo del oyente (Wightman & Kistler, 1989a, 1989b). En un entorno cerrado, como una habitación, al camino directo se le añaden una serie de reflectiones del sonido en las diferentes paredes y obstáculos. Al conjunto de filtros que caracterizan este caso se les conoce como BRIR (siglas del término en inglés Binaural Room Impulse Response) (Välimäki et al., 2012).

La simulación de audio binaural consiste en el procesado de estímulos sonoros monofónicos y anecoicos, para añadir los indicios auditivos descritos anteriormente. El audio se entregará al oyente mediante unos auriculares, y éste será capaz de percibir su localización, así como ciertas características de la sala en la que se encuentra (V. Algazi & Duda, 2011). La simulación de audio espacial utilizando tecnología binaural se considera muy cercana a la escucha natural (Langendijk & Bronkhorst, 2000; Martin et al., 2001). Se denomina escena virtual auditiva (más conocido como VAS, siglas del término en inglés Virtual Auditory Scene) a un entorno artificial en el que el oyente puede percibir diferentes sonidos virtuales como si fueran reales, situados en puntos concretos del espacio. Las señales de audio espacializado simuladas en un sistema binaural dependen de la posición relativa entre la fuente y el oyente. En un VAS inmersivo, tanto las fuentes como el oyente pueden estar en constante movimiento provocando un cambio en estas posiciones relativas y, por tanto, una modificación de las señales auditivas. El hecho de tratarse de sistemas que no son invariantes en el tiempo, hace necesario que este sea capaz de detectar constantemente las diferentes posiciones y realizar la simulación del audio espacial en tiempo real. Estos sistemas se denominan sistemas VAS dinámicos y en tiempo real (Xie, 2013). Un esquema detallado de este tipo de sistemas puede verse en Serafin et al. (2018).

En la Figura 85 se muestra una estructura básica de un sistema VAS dinámico binaural, que consta de tres partes:

- Información sobre el oyente (datos individuales como la HRTF), la fuente de sonido (los estímulos, la posición y orientación espacial y el nivel de sonido) y el

entorno (la geometría de la sala, las características de los materiales de las superficies y las características de absorción del aire).

- Procesamiento binaural en tiempo real, el cual toma toda la información anterior, junto con la posición y orientación de la cabeza del oyente, y simula el sonido espacializado. Este procesamiento puede dividirse en tres bloques: (1) simulación del sonido directo (sonido que va desde la fuente hasta el canal auditivo del oyente), cuya transformación se caracteriza con la HRTF, (2) la simulación de las reflectiones del sonido dentro de la habitación, caracterizados por la RIR (del inglés Room Impulse Response) o la BRIR si se tiene en cuenta el efecto combinado de sala y oyente y (3) la simulación de la distancia, cuyos indicios más destacados son: nivel de la señal (Shinn-Cunningham, 2000), familiaridad con el estímulo (Kolarik et al., 2016) y la relación entre el camino directo y el reverberante (Begault, 1994).
- La señal binaural (señales izquierda y derecha) son el resultado del procesado en tiempo real y se entrega al oyente a través de unos auriculares.

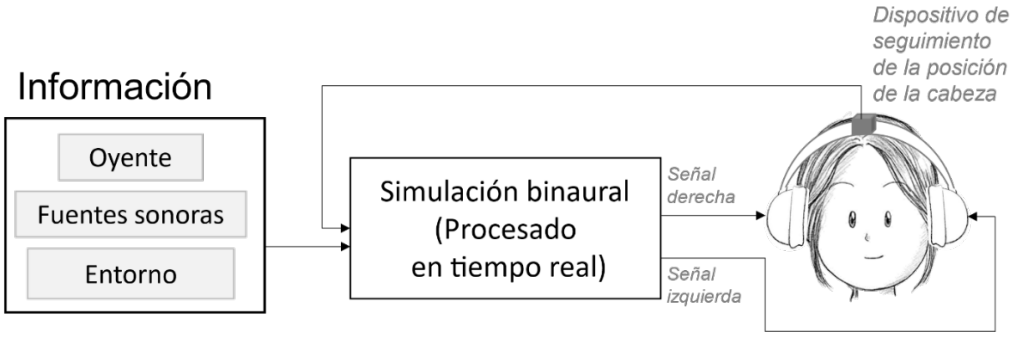


Figura 86. Estructura básica de un sistema VAS dinámico.

Esta tesis presenta el diseño y desarrollo de una herramienta que permite simular sistemas VAS dinámicos, implementada siguiendo la estructura presentada anteriormente. Dicha herramienta, denominada 3DTI Toolkit-BS, se introducirá en las siguientes subsecciones (B.1.1 y B.1.2) y se describe brevemente en la Sección B.2.

El procesado de audio en un sistema de tiempo real se puede volver muy complicado si se incorporan múltiples fuentes en movimiento. Esto conlleva una actualización continua de la información y el procesamiento de la señal, e incurre en un gran coste computacional. Además, al ser un sistema que varía en el tiempo, el cambio de posición hace que diferentes parámetros y filtros aplicados en el procesado tengan que cambiar. Si no se implementa con cuidado, esto podría provocar cambios abruptos en la señal, produciendo una serie de artefactos audibles no deseados, que provocan la incomodidad del oyente, así como la pérdida de naturalidad y presencia. Este problema se abordará

dentro de cada uno de los algoritmos de procesado en tiempo real implementados en el 3DTI Toolkit-BS, los cuales han sido evaluados en detalle en esta tesis y sus resultados son resumidos en Sección B.3.

Además de la localización de la fuente sonora, el sonido espacial también ayuda al oyente a centrase en un determinado sonido, cuando aparece más de uno en la escena. A la capacidad del cerebro para entender un discurso determinado en una situación con múltiples sonidos se le conoce como efecto de *cocktail party*. En este caso, el sistema auditivo permite al oyente centrar su atención en un sonido específico que llega desde una dirección concreta (sonido objetivo) cuando también le llegan uno o varios sonidos que interfieren (sonidos enmascaradores). El sistema auditivo puede aprovechar la separación espacial entre el sonido objetivo y los enmascaradores para detectar el objetivo con mayor eficacia. Este fenómeno se le conoce como desenmascaramiento espacial (en inglés *spatial release from masking* o *spatial unmasking)*. Como parte de esta tesis se ha llevado a cabo un experimento basado en el fenómeno de *spatial release from masking*, cuyo resumen se presenta en la Sección B.4.

Un ámbito en el que el sonido espacial ha encontrado una enorme y exitosa aplicación es el de los sistemas de Realidad Virtual (RV). Estos sistemas han experimentado un crecimiento y popularización constantes durante las últimas décadas, en las que la mayor parte del esfuerzo en investigación y desarrollo se ha realizado sobre la modalidad visual. Sin embargo, el mundo real está lleno de estímulos auditivos y estamos totalmente acostumbrados a recibir sonidos desde cualquier punto del espacio. De este modo, parece lógico que el audio espacial deba incluirse en las aplicaciones de RV, en aras de la inmersión y el realismo (Bormann, 2005). Afortunadamente, la situación está cambiando y el audio 3D en sistemas de RV se ha convertido en una importante y fuerte línea de investigación en los últimos años. Una muestra de ello puede verse en el hecho de que, mientras que el número de publicaciones sobre audio 3D en RV entre los años 1999 y 2009 se situaba en 2228, en los últimos 10 años (entre 2010 y 2020) éstas han aumentado hasta 4397[30]. El audio 3D también ha llamado la atención de los principales actores de la industria de la RV, como Google u Oculus. En 2017, Google lanza Resonance Audio (Google, n.d.) como código abierto, una herramienta para incluir audio 3D en escenarios de RV, convirtiéndose en uno de los renderizadores más utilizados en la actualidad tanto para aplicaciones comerciales como para la investigación. Otras herramientas populares

[30] Este dato ha sido obtenido en la Plataforma Scopus (www.scopus.com), realizando una búsqueda por título, abstract y keywords con las siguientes palabras: “audio” OR “sound” OR “auditory” OR “acoustic” OR “acoustics” OR “hearing”) AND (“Virtual Reality” OR “Augmented Reality” OR “Mixed Reality” OR “Extended Reality”).

son Oculus VR (2020) de Facebook/Meta o el motor de renderizado de audio de Microsoft (2020).

B.1.1 Contexto y motivación

Esta tesis doctoral se ha desarrollado dentro del grupo de investigación DIANA de la Universidad de Málaga, en el marco del proyecto 3D Tune-In[31], financiado por la Unión Europea. Este proyecto tenía como objetivo utilizar el sonido 3D y las técnicas de gamificación para apoyar a las personas que utilizan audífonos. Dentro del proyecto, el grupo de investigación DIANA se encargó de desarrollar el "3DTI Toolkit", una librería C++ de código abierto que integra funcionalidades de espacialización binaural, junto con otras características relacionadas con el audio, como un simulador de pérdida auditiva y audífonos. Dentro del desarrollo del 3DTI Toolkit, esta tesis doctoral se ha centrado en el diseño, desarrollo y evaluación del espacializador binaural llamado **3DTI Toolkit-BS** (3DTI Toolkit Binaural Spatialiser).

Es cierto que, con el auge de la RV en los últimos años, se han publicado un gran número de herramientas de renderizado binaural. Sin embargo, la principal motivación para el desarrollo de una librería de espacialización binaural personalizada y desarrollada desde cero fue la necesidad de cumplir una serie de requisitos y ofrecer varias características que no existían al principio del desarrollo de esta tesis doctoral, y que además e incluso hoy en día, todas juntas no se encuentran disponibles en otras herramientas existentes. Dichas características se enumeran a continuación:

- Soporte para múltiples plataformas, incluyendo web.
- Posicionamiento y libertad de movimiento en todo el espacio tridimensional de las fuentes y el oyente, incluyendo distancias muy cercanas y muy lejanas.
- Carga y personalización de HRTFs.
- Simulación de sonido reverberante espacializado, configurable para simular de forma realista una determinada habitación.
- Transiciones suaves en situaciones dinámicas sin artefactos audibles.

El 3DTI Toolkit-BS integra en un único paquete de código abierto varias técnicas y funcionalidades desarrolladas y evaluadas en los últimos 20 años de investigación en audio espacial. Durante la fase de desarrollo, se ha prestado especial atención a los aspectos de la espacialización relacionados con entornos dinámicos, lo que ha dado lugar

[31] This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644051.

a una simulación realista y fluida de las fuentes sonoras en movimiento. La implementación de todas estas funcionalidades dentro de una herramienta de código abierto proporciona el control total sobre el proceso de espacialización y lo abre a futuros desarrollos dentro de las comunidades de audio 3D.

Además, y gracias a los requisitos que cumple la librería, se pretende posicionar como una herramienta que sea utilizada como base para realizar experimentos de psicoacústica. Para demostrarlo el 3DTI Toolkit-BS se ha utilizado en la realización de un experimento de psicoacústica, presentado en la Sección B.4, donde se ha estudiado la influencia de una HRTF no individual en la inteligibilidad del habla. Se sabe que las HRTFs utilizadas para la simulación del audio tienen un impacto en la inteligibilidad del habla, sin embargo, aún no se ha investigado en profundidad cómo afectan estas funciones a cada individuo y el impacto de la elección de una HRTF en concreto para un individuo específico dentro de un entorno de *cocktail party*. La realización de este estudio permitió tanto evaluar el rendimiento del 3DTI Toolkit-BS como profundizar en el estudio de las HRTFs y su relación con la inteligibilidad del habla.

B.1.2 Objetivos

El principal objetivo de esta tesis doctoral es el diseño y desarrollo de una herramienta de espacialización binaural, que permita integrar el audio 3D de la manera efectiva y flexible en un entorno de RV. Para ello, la herramienta debe:

1. Simular la propagación del sonido directo entre la fuente y el oyente de forma precisa, basándose en las características individuales del oyente y teniendo en cuenta todos los indicios que conducen a la percepción de una fuente proveniente de un determinado punto del entorno.

2. Simular la reverberación del entorno con precisión, recogiendo las características direccionales del entorno reverberante. Sin embargo, hay que hacerlo de forma eficaz, para no obtener un coste computacional muy alto.

3. Admitir fuentes estáticas y en movimiento para simular escenarios dinámicos.

4. Procesar el audio 3D en tiempo real en un "PC comercial" sin un hardware específico y sin latencia perceptible.

5. Garantizar la suavidad de los cambios de audio en situaciones dinámicas cuando se modifican algunas características del escenario.

Adicionalmente, para evaluar el espacializador binaural, testear su uso como base de experimentos de psicoacústica y profundizar en el estudio de las HRTFs y su relación con las características individuales del oyente, se ha realizado un estudio perceptivo cuyos objetivos principales son:

6. Estudiar el impacto de la HRTF en la inteligibilidad del habla.

7. Estudiar el impacto, en diferentes oyentes, de HRTFs no personalizadas en la inteligibilidad del habla dentro de un contexto de *cocktail party*.

8. Evaluar el uso del 3DTI Toolkit-BS en un experimento psicoacústico virtual.

B.2 El espacializador binaural 3DTI Toolkit

Como se ha mencionado anteriormente, el 3DTI Toolkit-BS es un renderizador de audio, desarrollado en C++, de código abierto y multiplataforma, que permite el diseño y la creación de escenarios virtuales auditivos. El 3DTI Toolkit-BS espacializa las fuentes de sonido incluidas en la escena de forma separada para el camino directo y el camino de reverberación; caminos que recorre la señal desde la fuente hasta el oyente. Esta estructura permite una resolución espacial muy alta para el camino directo, donde cada fuente se procesa de forma independiente, y una simulación eficiente de la reverberación, donde todas las fuentes se procesan conjuntamente mediante un sistema Ambisónico virtual, manteniendo ciertas características dependientes de la posición de la fuente, pero con menor resolución. Los algoritmos que se han implementado en cada uno de los caminos se muestran en la Figura 86 y son descritos en las siguientes subsecciones.

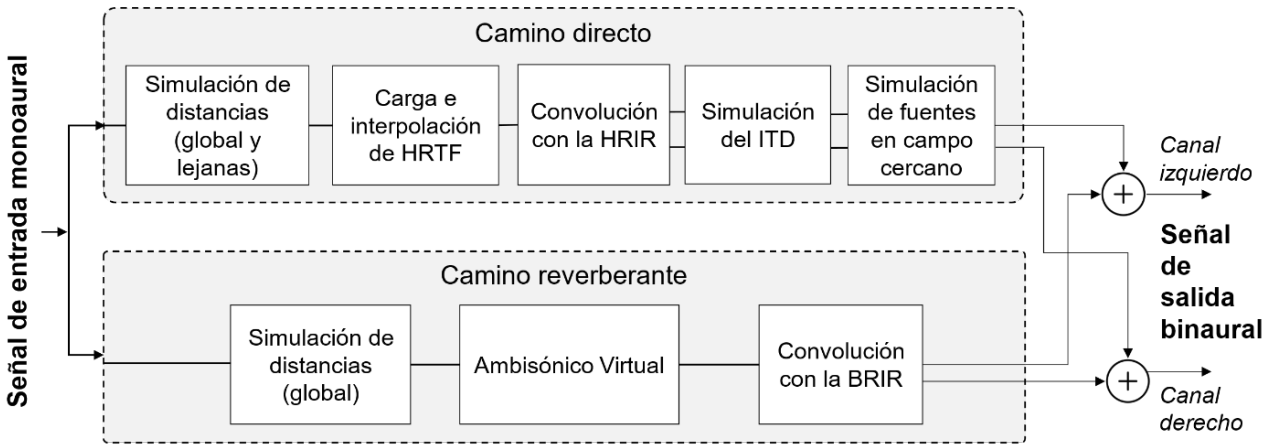


Figura 87. Estructura del 3DTI Toolkit-BS para la espacialización binaural de una fuente de audio mono.

B.2.1 Simulación del camino directo

- **Simulación de distancias.**

Para simular la distancia que recorre la señal de audio desde la fuente al oyente, el 3DTI Toolkit-BS implementa un algoritmo de atenuación de la señal para el desplazamiento de una onda esférica, donde ésta es atenuada 6dB cada vez que se dobla la distancia. Este parámetro de 6dB es configurable en el Toolkit. Para distancias muy grandes (mayores de 15 m.), el Toolkit además simula la atenuación producida por la absorción del aire con un filtrado paso bajo, diseñado siguiendo el estándar ISO 9613-1 (1993).

Para evitar artefactos audibles en situaciones dinámicas, se ha implementado el siguiente mecanismo de suavizado. A cada muestra del buffer de audio se le aplica un valor de atenuación adaptativo a_i, acercándose asintóticamente a la nueva atenuación deseada $A(d)$, mediante la siguiente ley:

$$a_i = (1 - \rho) \cdot a_{i-1} + \rho \cdot A(d) \tag{B.1}$$

donde ρ se calcula como

$$\rho = 1 - \exp \frac{\log 0.01}{t_a \cdot f_s} \tag{B.2}$$

Siendo, t_a (tiempo de ataque), el tiempo en el que la envolvente alcanza el 99%, y f_s la frecuencia de muestreo.

- **Carga, interpolación y convolución de HRTF.**

El 3DTI Toolkit-BS permite la carga de una HRTF medida en posiciones arbitrarias en formato SOFA (*SOFA General Purpose Database*, 2017). Las HRTFs se miden generalmente a una única distancia del oyente, formando una esfera alrededor de este, para un conjunto limitado de acimuts y elevaciones. Para poder simular una fuente proveniente de cualquier punto del espacio 3D, en primer lugar, se hace una corrección del efecto de paralaje. Esta corrección se presenta en Romblom & Cook (2008) y se basa en la modificación de los ángulos relativos entre la fuente sonora y cada uno de los dos oídos. En segundo lugar, se lleva a cabo una interpolación baricéntrica que calcula la HRIR en el punto deseado utilizando las tres HRIRs conocidas más cercanas. Para obtener estas HRIRs dentro de la esfera donde las HRIRs han sido medidas, se ha utilizado la fórmula de Haversine (C. C. Robusto, 1957) para el cálculo de las distancias. Una vez tenemos la distancia entre todos los puntos nos quedamos con los tres puntos

más cercanos y que formen un triángulo alrededor del punto deseado y los utilizamos para llevar a cabo la interpolación baricéntrica.

En la práctica, encontrar la HRIR más cercana en un conjunto arbitrario es un proceso costoso. Por esta razón, el proceso de interpolación ha sido dividido en dos fases. Una primera fase que se realiza "fuera de línea" (no en tiempo real), lo que da como resultado una tabla HRTF remuestreada. En esta primera fase, la HRTF se remuestrea en una cuadrícula regular (5 grados por defecto tanto en acimut como en elevación) utilizando una interpolación baricéntrica. A continuación, se aplica una FFT a cada HRIR y se almacena en memoria. El objetivo principal de esta primera fase es obtener una tabla HRTF regular para simplificar y acelerar la segunda fase del proceso, que se realiza en línea, en tiempo real. En la segunda fase, para llevar a cabo la interpolación baricéntrica se utilizan las tres HRIRs más cercanas de la tabla HRTF remuestreada. Esta operación es ahora mucho más sencilla, ya que la tabla tiene una base regular y sabemos a qué distancia se encuentra las HRIRs más cercanas. Para evitar el "efecto de filtro de peine" (el cual produce artefactos audibles) que se produce cuando interpolamos señales con un módulo similar pero distinta fase, las HRIRs utilizadas en la interpolación han sido alineadas, extrayendo el retardo inicial de cada una de ellas. Este retardo se interpolará por separado, utilizando la misma técnica de interpolación, y se añadirá tras el proceso de convolución. Finalmente, se convoluciona cada fuente con su correspondiente HRIR en el dominio de la frecuencia.

- **Simulación del ITD.**

Como se ha comentado anteriormente, las HRIRs deben ser cargadas en el Toolkit con el retardo inicial (o ITD si tenemos en cuenta la diferencia de retardo entre los dos oídos) eliminado y almacenado en un campo diferente del fichero SOFA. Tras la interpolación, el Toolkit añade el ITD a la señal que estamos procesando. Este ITD puede ser calculado con una interpolación baricéntrica de los ITDs de las HRIRs más cercanas, o bien calculado en base al radio de la cabeza del oyente, utilizando la fórmula de Woodworth et al. (1954), y de este modo utilizar un ITD personalizado para el oyente. Además, se implementa un algoritmo para evitar las distorsiones en la señal que se producen cuando el ITD cambia, lo que ocurrirá en entornos dinámicos. Este algoritmo implementa una compresión o expansión de las muestras del buffer de audio según el cambio de valor del ITD.

- **Simulación de fuentes en el campo cercano.**

El 3DTI Toolkit realiza una corrección de la HRIR para simular fuentes que se encuentran en el campo cercano (distancias menores a 2 m.), donde el ILD tiene un efecto diferente a otras distancias. Para ello implementa un algoritmo basado en un filtro de diferencias, el cual sigue el modelo de cabeza esférica presentado por Duda & Martens (1998). Este filtro predice las diferencias espectrales entre una fuente de campo cercano y una fuente situada en la misma dirección, pero a la distancia en que se midió la HRIR (generalmente 2 m.). Estos filtros se calculan previamente y se almacenan en un archivo en forma de tabla de consulta. Nos referimos a este proceso como corrección HRIR porque se aplica en serie con el HRIR seleccionado e interpolado en las etapas anteriores. En este caso, para minimizar los artefactos auditivos que se producen en entornos dinámicos, se lleva a cabo una fusión cruzada (o cross-fading en inglés) lineal de los coeficientes del filtro de diferencias.

B.2.2 Simulación del camino reverberante

En el camino de reverberación, la simulación de distancias se lleva a cabo con la misma técnica descrita anteriormente para el camino directo. En este caso, las fuentes no se espacializan de forma separada, como en el camino directo, sino que se procesan todas juntas siguiendo una aproximación Ambisónica, tal y como se explica a continuación.

- **Ambisónico Virtual.**

Haciendo uso de una aproximación Ambisónica virtual (M. Noisternig et al., 2003), las señales de audio se codifican juntas en un formato Ambisónico de primer orden. Esto mantiene parte de la información espacial de las fuentes, aunque con una baja resolución espacial. De este modo, la información direccional de todo el campo sonoro se incluye en los cuatro primeros canales Ambisónicos (W, X, Y, Z), que luego se decodifican en una serie de altavoces virtuales situados en un conjunto de posiciones conocidas. Por último, las señales de los altavoces virtuales se convierten al dominio binaural convolucionándolas con las BRIRs correspondientes a cada una de las posiciones de los altavoces. Las convoluciones se llevan a cabo utilizando un algoritmo de convolución uniformemente particionada con Overlap-Save (UPOLS) (Wefers, 2015). Esta técnica, basada en FFTs, particiona la respuesta al impulso del filtro en un conjunto de bloques

del mismo tamaño que el buffer de entrada de audio, lo que permite que las operaciones de convolución se realicen de manera muy eficiente, lo cual es importante en el caso de las BRIRs, que pueden ser muy largas.

B.3 Evaluación del 3DTI Toolkit-BS

Para evaluar el correcto funcionamiento de los algoritmos implementados y el rendimiento del 3DTI Toolkit-BS, se han realizado una serie de pruebas y una evaluación objetiva. Un breve resumen de los resultados se muestra en las siguientes subsecciones.

B.3.1 Evaluación de la técnica de interpolación

La técnica de interpolación implementada por el 3DTI Toolkit-BS se ha evaluado cogiendo una HRTF de una base de datos y eliminando una HRIR de una de las posiciones ya conocidas. Seguidamente, esta posición ya conocida se compara con una HRIR en la misma posición, pero calculada mediante el proceso de interpolación. La interpolación se ha realizado con HRIRs que incluyen el retardo inicial (HRIRs no alineadas) y con las que no incluyen dicho retardo (HRIRs alineadas). La Figura 87 muestra la comparación de las tres HRIRs.

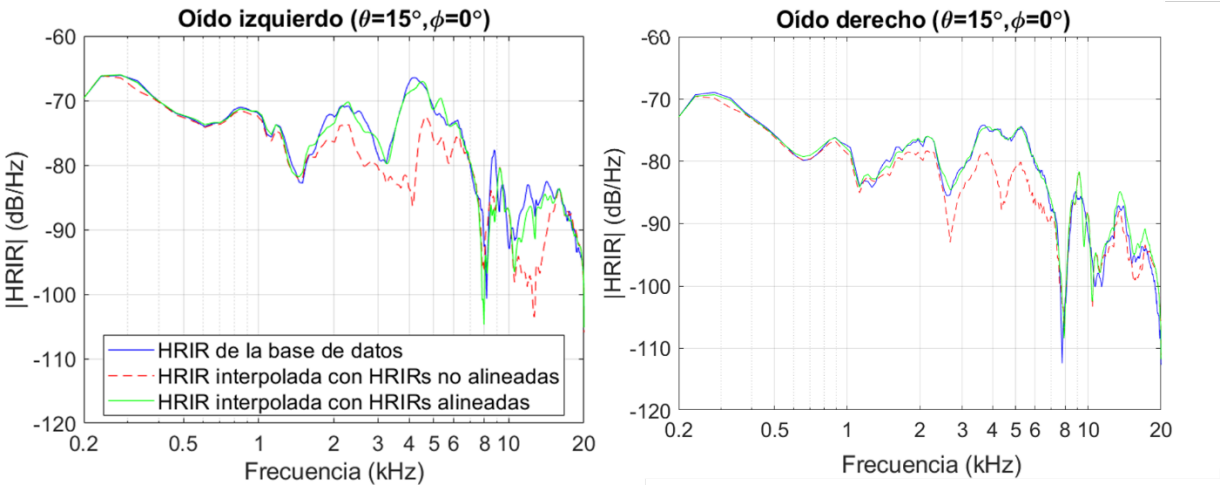


Figura 88. Densidad espectral de potencia para una señal de barrido situada a 15° de acimut y 0° de elevación, para el oído izquierdo y derecho, comparando tres condiciones: (1) HRIR original de la base de datos (línea azul), (2) HRIR interpolada usando HRIRs no alineados (línea roja punteada), y (3) HRIR interpolada usando HRIRs alineados (línea verde). Se ha utilizado la HRTF 1008 de la base de datos LISTEN.

Como puede observarse, la interpolación utilizando HRIRs no alineadas produce una coloración importante debido al efecto de filtrado de peine, lo que hace que aparezcan valles adicionales en diferentes frecuencias. Esto no ocurre cuando utilizamos las HRIRs alineadas, donde la HRIR interpolada es muy similar la HRIR original de la base de datos.

B.3.2 Evaluación de la simulación de campo cercano

Para simular las fuentes situadas en el campo cercano del oyente, el 3DTI Toolkit-BS implementa una corrección de la HRTF. Siguiendo el trabajo de Duda y Martens (1998), se ha creado un conjunto de filtros, basados en un modelo de cabeza esférica (SHM), para simular los incrementos de ILD cuando la fuente se acerca al oyente para todo el rango de frecuencias. El trabajo de Duda y Martens se presenta en la Figura 88.

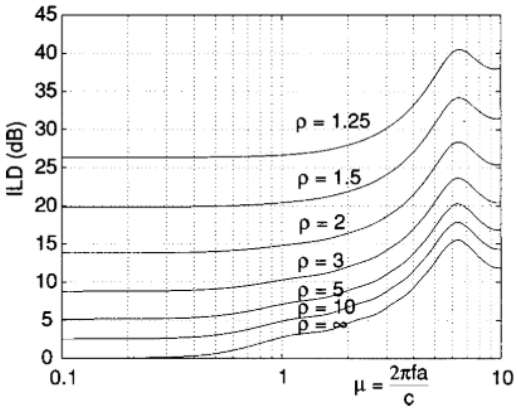


Figura 89. Imagen de Duda & Martens (1998). ILD para una fuente situada en (100°, 0°).ρ se calcula como d/a y es la distancia de la fuente al centro de la cabeza del oyente (d) normalizada con el radio de la cabeza de este (a). El eje horizontal es la frecuencia normalizada utilizando el radio de la cabeza del oyente (a) y la velocidad del sonido (c), lo que significa que el valor 1 corresponde a una longitud de onda igual al radio de la cabeza.

Para comprobar que el 3DTI Toolkit-BS sigue este modelo, la Figura 89 presenta el comportamiento de la herramienta con la simulación de campo cercano desactivada y activada.

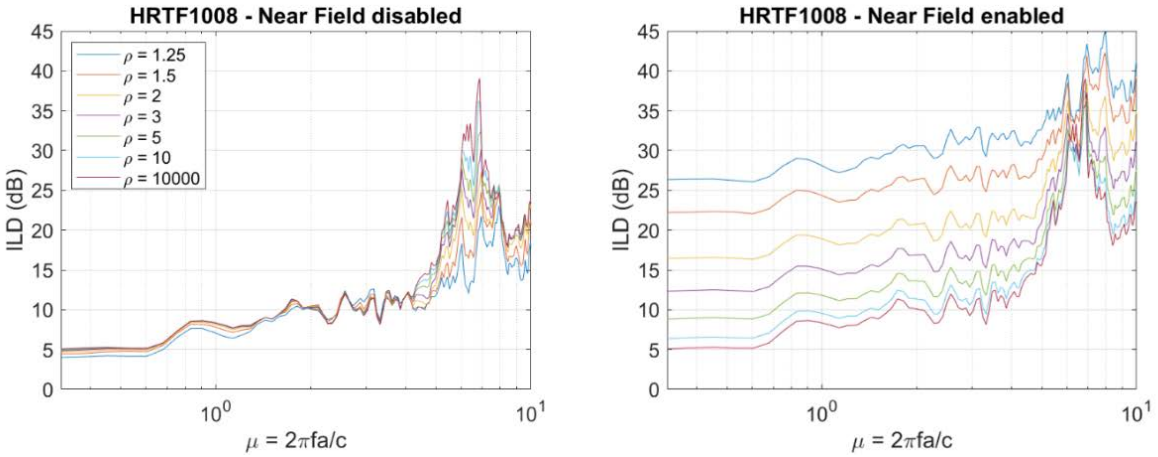


Figura 90. IDL (dB) para la HRIR 1008 para una fuente en (100°, 0°). Simulación de campo cercano desactivada en la figura de la izquierda y activada en la derecha.

Cuando se habilita el campo cercano (Figura 89 derecha) se observa el mismo efecto que en el de SHM (Figura 88), donde el ILD aumenta para todas las frecuencias y es mayor a medida que la fuente se acerca al oyente (ρ cercano a 1). Además, se puede observar la misma forma en las curvas para los valores de μ entre 6 y 8, que corresponden a 3,7 kHz y 5 kHz. Este efecto también aparece en el gráfico en el que se desactiva el campo cercano, lo que nos hace pensar que es un efecto causado por el filtrado de la cabeza caracterizado en la HRTF y no depende de la distancia al oyente.

B.3.3 Evaluación de la técnica de simulación con BRIR

En esta sección se evalúa la BRIR utilizada por el Toolkit en diferentes posiciones de la fuente, los cuales pueden coincidir o no con la ubicación de los altavoces virtuales utilizadas en la implementación del Ambisónico virtual. Para ello, se han creado unas BRIRs medidas en algunas direcciones del plano horizontal, con acimuts entre 0° y 90° cada 10 grados. Estas BRIRs sintetizadas se han comparado con la respuesta al impulso del Toolkit (llamada aquí Toolkit BRIR). Para obtener las BRIRs del Toolkit, este ha sido estimulado con impulso (delta) situada en las mismas posiciones que las BRIR medidas. Para comparar ambas señales, se ha realizado una correlación cruzada y los resultados se muestran en la Figura 90.

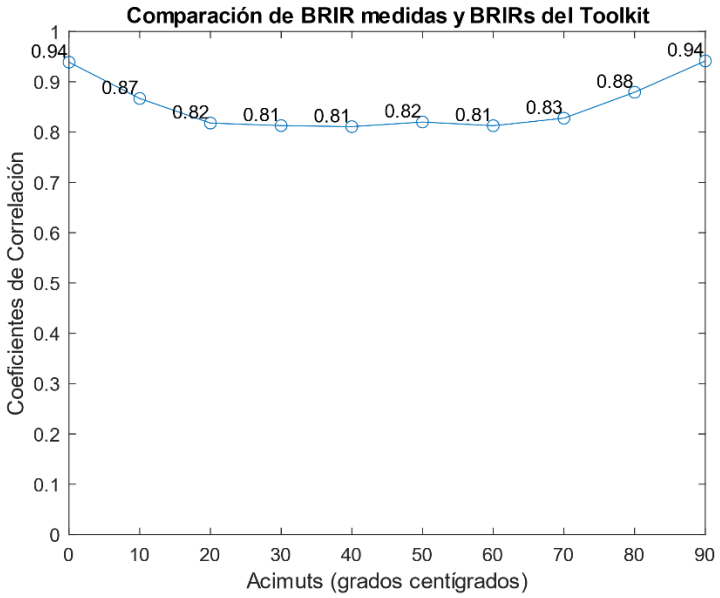


Figura 91. Correlación cruzada entre una BRIR medida y la simulado por el 3DTI Toolkit-BS para el oído izquierdo y diferentes acimuts en el plano horizontal.

La correlación máxima se alcanza cuando comparamos la BRIR medida con la BRIR del Toolkit en las posiciones de los altavoces virtuales (0° y 90°). A medida que el valor del acimut se aleja de estas posiciones, el coeficiente de correlación disminuye, tomando su valor mínimo entre 40° y 50°, las posiciones más alejadas de los altavoces virtuales. Aun así, estos valores están en torno a 0,8, lo que sugiere que, incluso en el peor de los casos, existe una buena correlación entre la BRIR medida y la que obtenemos del Toolkit. Además, debemos mencionar que el uso de una aproximación Ambisónica hace que la BRIR del Toolkit en las posiciones de los altavoces virtuales y la BRIR medida en esta posición no sean exactamente iguales (la correlación es inferior a 1).

B.3.4 Reducción de los artefactos no lineales

En el diseño y desarrollo de los algoritmos del 3DTI Toolkit-BS se ha prestado especial atención a reducir las distorsiones que aparecen en las señales de audio en situaciones dinámicas. En esta sección se resume la evaluación del comportamiento de la herramienta en estas situaciones y se muestran cómo se minimizan dichas distorsiones. El análisis se basa en la medición de las distorsiones no lineales causadas por la no invarianza en el tiempo del sistema, cuando una fuente se mueve a diferentes velocidades. Para ello, se estimula el sistema con una señal compuesta por tres tonos representativos

(859,65 Hz, 4298 Hz y 8596 Hz), estimando el porcentaje de *energía fuera de banda* (EoB), siguiendo el enfoque descrito en (Belloch et al., 2013).

La EoB se ha calculado para diferentes combinaciones de distancia y velocidades angulares, obteniendo los resultados que se muestran en la Figura 91. La Figura 91a y la Figura 91b muestran la EoB cuando el único procesado es la convolución con las HRIRs alineadas. Puede observarse que al aumentar la velocidad se produce un aumento de la EoB, como era de esperar, pero incluso a alta velocidad (9 rad/s) la distorsión global es relativamente pequeña. Las Figura 91c y Figura 91b se refieren al procesado que incluye la convolución y la simulación del ITD. La distorsión estimada muestra un aumento muy ligero si se compara con la condición anterior, a pesar de que se aplica un retardo de hasta 30 muestras, que disminuye dinámicamente hasta 0 y vuelve a subir en el otro oído por cada vuelta completa que recorre la fuente. Por último, las Figura 91e y Figura 91f incluyen el procesado de la convolución, simulación del ITD y corrección de ILD de campo cercano, para distancias menores de 2m. En este caso, la EoB no aumenta al añadir la corrección de campo cercano; por el contrario, se observa una pequeña disminución de la distorsión general. Esto se debe probablemente al hecho de que la distorsión no lineal es mayor en el oído contralateral, donde los filtros de corrección de campo cercano aplican una mayor atenuación. En cualquier caso, la distorsión introducida por el comportamiento dinámico de estos filtros puede considerarse insignificante.

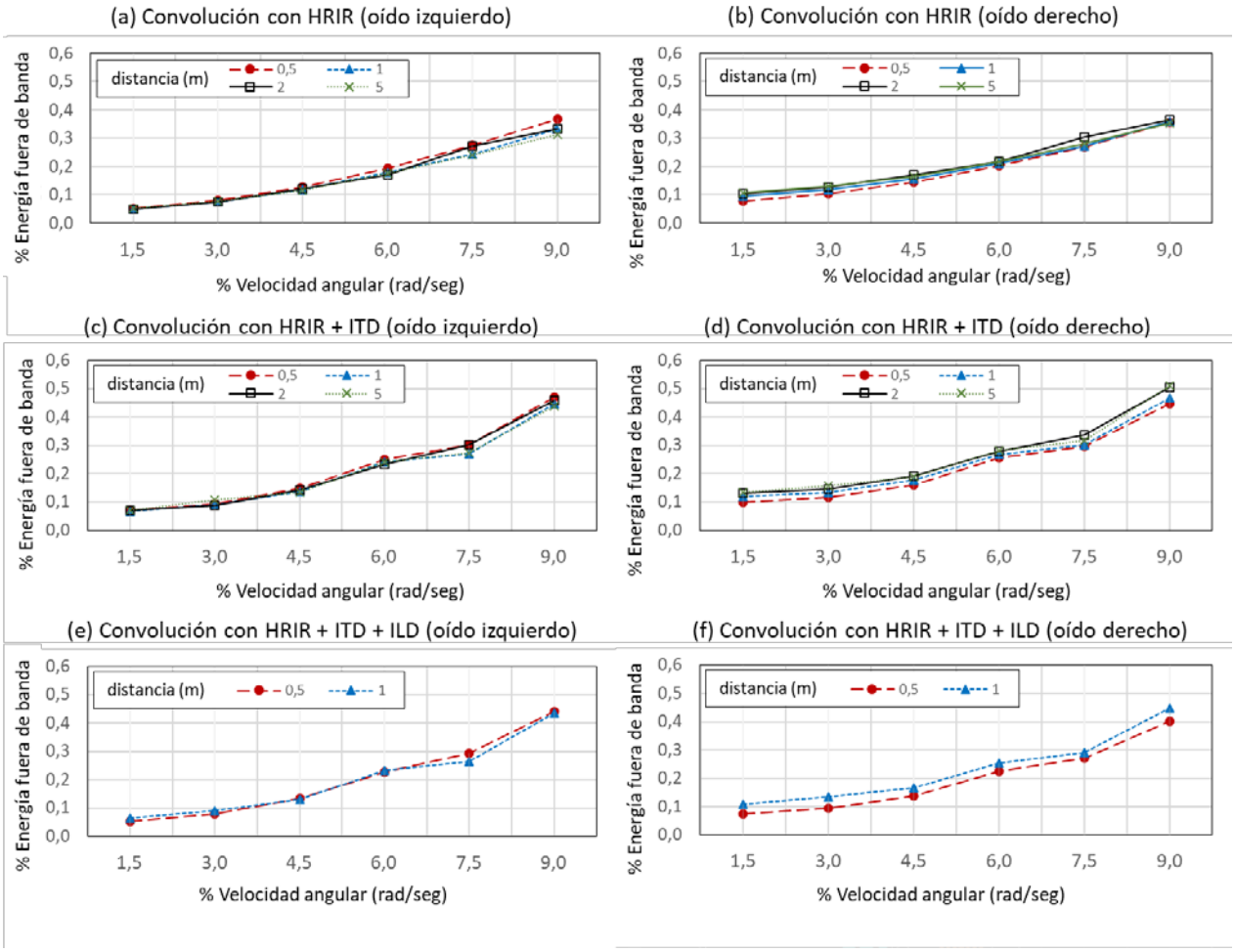


Figura 92. Energía fuera de banda producida por el proceso de espacialización para diferentes configuraciones. (a) y (b): sólo convolución de HRIRs alineadas, (c) y (d): convolución con HRIRs alineadas y simulación del ITD, (e) y (f) convolución con HRIRs alineadas y simulación de ITD e ILDs. Cada configuración para ambos oídos. El tamaño del *frame* es de 512 muestras, frecuencia de muestreo 44,1 kHz y HRTF número 1013 de la base de datos LISTEN.

B.3.5 Rendimiento en tiempo real

El procesado del 3DTI Toolkit-BS tiene que compartir el tiempo de la CPU con otros procesos, por lo que el tiempo que el Toolkit tarda en procesar un *frame* es realmente el dedicado a producir el audio espacializado más el tiempo que tardan los otros procesos en caso de que se produzcan interrupciones. Si este tiempo total excede el tiempo del *frame*, el *frame* de audio completo será descartado, produciendo un artefacto audible. En esta sección se muestra el rendimiento del Toolkit según el porcentaje del tiempo total del *frame* utilizado. La Figura 92 muestra el porcentaje de tiempo de un *frame* que

el Toolkit utiliza para procesar diferentes números de fuentes, en la simulación del camino directo. La medida se ha repetido para diferentes tamaños de *frame*. Como era de esperar, este porcentaje aumenta linealmente con el número de fuentes, lo que permite renderizar un número relativamente grande de fuentes en un ordenador de sobremesa común. También hay algunos valores atípicos que podemos atribuir a la presencia de las interrupciones comentadas anteriormente.

Del mismo modo, el gráfico presentado en la Figura 93 muestra el ciclo de trabajo para el proceso de reverberación. Se observa que este proceso es casi independiente del número de fuentes implicadas y que a medida que aumenta el tamaño del *frame* el tiempo de procesado disminuye. Una cuidadosa selección del tamaño del *frame* ofrece la posibilidad de renderizar con baja latencia reverberaciones muy largas, por ello es conveniente buscar un compromiso entre la latencia y el coste computacional soportado.

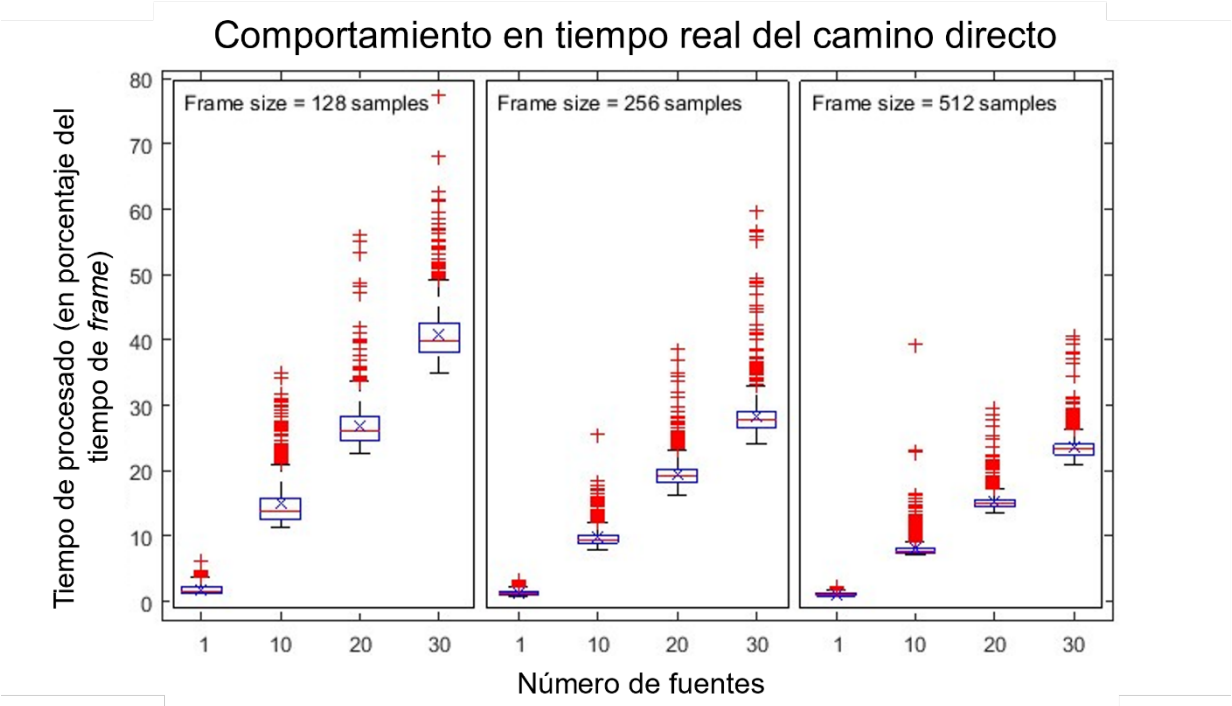


Figura 93. Rendimiento del procesado del camino directo en función del tamaño del *frame*. El eje horizontal muestra el número de fuentes y el vertical el porcentaje del tamaño total del *frame*.

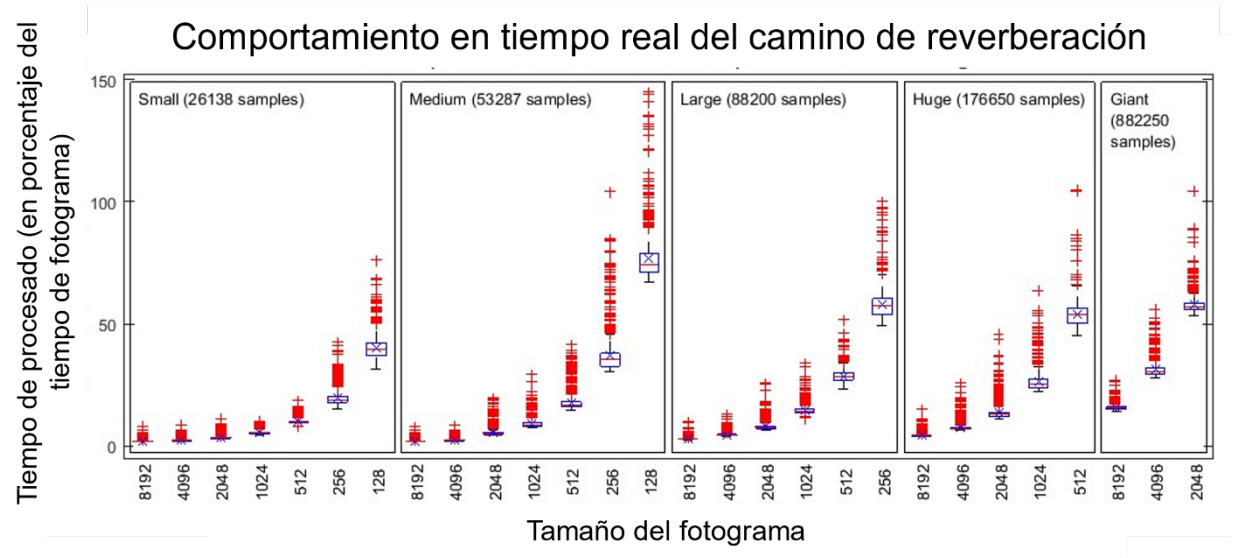


Figura 94. Rendimiento del proceso de reverberación en función de la longitud del BRIR. El eje horizontal muestra diferentes tamaños de *frame* y el eje vertical el porcentaje del tiempo total del *frame*.

B.4 Estudio del impacto de HRTF no individualizadas en la inteligibilidad del habla

En este apartado se resume un estudio realizado cuyo objetivo principal es, utilizando el 3DTI Tollkit-BS, analizar el impacto de diferentes HRTFs no individualizadas sobre la inteligibilidad del habla en un contexto de *cocktail party* en un entorno de RV. Si el sistema atencional humano utiliza la propia experiencia del oyente con su HRTF individual para mejorar el reconocimiento del habla, en un entorno en el que el objetivo y los enmascaradores están situados en diferentes posiciones, deberíamos ser capaces de encontrar experimentalmente un efecto de la elección de la HRTF en dicho reconocimiento del habla. Además, si la HRTF es una característica idiosincrásica de cada oyente, también deberíamos encontrar que este efecto es diferente para diferentes sujetos. Las hipótesis del estudio son las siguientes:

- H1: Existe un efecto significativo de la elección de la HRTF sobre el reconocimiento del habla en el contexto de *coktail party* en un escenario virtual. Es decir, para un sujeto determinado, diferentes HRTFs proporcionan diferentes rendimientos en términos de reconocimiento del habla de las palabras objetivo, en condiciones de enmascaramiento con ruido.

- H2: El efecto de una HRTF determinada en el reconocimiento del habla es diferente para distintos sujetos; por lo tanto, no hay HRTFs que sean universalmente mejores o peores que otras cuando se evalúan en esta tarea específica.

B.4.1 Recogida y análisis de datos

Durante el experimento se estimó la SRT (siglas del término en inglés Speech Reception Threshold) para evaluar el efecto de un conjunto de HRTFs no individualizadas sobre la inteligibilidad del habla. Se colocó un sonido objetivo de voz frente al oyente (0° acimut), y dos enmascaradores de ruido a la derecha y a la izquierda (90° y 270° acimut), todos ellos en el plano horizontal. La simulación de audio en 3D se realizó con el 3DTI Toolkit-BS, utilizando 7 HRTFs medidas de la base de datos LISTEN (denominadas HRTF1 - HRTF7) más una HRTF sintética como condición de anclaje (denominada HRTFA). El HRTFA se generó a partir de un modelo de cabeza esférica, que no proporcionaba los efectos de los pabellones auriculares, pero incluía ITD e ILD correspondientes a una cabeza esférica sin orejas. Un total de 22 sujetos participaron en el experimento.

Es importante tener en cuenta que puede haber algunas características de las HRTFs, como diferencias en la relación de potencia entre los lados y el frente, que podrían hacer que algunos HRTFs fueran peores o mejores para todos los participantes en general, independientemente de las diferencias individuales. Por ejemplo, en el caso de una HRTF con una mayor atenuación dentro de las bandas espectrales del habla para fuentes en (0°, 0°), el objetivo se atenuaría más que cuando se utilizan otras HRTFs, lo que daría lugar a una mayor SRT. Teniendo en cuenta la segunda hipótesis del presente estudio (H2), y el objetivo de identificar las diferencias específicas del sujeto cuando se usan diferentes HRTFs, es importante cuantificar las posibles ventajas específicas de cada HRTF, y que afectan a todos los participantes de la misma manera. Para ello, utilizamos el modelo desarrollado por Lavandier & Culling (2010) y posteriormente revisado por Jelfs et al. (2011), incluido en el Matlab Auditory Modeling Toolbox (AMT_JELFS2011, n.d.). Este modelo predice el beneficio (en decibelios) que ofrece cada HRTF según la posición del objetivo y del enmascarador. Este beneficio es utilizado para compensar los valores del SRT obtenido en el experimento, con el fin de eliminar estos componentes específicos de cada HRTF.

B.4.2 Resultados y discusión

Para analizar inicialmente los datos sin tener en cuenta las diferencias individuales, se llevó a cabo un análisis global. Para ello se promediaron los datos recogidos para cada participante (utilizando los SRTs de las 20 sesiones), obteniendo un SRT promedio por HRTF por participante. La distribución de estos datos se muestra en el gráfico de la Figura 94a. La Figura 94b muestra los SRTs promedios de todos los participantes, junto con el intervalo de confianza (IC) del 95% para cada HRTF utilizada en el experimento. Los datos se muestran sin procesar (Raw) y compensados (Compensated).

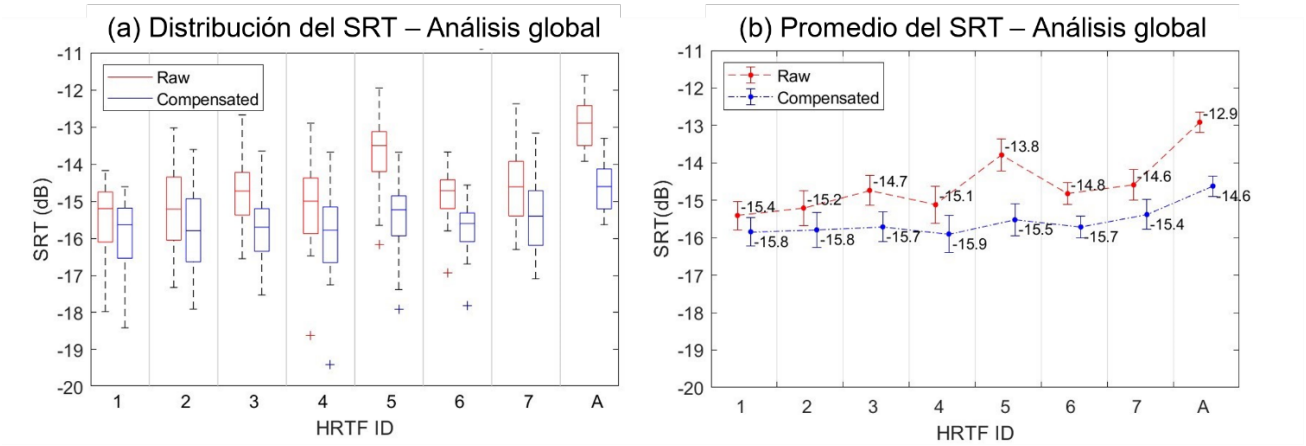


Figura 95. El gráfico de la izquierda muestra la distribución de las medias del SRT de cada participante en todas las sesiones. En cada caja, la marca horizontal central indica la mediana, y los bordes inferior y superior los percentiles 25 y 75, respectivamente. Los «bigotes» se extienden hasta los puntos de datos más extremos no considerados como valores atípicos, los cuales se representan individualmente con un «+». El gráfico de la derecha muestra la media del SRT en las sesiones y los participantes y los IC del 95%.

Al observar los datos sin procesar, se pueden identificar diferencias relevantes entre las distintas HRTFs, en particular las HRTF5 y HRTFA. Es evidente que, una vez aplicada la compensación, estas diferencias son menos acentuadas. Esto se corrobora con un análisis de varianza unidireccional (ANOVA). Los SRT sin procesar mostraron un efecto significativo de la HRTF sobre el SRT cuando se incluyó la HRTFA [$F(7; 168)=16{,}7861$; $p < 0{,}001$] y también cuando se eliminó del conjunto de datos [$F(6; 147) = 6{,}3972$; $p < 0{,}001$]. Los SRTs compensados mostraron un efecto significativo sólo cuando se incluyó la HRTFA [$F(7; 168)=4{,}1892$; $p < 0{,}001$] pero no cuando se eliminó del conjunto de datos [$F(6; 147)= 0{,}76083$; $p=0{,}602$].

En cuanto a la HRTFA, se puede dar una explicación plausible si se tiene en cuenta que no incluye ninguna señal espectral monoaural, como el resto de HRTFs, ya que el

modelo no tiene en cuenta el pabellón auditivo ni otras características antropométricas relevantes más allá de una cabeza esférica aproximada. Por otro lado, el análisis de los datos compensados nos da algunas pistas para entender también estos resultados, ya que tanto el HRTF5 como el HRTFA dieron lugar a un factor de compensación significativamente mayor, lo que nos hace pensar que estas dos HRTFs son «universalmente» peores en términos de SRT si se comparan con los demás.

Seguidamente, los datos se han analizado por separado para cada individuo. Para el 82% de los participantes (18 de 22) se encontró un efecto significativo de la HRTF sobre la SRT no procesada. Esto sugiere que la elección de la HRTF puede tener un impacto significativo en los SRTs para la gran mayoría de los participantes evaluados, lo que confirma la primera de nuestras hipótesis iniciales (H1). Por otro lado, se ha encontrado una reducción del número de participantes que muestran un efecto significativo de la HRTF al eliminar la HRTFA de la comparación (de 18 a 9 de 22), pero es importante observar cómo todavía se puede encontrar un efecto significativo para el 41% de los participantes. Aunque con menos fuerza, este resultado sigue apoyando H1, por lo tanto esto nos sugiere que, para un sujeto dado, diferentes HRTFs proporcionan diferentes rendimientos en términos de reconocimiento del habla. Las comparaciones *post hoc* por pares para las diferentes condiciones de HRTFs se llevaron a cabo mediante la prueba de diferencia mínima significativa (LSD) de Fisher. La Figura 95a muestra el número de participantes con diferencias significativas en cada comparación por pares ($p < 0{,}05$) para los datos del SRT sin procesar. Puede observarse que, de nuevo, la HRTFA y la HRTF5 muestran un mayor número de participantes con diferencias significativas.

El análisis de los datos de SRTs compensados da resultados bastante diferentes. El número de participantes que muestran un efecto significativo de la HRTF disminuye a cinco, y a uno cuando se excluye la HRTFA. Al examinar H1, mientras que los datos sin procesar apoyan el hecho de que existe un efecto significativo de la elección de la HRTF para un gran número de individuos, esto no puede evidenciarse tan claramente al examinar los datos compensados. Sin embargo, es cierto que también en este último caso se ha encontrado un cierto número de diferencias significativas entre pares (Figura 95b), donde el número de pares que muestran diferencias significativas es mayor de lo esperado por azar. Para el análisis SRT compensado, el HRTFA es el único que muestra un gran número de participantes con diferencias significativas al compararlo con otras condiciones. No obstante, también cuando se consideran sólo los HRTFs medidos (HRTF1-7), el número de pares con diferencias significativas debe tenerse en cuenta ya que muestra signos claros de que los individuos actúan de forma diferente con diferentes HRTFs.

Una situación simétricamente diferente se encuentra cuando se observa H2; en este caso, los datos compensados ofrecen mejores resultados, si se comparan con los datos no procesados, al apoyar la hipótesis de que no hay HRTFs medidas individualmente que sean universalmente mejores o peores que otras. Sin embargo, a la hora de hacer estas consideraciones, es importante tener en cuenta la naturaleza de la compensación, cuyo objetivo es equilibrar aquellas diferencias que podrían hacer que algunas HRTF sean peores o mejores para la muestra global de participantes, independientemente de las diferencias individuales. Está claro que hay algunas HRTFs que son generalmente mejores o peores que otras cuando se analizan las actuaciones en el reconocimiento virtual del habla en el ruido. Pero también podemos observar que hay características individuales de las HRTFs que permiten a ciertos sujetos un mejor comportamiento, y peor para otros, en el reconocimiento del habla.

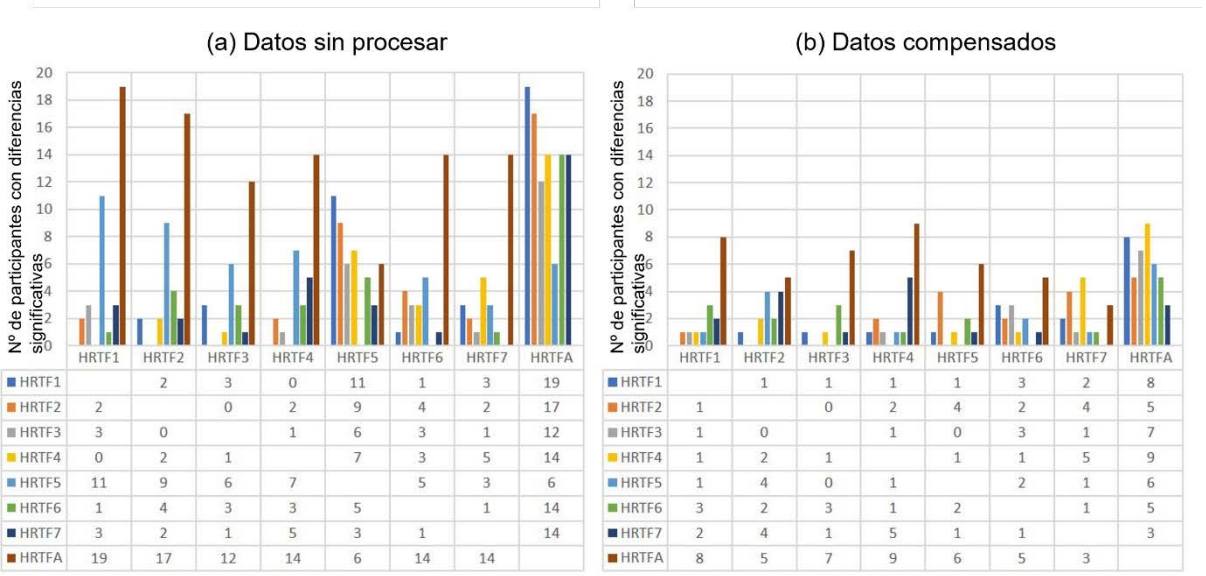


Figura 96. Comparaciones *post hoc* por pares para las diferentes condiciones de HRTF con la prueba de diferencia mínima significativa (LSD). Las tablas indican el número de participantes con diferencias significativas entre la condición HRTF indicada en el encabezado y las de la columna de la izquierda. Los gráficos indican el número de participantes con diferencias significativas entre la condición HRTF indicada en el eje horizontal y la correspondiente al color en la leyenda.

En resumen, con este estudio se ha demostrado que, dentro de las condiciones probadas y observando los datos de SRTs tanto sin procesar como compensados, puede haber un efecto significativo de la elección de la HRTF en el reconocimiento del habla, y este efecto puede ser diferente para diferentes sujetos. Las implicaciones de estos resultados podrían ser relevantes para varias áreas de investigación. A la luz de los resultados, debe tenerse en cuenta que, al modelar la percepción binaural del habla en

el ruido, deben tenerse en cuenta tanto las señales monoaurales como las binaurales; y por tanto, debe considerarse cuidadosamente la elección de la HRTF.

B.5 Conclusiones

Las principales aportaciones de esta tesis de dividen en dos bloques. El primer bloque consiste en el desarrollo y evaluación de una librería de espacialización de audio binaural para entornos virtuales en tiempo real, denominada 3DTI Toolkit-BS. La motivación de este trabajo surgió de la necesidad de una herramienta que funcione como plataforma para realizar experimentos psicoacústicos, para estudiar la percepción auditiva en un entorno virtual. Este tipo de experimentos requiere un conjunto de características que se han tenido en cuenta en el desarrollo del Toolkit que no se han encontrado todas juntas en otras herramientas disponibles y se enumeran a continuación. (1) Modularidad, (2) la inclusión de una serie de algoritmos como la lectura de ficheros SOFA, la interpolación de HRIRs alineadas, la personalización del ITD, la simulación de fuentes en campo cercano y la simulación de la reverberación espacializada. Además de (3) la distribución en código abierto, (4) ser multipataforma, (5) ofrecer una simulación de baja latencia y (6) tener un buen comportamiento en entornos dinámicos sin distorsiones en la señal de audio. Esta herramienta de espacialización ha sido evaluada dentro de la tesis pudiendo comprobar que cumple los requisitos establecidos. Además a día de hoy ya ha sido utilizada en diferentes estudios y aplicaciones fuera del ámbito de esta tesis, como los presentados por Engel et al. (2021), Reyes-Lecuona, et al. (2021), Comunità et al. (2020) and Lerner et al. (2021). Además, existen dos proyectos activos actualmente que hacen uso del 3DTI Toolkit-BS: SONICOM[32] y SAVLab[33]. Estos proyectos garantizan el uso y mantenimiento de la herramienta en los próximos años, así como la introducción de mejoras, nuevos experimentos, publicaciones y una mayor difusión. Por último, destacar que la herramienta está disponible en https://github.com/3DTune-In/3dti_AudioToolki (ultimo acceso en enero de 2022).

El segundo bloque presenta un estudio que demuestra el uso del 3DTI Toolkit-BS como renderizador base para experimentos de psicoacústica y nos ayuda a comprender la influencia de las HRTFs no individualizadas en la inteligibilidad del habla en escenarios de cocktail party. Para ello se midió el umbral de recepción del habla (SRT)

[32] SONICOM es un proyecto financiado por la Unión Europea dentro del programa Horizonte 2020, (no.101017743). Web: https://www.sonicom.eu/ (ultimo acceso en enero de 2022).

[33] SAVLab es un proyecto financiado por el Plan Nacional de I+D.

para diferentes HRTFs en 22 participantes. Además, utilizando el SRT predicho por un modelo de percepción del habla existente, se compensaron los valores medidos en el intento de eliminar los beneficios globales específicos de cada HRTF. Los resultados mostraron diferencias globales significativas entre los SRTs medidos utilizando diferentes HRTFs, en consonancia con los resultados predichos por el modelo. También se encontraron diferencias individuales entre los participantes, relacionadas con sus SRTs obtenidos utilizando diferentes HRTFs, pero su importancia se redujo después de la compensación. Las implicaciones de estos resultados son relevantes para varias áreas de investigación relacionadas con la audición espacial y la percepción del habla, sugiriendo que, en un escenario virtual, para la inteligibilidad del habla dentro de un entorno ruidoso, la elección de la HRTF para cada individuo debe ser considerada cuidadosamente. Este estudio abre las puertas a múltiples líneas futuras, en algunas de las cuales ya se está trabajando, como es por un lado el estudio de la influencia de los HRTFs (pero esta vez personalizados para cada sujeto) en la inteligibilidad del habla cuando la separación de las fuentes objetivo y de enmascaramiento se encuentran en el plano vertical y, por otro lado, la repetición de este mismo estudio pero en entornos dinámicos donde tanto las fuentes como el oyente pueden estar en movimiento.

Bibliography

Abhijit Patait. (2017). *VRWorks Audio SDK in-depth | NVIDIA Developer*. https://developer.nvidia.com/vrworks-audio-sdk-depth

Ahrens, A., Cuevas-Rodriguez, M., & Brimijoin, W. O. (2021). Speech intelligibility with various head-related transfer functions: A computational modelling approach. *JASA Express Letters*, *1*(3), 034401. https://doi.org/10.1121/10.0003618

Algazi, R., Duda, R. O., Thompson, D. M., & Algazi, V. R. (2002). *The Use of Head-and-Torso Models for Improved Spatial Sound Synthesis*. www.aes.org.

Algazi, V., & Duda, R. (2011). Headphone-based spatial sound. *IEEE Signal Processing Magazine*, *28*(1), 33–42. https://doi.org/10.1109/MSP.2010.938756

Algazi, V. Ralph, Avendano, C., & Duda, R. O. (2001). Estimation of a Spherical-Head Model from Anthropometry. *Journal of the Audio Engineering Society*, *49*(6), 472–479.

Algazi, V. Ralph, Avendano, C., & Duda, R. O. (2002). Elevation localization and head-related transfer function analysis at low frequencies. *The Journal of the Acoustical Society of America*, *109*(3), 1110–1122. https://doi.org/10.1121/1.1349185

Algazi, V. Ralph, Duda, R. O., Duraiswami, R., Gumerov, N. A., & Tang, Z. (2002). Approximating the head-related transfer function using simple geometric models of the head and torso. *The Journal of the Acoustical Society of America*, *112*(5), 2053–2064. https://doi.org/10.1121/1.1508780

Algazi, V.R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001). The CIPIC HRTF database. *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, *October*, 99–102. https://doi.org/10.1109/ASPAA.2001.969552

Algazi, V.Ralph, Avendano, C., & Thompson, D. (1999). Dependenceof Subject and MeasurementPosition in Binaural Signal Acquisition. *Journal of the Audio Engineering Society*, *47*(11), 937–947.

Algazi, V.Ralph, Divenyi, P. L., Martinez, V. A., & Duda, R. O. (1997). Subject dependent transfer functions in spatial hearing. In Anon (Ed.), *Proceedings of the 1997 40th Midwest Symposium on Circuits and Systems. Part 1 (of 2)* (Vol. 2, pp.

877–880). IEEE.

Allen, J. B., & Berkley, D. A. (1979). Image method for efficiently simulating small-room acoustics. *Journal of the Acoustical Society of America*, *65*(4), 943–950. https://doi.org/10.1121/1.382599

Alon, D. Lou, Ben-Hur, Z., Rafaely, B., & Mehra, R. (2018). Sparse Head-Related Transfer Function representation with spatial aliasing cancellation. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 6792–6796.

Alon, D. Lou, Cuevas-Rodríguez, M., Mehra, R., & Robinson, P. W. (2020). *Compensating for effects of headset on head related transfer functions* (Patent No. 10798515). https://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&p=1&u=%2Fnetahtml%2FPTO%2Fsearch-bool.html&r=1&f=G&l=50&co1=AND&d=PTXT&s1=10%2C798%2C515&OS=10%2C798%2C515&RS=10%2C798%2C515

AMT_JELFS2011. (n.d.). *JELFS2011 - Predicted binaural advantage for speech in reverberant conditions*. Retrieved May 11, 2020, from http://amtoolbox.sourceforge.net/amt-0.9.9/doc/models/jelfs2011.php

Andreopoulou, A., & Katz, B. F. (2016a). Investigation on Subjective HRTF Rating Repeatability. *140th Audio Engineering Society Convention*, 1–10. http://www.aes.org/e-lib/browse.cfm?elib=18295

Andreopoulou, A., & Katz, B. F. G. (2016b). Subjective HRTF evaluations for obtaining global similarity metrics of assessors and assessees. *Journal on Multimodal User Interfaces*, *10*(3), 259–271. https://doi.org/10.1007/s12193-016-0214-y

ANSI. (1997). *ANSI/ASA S3.5-1997 (R2017) - Methods for Calculation of the Speech Intelligibility Index*. https://webstore.ansi.org/standards/asa/ansiasas31997r2017

ARI HRTF Database. (n.d.). Retrieved November 4, 2018, from https://www.kfs.oeaw.ac.at/index.php?view=article&id=608&lang=en

Ashmore, J. (2008). Cochlear outer hair cell motility. In *Physiological Reviews* (Vol. 88, Issue 1, pp. 173–210). American Physiological Society. https://doi.org/10.1152/physrev.00044.2006

Aspöck, L., Mösch, L., Stienen, J., & Vorländer, M. (2019). Auralization of interactive virtual scenes containing numerous sound sources. *23rd International Congress on Acoustics*, 6401–6408.

Audio, F. (2020). *Facebook Audio 360 documentation*. https://facebookincubator.github.io/facebook-360-spatial-workstation/Documentation/SpatialWorkstation/SpatialWorkstation.html

Baalman, M. A. (2005). Updates of the WONDER software interface for using Wave Field Synthesis. *LAC2005 Proceedings*, 69–72.

Batteau, D. W. (1967). The role of the pinna in human localization. *Royal Society of London*, *168*(1011), 158–180.

Baumgartner, R., Majdak, P., & Laback, B. (2013). Assessment of Sagittal-plane Sound Localization Performance in Spatial-audio Applications. In J. Blauert (Ed.), *The Technology of Binaural Listening* (Issue January, pp. 1–511). Springer, Berlin–Heidelberg–New York NY. https://doi.org/10.1007/978-3-642-37762-4

Baumgartner, Robert, Majdak, P., & Laback, B. (2014). Modeling sound-source localization in sagittal planes for human listeners. *The Journal of the Acoustical Society of America*, *136*(2), 791–802. https://doi.org/10.1121/1.4887447

Begault, D. R. (1994). *3-D Sound for Virtual Reality and Multimedia*. Academic Press Professional, Inc. http://www.jstor.org/stable/3680997?origin=crossref

Begault, D. R., & Wenzel, E. M. (1990). Techniques and Applications for Binaural Sound Manipulation. *The International Journal of Aviation Psychology*, *2*(1), 1–22. https://doi.org/10.1207/s15327108ijap0201_1

Begault, D. R., & Wenzel, E. M. (1993). Headphone localization of speech. *Human Factors*, *35*(2), 361–376. https://doi.org/10.1177/001872089303500210

Begault, D. R., Wenzel, E. M., & Anderson, M. R. (2001). Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source. *Journal of the Audio Engineering Society*, *49*(10), 904–916.

Békésy, G. von, & Wever, E. G. (1960). *Experiments in Hearing*. McGraw-Hill.

Belloch, J. A., Ferrer, M., Gonzalez, A., Martinez-Zaldivar, F. J., & Vidal, A. M. (2013). Headphone-based virtual spatialization of sound with a GPU accelerator. *AES: Journal of the Audio Engineering Society*, *61*(7–8), 546–561.

Benichoux, V., Rébillat, M., & Brette, R. (2016). On the variation of interaural time differences with frequency. *The Journal of the Acoustical Society of America*, *139*(4), 1810–1821. https://doi.org/10.1121/1.4944638

Beranek, L., & Martin, D. W. (1996). Concert & Opera Halls: How They Sound. *The Journal of the Acoustical Society of America*, *99*, 2637. https://doi.org/10.1121/1.414882

Bergstrom, I., Azevedo, S., Papiotis, P., Saldanha, N., & Slater, M. (2017). The Plausibility of a String Quartet Performance in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics*, *23*(4), 1352–1359. https://doi.org/10.1109/TVCG.2017.2657138

Berkhoul, A. J., De Vries, D., & Vogel, P. (1993). Acoustic control by wave field synthesis. *Journal of the Acoustical Society of America*, *93*(5), 2764–2778. https://doi.org/10.1121/1.405852

Bernschütz, B. (2013). A Spherical Far Field HRIR/HRTF Compilation of the Neumann

KU 100. *Fortschritte Der Akustik -- AIA-DAGA 2013*, 592--595. http://www.audiogroup.web.fh-koeln.de/FILES/AIA-DAGA2013_HRIRs.pdf

Best, V., Baumgartner, R., Lavandier, M., Majdak, P., & Kop, N. (2020). *Sound Externalization : A Review of Recent Research*. https://doi.org/10.1177/2331216520948390

Blauert, J., & Cobben, W. (1978). Some consideration of binaural cross correlation analysis. *Acta Acustica United with Acustica*, *32*(2), 96–104.

Blauert, Jens. (1997). *Spatial hearing : the psychophysics of human sound localization*. MIT Press. https://mitpress.mit.edu/books/spatial-hearing

Blauert, Jens. (1983). Spatial Hearing - The Psychophysics of Human of sound. *Journal of the Acoustical Society of America*, *77*(1), 334–335. https://doi.org/doi:http://dx.doi.org/10.1121/1.392109

Blauert, Jens. (1994). An introduction to binaural technology. In R. H. Gilkey & T. R. Anderson (Eds.), *Binaural and spatial hearing in real and virtual environments* (pp. 593–609). Lawrence Erlbaum Associates, Inc.

Blauert, Jens. (2013). The Technology of Binaural Listening. In *The Technology of Binaural Listening, Modern Acoustics and Signal Processing*. https://doi.org/10.1007/978-3-642-37762-4

Blauert, Jens, Bfiggen, M., Hartun, H., Bronkhorst, A. W., Drullmarm2, R., Reynaud, G., Pellieux, L., Fiebber, W., & Sottek4, R. (1998). *The AUDIS Catalog of Human HRTFs*. 2901–2902.

Blauert, Jens, Braasch, J., Buchholz, J., Colburn, H. S., Jekosch, U., Kohlrausch, A., Mourjopoulos, J., Pulkki, V., & Raake, A. (2010). Aural assessment by means of binaural algorithms — The AABBA project —. *ISAAR 2009: Binaural Processing and Spatial Hearing, August 2009*.

Blauert, Jens, Lehnert, H., Sahrhage, J., & Strauss, H. (2000). An interactive virtual-environment generator for psychoacoustic research. I: Architecture and implementation. *Acustica*, *86*(1), 94–102.

Blue Ripple Sound. (2016). *Rapture3D*. http://www.blueripplesound.com/

Blum, A., Warusfel, O., & Katz, B. F. G. (2004). *Eliciting adaptation to non-individual HRTF spectral cues with multi-modal training presence Eliciting adaptation to non-individual HRTF spectral cues with multi-modal training*. https://www.researchgate.net/publication/280756922

Bormann, K. (2005). Presence and the utility of audio spatialization. *Presence: Teleoperators and Virtual Environments*, *14*(3), 278–297. https://doi.org/10.1162/105474605323384645

Borß, C., & Martin, R. (2009). *An Improved Parametric Model for Perception-Based Design of Virtual Acoustics*. Audio Engineering Society.

Braasch, J., & Hartung, K. (2002). Localization in the presence of a distracter and reverberation in the frontal horizontal plane. *Psychoacoustical Data. Acta Acustica United with Acustica*, *88*(6), 942–955.

Bradley, J. S., Sato, H., & Picard, M. (2003). On the importance of early reflections for speech in rooms. *The Journal of the Acoustical Society of America*, *113*(6), 3233. https://doi.org/10.1121/1.1570439

Breebaart, J., van de Par, S., & Kohlrausch, A. (1999). The contribution of static and dynamically varying ITDs and IIDs to binaural detection. *The Journal of the Acoustical Society of America*, *106*(2), 979–992. https://doi.org/10.1121/1.427110

Brinkmann, F., Dinakaran, M., Pelzer, R., Grosche, P., Voss, D., & Weinzierl, S. (2019). A cross-evaluated database of measured and simulated HRTFs including 3d head meshes, anthropometric features, and headphone impulse responses. *AES: Journal of the Audio Engineering Society*, *67*(9), 705–718. https://doi.org/10.17743/jaes.2019.0024

Brinkmann, F., Lindau, A., Weinzierl, S., Van De Par, S., Müller-Trapet, M., Opdam, R., & Vorländer, M. (2017). A high resolution and full-spherical head-related transfer function database for different head-above-torso orientations. *AES: Journal of the Audio Engineering Society*, *65*(10), 841–848. https://doi.org/10.17743/jaes.2017.0033

Brinkmann, F., Roden, R., Lindau, A., & Weinzierl, S. (2015). Audibility and Interpolation of Head-Above-Torso Orientation in Binaural Technology. *IEEE Journal on Selected Topics in Signal Processing*, *9*(5), 931–942. https://doi.org/10.1109/JSTSP.2015.2414905

Brix, S., Melchior, F., Roder, T., Wabnik, S., & Riegel, C. (2003). *Authoring Systems for Wave Field Synthesis Content Production*. Audio Engineering Society.

Bronkhorst, A. W., & Plomp, R. (1988). The effect of head-induced interaural time and level differences on speech intelligibility in noise. *J.Acoustic Soc.America*, *83*(April). https://doi.org/10.1121/1.2024170

Bronkhorst, A. W., & Plomp, R. (1992). Effect of multiple speechlike maskers on binaural speech recognition in normal and impaired hearing. *The Journal of the Acoustical Society of America*, *92*(6), 3132–3139. https://doi.org/10.1121/1.404209

Bronkhorst, Adelbert W. (1995). Localization of real and virtual sound sources. *Journal of the Acoustical Society of America*, *98*(5), 2542–2553. https://doi.org/10.1121/1.413219

Bronkhorst, Adelbert W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, *2015*, 1465–1487. https://doi.org/10.3758/s13414-015-0882-9

Bronkhorst, Adelbert W., & Houtgast, T. (1999). Auditory distance perception in rooms.

Nature, *397*(6719), 517–520. https://doi.org/10.1038/17374

Bronkhorst, Adelbert W. (2000). The Cocktail Party Phenomenon: A Review of Research on Speech Intelligibility in Multiple-Talker Conditions. *Acta Acust. United with Acust.*, *86*(October), 117–128. https://doi.org/10.1306/74D710F5-2B21-11D7-8648000102C1865D

Brown, C. P., & Duda, R. O. (1997). An efficient HRTF model for 3-D sound. *Proceedings of 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, *Ild*, 4. https://doi.org/10.1109/ASPAA.1997.625596

Brungart, D., Kordik, A. J., & Simpson, B. D. (2006). Effects of headtracker latency in virtual audio displays. *Ournal of the Audio Engineering Society*, *54*(1/2), 32–44.

Brungart, D. S., & Rabinowitz, W. M. (1999). Auditory localization of nearby sources. Head-related transfer functions. *The Journal of the Acoustical Society of America*, *106*(3), 1465–1479. https://doi.org/10.1121/1.427180

Brungart, D.S. (1999). Auditory localization of nearby sources. III. Stimulus effects. *The Journal of the Acoustical Society of America*, *106*(6), 3589–3602. https://doi.org/10.1121/1.428212

Brungart, D.S., Durlach, N. I., & Rabinowitz, W. M. (1999). Auditory localization of nearby sources. II. Localization of a broadband source. *The Journal of the Acoustical Society of America*, *106*(4), 1956–1968. https://doi.org/10.1121/1.427943

Brungart, Douglas S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*, 2112. https://doi.org/10.1121/1.1345696

Bruschi, V., Nobili, S., Cecchi, S., & Piazza, F. (2020, May 28). An Innovative Method for Binaural Room Impulse Responses Interpolation. *Audio Engineering Society Convention 148.*

C. C. Robusto. (1957). The Cosine-Haversine Formula. *The American Mathematical Monthly*, *64*(1), 38–40. http://www.jstor.org/stable/2309088

Carlile, S. (1996). Virtual Auditory Space: Generation and Applications. In *Virtual auditory space: Generation and applications* (Issue September). https://doi.org/10.1007/978-3-662-22594-3

Carlile, S., Jin, C. T., & Van Raad, V. (2000). Continuous virtual auditory space using HRTF interpolation: acoustic and psychophysical errors. *IEEE PacificRim Conference on Multimedia*, *August*, 220–223.

Carpentier, T. (2018). A new implementation of spat in Max. *Proceedings of the 15th Sound and Music Computing Conference: Sonic Crossings, SMC 2018*, *Umr 9912*, 184–191.

Carpentier, T., Noisternig, M., & Warusfel, O. (2015). *Twenty Years of Ircam Spat: Looking Back, Looking Forward*. 270–277. https://hal.archives-ouvertes.fr/hal-

01247594v1

Cherry, E. C. (1953a). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. https://doi.org/10.1121/1.1907229

Cherry, E. C. (1953b). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. https://doi.org/10.1121/1.1907229

Chowning, J. M. (1971). The simulation of moving sound sources. *Ournal of the Audio Engineering Society*, *9*(1), 2–6.

Colburn, H. S., & Durlach, N. I. (1978). Models of binaural interaction. In *Handbook of perception* (pp. 467–518).

Coleman, P. (1968). Dual Role of Frequency Spectrum in Determination of Auditory Distance. *Journal of the Acoustical Society of America*, *44*(2), 631–632. https://doi.org/10.1121/1.1911132

Colin Cherry, E., & Sayers, B. A. (1956). "Human 'Cross-Correlator'"-A Technique for Measuring Certain Parameters of Speech Perception. *Journal of the Acoustical Society of America*, *28*(5), 889–895. https://doi.org/10.1121/1.1908506

Comunità, M., Gerino, A., Lim, V., & Picinali, L. (2020). *PlugSonic: a web- and mobile-based platform for binaural audio and sonic narratives*. arXiv preprint arXiv:2008.04638.

Cooper, D. H., & Bauck, J. L. (1989). Prospects for Transaural Recording. *Journal of the Audio Engineering Society*, *37*(1/2), 3–19.

Cowan, B., & Kapralos, B. (2009). Real-time GPU-based convolution: A follow-up. *FuturePlay 2009 at GDC Canada International Conference on the Future of Game Design and Technology*, 25–26. https://doi.org/10.1145/1639601.1639617

Cowan, B., & Kapralos, B. (2008). Spatial sound for video games and virtual environments utilizing real-time gpu-based convolution. *ACM Future Play 2008 International Academic Conference on the Future of Game Design and Technology, Future Play: Research, Play, Share*, 166–172. https://doi.org/10.1145/1496984.1497012

Cruz-Neira, C., Sandin, D. J., & Defanti, T. A. (1993). Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE. *20th Annual Conference on Computer Graphics and Interactive Techniques*, 135–142.

Cuevas-Rodríguez, M., Picinali, L., González-Toledo, D., Garre, C., de la Rubia-Cuestas, E., Molina-Tanco, L., & Reyes-Lecuona, A. (2019). 3D Tune-In Toolkit: An open-source library for real-time binaural spatialisation. *PLOS ONE*, *14*(3), e0211899. https://doi.org/10.1371/journal.pone.0211899

Cuevas-Rodriguez, Maria, Alon, D. Lou, Clapp, S. W., Robinson, P. W., & Mehra, R.

(2019). Evaluation of the effect of head-mounted display on individualized head-related transfer functions. *Proceeding of the 23rd International Congress on Acoustic (9 to 13 September)*, 2635–2642.

Cuevas-Rodriguez, María, Gonzalez-Toledo, D., de La Rubia-Cuestas, E., Garre, C., Molina-Tanco, L., Reyes-Lecuona, A., Poirier-Quinot, D., & Picinali, L. (2017). An open-source audio renderer for 3D audio with hearing loss and hearing aid simulations. *142nd Convention Audio Engineering Society*, 1–8.

Cuevas-Rodriguez, Maria, Gonzalez-Toledo, D., La Rubia-Cuestas, E. De, Garre, C., Molina-Tanco, L., Reyes-Lecuona, A., Poirier-Quinot, D., & Picinali, L. (2018). The 3D Tune-In Toolkit - 3D audio spatialiser, hearing loss and hearing aid simulations. *2018 IEEE 4th VR Workshop on Sonic Interactions for Virtual Environments, SIVE 2018*, 1–3. https://doi.org/10.1109/SIVE.2018.8577076

Cuevas-Rodriguez, Maria, Gonzalez-Toledo, D., Reyes-Lecuona, A., & Picinali, L. (2021). Impact of non-individualised head-related transfer functions on speech-in-noise performances within a synthesised virtual environment. *Journal of the Acoustical Society of America*, *149*(April), 2573–2586. https://doi.org/10.1121/10.0004220

Culling, J., Hawley, M., & Litovsky, R. (2004). The role of head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources. *Journal of the Acoustical Society of America*, *116*(2), 1057–1065. https://doi.org/10.1121/1.1772396

Culling, J., Hawley, M., & Litovsky, R. (2005). Erratum: The role head-induced interaural time and level differences in the speech reception threshold for multiple interfering sound sources [J. Acoust. Soc. Am. 116, 1057 (2004)]. *The Journal of the Acoustical Society of America*, *118*(1), 552–552. https://doi.org/10.1121/1.1925967

Culling, J., & Summerfield, Q. (1995). Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay. *The Journal of the Acoustical Society of America*, *98*(2), 785–797. https://doi.org/10.1121/1.413571

Daniel, Jérôme, Nicol, R., & Moreau, S. (2003, May 1). Further Investigations of High Order Ambisonics and Wavefield Synthesis for Holophonic Sound Imaging. *114th Convention AES*.

Daniel, Jérome, Rault, J.-B., & Polack, J.-D. (1998). *Ambisonics Encoding of Other Audio Formats for Multiple Listening Conditions*. Audio Engineering Society.

de Cárdenas, M. R., & Marrero Aguiar, V. (1994). *Cuaderno de logoaudiometría. Guía de referencia rápida*. Universidad Nacional de Educaci{ó}n a Distancia, UNED.

de Vries, D., Start, E. W., & Valstar, V. G. (1994). *The Wave-Field Synthesis Concept Applied to Sound Reinforcement Restriction and Solutions*. Audio Engineering

Society.

Dietz, M., Ewert, S. D., & Hohmann, V. (2011). Auditory model based direction estimation of concurrent speakers from binaural signals. *Speech Communication*, *53*(5), 592–605. https://doi.org/10.1016/j.specom.2010.05.006

Duda, R. O. (1993). Modeling head related transfer functions. *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference On*, 996–1000.

Duda, R. O., Avendano, C., & Algazi, V. R. (1999). Adaptable ellipsoidal head model for the interaural time difference. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, *2*, 965–968. https://doi.org/10.1109/icassp.1999.759855

Duda, R. O., & Martens, W. L. (1998). Range dependence of the response of a spherical head model. *J. Acoust. Soc. Am.*, *104*(5), 3048–3058.

Duraiswaini, R., Zotkin, D. N., & Gumerov, N. A. (2004). Interpolation and range extrapolation of HRTFs [head related transfer functions]. *IEEE International Conference on Acoustics, Speech, and Signal Processing.*

Durlach, N. I. (1963). Equalization and Cancellation Theory of Binaural Masking-Level Differences. *The Journal of the Acoustical Society of America*, *35*(8), 1206–1218. https://doi.org/10.1121/1.1918675

Durlach, N. I., Rigopulos, A., Pang, X. D., Woods, W. S., Kulkarni, A., Colburn, H. S., & Wenzel, E. M. (1992). On the Externalization of Auditory Images. *Presence: Teleoperators and Virtual Environments*, *1*(2), 251–257. https://doi.org/10.1162/pres.1992.1.2.251

Edmonds, B. A., & Culling, J. F. (2006). The spatial unmasking of speech: Evidence for better-ear listening. *The Journal of the Acoustical Society of America*, *120*(3), 1539–1545. https://doi.org/10.1121/1.2228573

Engel, I., Alon, D. Lou, Robinson, P. W., & Mehra, R. (2019). The effect of generic headphone compensation on binaural renderings. *AES International Conference*, *2019-March*.

Engel, I., Craig, H., Amengual Garí, S. V., Robinson, P. W., & Picinali, L. (2021). Perceptual implications of different Ambisonics-based methods for binaural reverberation. *J. Acoust. Soc. Am.*, *149*(2), 895–819. https://doi.org/10.1121/10.0003437

Engel, I., Goodman, D. F. M., & Picinali, L. (2022). Assessing HRTF preprocessing methods for Ambisonics rendering through perceptual models. *Acta Acustica*, *6*(2). https://doi.org/10.1051/aacus/2021055

Enzer, G., Antweiler, C., & Spors, S. (2013). Trends in adfquisition of individualized head-related transfer functions. In J. Blauert (Ed.), *The Technology of Binaural*

Listening. Springer.

Farina, A. (2000, February 1). Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique. *Audio Engineering Society Convention 108. Audio Engineering Society, 2000.*

Farina, A. (2007). Advancements in impulse response measurements by sine sweeps. *122nd AES Convention*, 1–21. http://www.aes.org/e-lib/browse.cfm?elib=14106

Fels, J., & Vorländer, M. (2009). Anthropometric parameters influencing head-related transfer functions. *Acta Acustica United with Acustica*, *95*(2), 331–342. https://doi.org/10.3813/AAA.918156

Foster, S. H., & Wenzel, E. M. (1992). The Convolvotron: Real-time demonstration of reverberant virtual acoustic environments. *The Journal of the Acoustical Society of America*, *92*(4), 2376–2376. https://doi.org/10.1121/1.404833

Foster, S. H., Wenzel, E. M., & Taylor, M. R. (1991). Real time synthesis of complex acoustic environments. *1991 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*. https://doi.org/10.1109/ASPAA.1991.634098

Frank, M., Zotter, F., & Sontacchi, A. (2015). *Producing 3D Audio in Ambisonics*. Audio Engineering Society.

Freeland, F. P., Biscainho, L. W. P., & Diniz, P. S. R. (2004). Interpositional Transfer Function for 3D-Sound Generation. *Journal of the Audio Engineering Society*, *52*(9), 915–930. http://www.aes.org/e-lib/browse.cfm?elib=13019

Freyman, R. L., Helfer, K. S., Mccall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *The Journal of the Acoustical Society of America*, *106*, 833. https://doi.org/10.1121/1.428211

Fu, Q.-J., & Galvin, J. J. (2003). The effects of short-term training for spectrally mismatched noise-band speech. *The Journal of the Acoustical Society of America*, *113*, 1065. https://doi.org/10.1121/1.1537708

Gamper, H. (2013). Head-related transfer function interpolation in azimuth, elevation, and distance. *The Journal of the Acoustical Society of America*, *134*(6), EL547–EL553. https://doi.org/10.1121/1.4828983

Garcia-Gomez, V., & Lopez, J. J. (2018). *Binaural Room Impulse Responses Interpolation for Multimedia Real-Time Applications*. Audio Engineering Society.

Garcia, G. (2002). Optimal filter partition for efficient convolution with short input/output delay. *Audio Engineering Society Convention 113rd*, 1–9. http://www.aes.org/e-lib/browse.cfm?elib=11275

Gardner, W. G. (1994). Efficient convolution without input/output delay. *Audio Engineering Society Convention 97th*.

Gardner, W. G., & Martin, K. D. (1995). HRTF measurements of a KEMAR. *Journal of the Acoustical Society of America*, *97*(6), 3907–3908.

https://doi.org/10.1121/1.412407

Gauthier, P.-A., & Berry, A. (2006). Adaptive wave field synthesis with independent radiation mode control for active sound field reproduction: Theory. *The Journal of the Acoustical Society of America*, *119*(5), 2721–2737. https://doi.org/10.1121/1.2186514

Geier, M., Ahrens, J., & Spors, S. (2010). Object-based audio reproduction and the audio scene description format. *Organised Sound*, *15*(3), 219–227. https://doi.org/10.1017/S1355771810000324

Geier, M., & Spors, S. (2012). Spatial Audio with the SoundScape Renderer. *27th Tonmeistertagung - VDT International Convention*, 2012.

Gerzon, M. A. (1973). Periphony: With-Height Sound Reproduction. *Journal of the Audio Engineering Society*, *21*(1), 2–10.

Gerzon, M. A. (1985). Ambisonics in Multichannel Broadcasting and Video. *Journal of the Audio Engineering Society*, *33*(11), 859–871. http://www.aes.org/e-lib/browse.cfm?elib=4419

Google. (n.d.). *Resonance Audio*. Retrieved March 6, 2018, from https://developers.google.com/resonance-audio/develop/overview

Gorzel, M., Allen, A., Kelly, I., Kammerl, J., Gungormusler, A., Yeh, H., & Boland, F. (2019). *Efficient Encoding and Decoding of Binaural Sound with Resonance Audio*. Audio Engineering Society. http://www.aes.org/e-lib

Guldenschuh, M., Sontacchi, A., Zotter, F., & Höldrich, R. (2008). HRTF modelling in due consideration variable torso reflections. *Acoustical Society of America*, *123*(5), 99–104. https://doi.org/10.1121/1.2932888

Gupta, N., Barreto, a., & Ordonez, C. (2002). Improving sound spatialization by modifying head-related transfer functions to emulate protruding pinnae. *Proceedings IEEE SoutheastCon 2002 (Cat. No.02CH37283)*, 446–450. https://doi.org/10.1109/.2002.995637

Gupta, Navarun, Barreto, A., Joshi, M., & Agudelo, J. C. (2010). HRTF database at FIU DSP Lab. *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 169–172. https://doi.org/10.1109/ICASSP.2010.5496084

Hartmann, W. M., & Wittenberg, A. (1996). On the externalization of sound images. *The Journal of the Acoustical Society of America*, *99*(6), 3678–3688. https://doi.org/10.1121/1.414965

Hartung, K., Braasch, J., & Sterbing, S. J. (1999, March 1). Comparison of Different Methods for the Interpolation of Head-Related Transfer Functions. *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*.

Hawley, M. L., Litovsky, R. Y., & Culling, J. F. (2004). The benefit of binaural hearing

in a cocktail party: Effect of location and type of interferer. *The Journal of the Acoustical Society of America*, *115*(2), 833–843. https://doi.org/10.1121/1.1639908

Haykin, S., & Chen, Z. (2005). The cocktail party problem. *Neural Computation*, *17*(9), 1875–1902. https://doi.org/10.1162/0899766054322964

Hendrix, C., & Barfield, W. (1996). The sense of presence within auditory virtual environments. *Presence: Teleoperators and Virtual Environments*, *5*(3), 290–301. https://doi.org/10.1162/pres.1996.5.3.290

Herre, J., Hilpert, J., Kuntz, A., & Plogsties, J. (2015). MPEG-H 3D Audio - The New Standard for Coding of Immersive Spatial Audio. *IEEE Journal on Selected Topics in Signal Processing*, *9*(5), 770–779. https://doi.org/10.1109/JSTSP.2015.2411578

Hofman, P. M., Van Riswick, J. G. A., & Van Opstal, A. J. (1998). Relearning sound localization with new ears. *Nature Neuroscience*, *1*(5), 417–421. https://doi.org/10.1038/1633

Hosoe, S., Nishino, T., Itou, K., & Takeda, K. (2005). Measurement of Head-Related Transfer Functions in the Proximal Region. *Hosoe, S., Nishino, T., Itou, K., & Takeda, K.*, 2539–2542.

Hulsebos, E., De Vries, D., & Bourdillat, E. (2002). Improved microphone array configurations for auralization of sound fields by wave-field synthesis. *AES: Journal of the Audio Engineering Society*, *50*(10), 779–790.

Iec. (2002). *Electroacoustics-Sound level meters-Part 1: Specifications Electroacoustics-Sound level meters-Part 1: Specifications including photocopying and microfilm, without permission in writing from the publisher.* www.iec.ch

Inanaga, K., Yamada, Y., & Koizumi, H. (1995). *Headphone System with Out-of-Head Localization Applying Dynamic HRTF (Head-Related Transfer Function)*. Audio Engineering Society. http://www.aes.org/e-lib/browse.cfm?elib=7755

Ince, D. C., Hatton, L., & Graham-Cumming, J. (2012). The case for open computer programs. *Nature*, *482*(7386), 485–488. https://doi.org/10.1038/nature10836

IoSR. (n.d.). *MatlabToolbox/+iosr at master · IoSR-Surrey/MatlabToolbox*. Retrieved March 3, 2021, from https://github.com/IoSR-Surrey/MatlabToolbox

ISO 9613-1. (1993). *Attenuation of sound during propagation outdoors.*

ITU-R. (1993). *Recommendation BS.775: Multi-channel stereophonic sound system with or without accompanying picture*. International Telecommunications Union.

Iwaya, Y. (2006). Individualization of head-related transfer functions with tournament-style listening test: Listening with other's ears. *Acoustical Science and Technology*, *27*(6), 340–343. https://doi.org/10.1250/ast.27.340

Iwaya, Y., & Katz, B. F. G. (2018). Distributed signal processing architecture for real-time convolution of 3d audio rendering for mobile applications. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and*

Lecture Notes in Bioinformatics), *11162 LNCS*, 148–157. https://doi.org/10.1007/978-3-030-01790-3_9

Jeffress, L. A. (1948). A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, *41*(1), 35–39. https://doi.org/10.1037/h0061495

Jelfs, S., Culling, J. F., & Lavandier, M. (2011). Revision and validation of a binaural model for speech intelligibility in noise. *Hearing Research*, *275*(1–2), 96–104. https://doi.org/10.1016/j.heares.2010.12.005

Jenny, C., Majdak, P., & Reuter, C. (2018). SOFA Native spatializer plugin for Unity - Exchangeable HRTFs in virtual reality. *Audio Engineering Society 144th Convention*, 1–4. http://www.aes.org/e-lib/browse.cfm?elib=19519

Jeub, M., Schäfer, M., & Vary, P. (2009). A binaural room impulse response database for the evaluation of dereverberation algorithms. *DSP 2009: 16th International Conference on Digital Signal Processing, Proceedings*. https://doi.org/10.1109/ICDSP.2009.5201259

Jones, G. L., & Litovsky, R. Y. (2011). A cocktail party model of spatial release from masking by both noise and speech interferers. *The Journal of the Acoustical Society of America*, *130*(3), 1463–1474. https://doi.org/10.1121/1.3613928

Jot, J.-M., Larcher, V., & Warusfel, O. (1995). Digital Signal Processing Issues in the Context of Binaural and Transaural Stereophony. *98th AES Convention*, Convnetion Paper 3980. http://www.aes.org/e-lib/browse.cfm?elib=7786

Jot, J. M. (1999). Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces. *Multimedia Systems*, *7*(1), 55–69. https://doi.org/10.1007/s005300050111

Kahana, Y., & Nelson, P. A. (2007). Boundary element simulations of the transfer function of human heads and baffled pinnae using accurate geometric models. *Journal of Sound and Vibration*, *300*(3–5), 552–579. https://doi.org/10.1016/j.jsv.2006.06.079

Kaplanis, N., Bech, S., Jensen, S. H., & Van Waterschoot, T. (2014). Perception of reverberation in small rooms: A literature study. *Proceedings of the AES International Conference*, *2014-Janua*(April 2018).

Kapralos, B., Zikovitz, D., Jenkin, M. R., & Harris, L. R. (2004, May 1). Auditory Cues in the Perception of Self Motion. *Audio Engineering Society Convention 116th*.

Katz, B. F. G. (2001a). Boundary element method calculation of individual head-related transfer function. I. Rigid model calculation. *The Journal of the Acoustical Society of America*, *110*(5), 2440–2448. https://doi.org/10.1121/1.1412440

Katz, B. F. G. (2001b). Boundary element method calculation of individual head-related transfer function. II. Impedance effects and comparisons to real measurements. *The Journal of the Acoustical Society of America*, *110*(5), 2449–2455.

https://doi.org/10.1121/1.1412441

Katz, B. F. G., & Begault, D. R. (2007). Round robin comparison of HRTF measurement systems: preliminary results. *19th Intl. Cong. on Acoustics (ICA 2007)*.

Katz, B. F. G., Kammoun, S., Parseihian, G., Gutierrez, O., Brilhault, A., Auvray, M., Truillet, P., Denis, M., Thorpe, S., & Jouffrais, C. (2012). NAVIG: Augmented reality guidance system for the visually impaired: Combining object localization, GNSS, and spatial audio. *Virtual Reality*, *16*(4), 253–269. https://doi.org/10.1007/s10055-012-0213-6

Katz, B. F. G., & Noisternig, M. (2014). A comparative study of interaural time delay estimation methods. *The Journal of the Acoustical Society of America*, *135*(6), 3530–3540. https://doi.org/10.1121/1.4875714

Katz, B. F. G., & Parseihian, G. (2012). Perceptually based head-related transfer function database optimization. *The Journal of the Acoustical Society of America*, *131*(2), EL99-105. https://doi.org/10.1121/1.3672641

Kobayashi, M., Ueno, K., & Ise, S. (2015). The Effects of Spatialized Sounds on the Sense of Presence in Auditory Virtual Environments: A Psychological and Physiological Study. *Presence: Teleoperators and Virtual Environments*, *24*(2), 163–174. https://doi.org/10.1162/PRES_a_00226

Koehnke, J., & Besing, J. M. (1996). A procedure for testing speech intelligibility in a virtual listening environment. *Ear and Hearing*, *17*(3), 211–217. https://doi.org/10.1097/00003446-199606000-00004

Kolarik, A. J., Moore, B. C. J., Zahorik, P., Cirstea, S., & Pardhan, S. (2016). Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. *Attention, Perception, and Psychophysics*, *78*(2), 373–395. https://doi.org/10.3758/s13414-015-1015-1

Kreuzer, W., Majdak, P., & Chen, Z. (2009). Fast multipole boundary element method to calculate head-related transfer functions for a wide frequency range. *The Journal of the Acoustical Society of America*, *126*(3), 1280–1290. https://doi.org/10.1121/1.3177264

Kulkarni, A., & Colburn, H. S. (2000). Variability in the characterization of the headphone transfer-function. *The Journal of the Acoustical Society of America*, *107*(2), 1071–1074. https://doi.org/10.1121/1.428571

Kuttruff, H. (2016). *Room Acoustics*. Crc Press.

Kyriakakis, C., Tsakalides, P., & Holman, T. (1999). Acquisition and Rendering Methods for Immersive Audio. *IEEE Signal Processing Magazine*.

Langendijk, E. H. A., & Bronkhorst, A. W. (2000). Fidelity of three-dimensional-sound reproduction using a virtual auditory display. *The Journal of the Acoustical Society of America*, *107*(1), 528–537. https://doi.org/10.1121/1.428321

Langendijk, E. H. A., & Bronkhorst, A. W. (2002). Contribution of spectral cues to human sound localization. *The Journal of the Acoustical Society of America*, *112*(4), 1583–1596. https://doi.org/10.1121/1.1501901

Larsson, P., Västfjäll, D., & Kleiner, M. (2002). Better Presence and Performance in Virtual Environments By Improved Binaural Sound Rendering. *Journal of the Audio Engineering Society*, 1–8.

Lavandier, M., & Culling, J. F. (2010). Prediction of binaural speech intelligibility against noise in rooms. *The Journal of the Acoustical Society of America*, *127*(November 2009), 387–399. https://doi.org/10.1121/1.3268612

Lazzarini, V., & Carty, B. (2008). New Csound Opcodes for Binaural Processing. *Proc. 6th Linux Audio Conference, Köln*.

Lehnert, H., & Blauert, J. (1992). Principles of binaural room simulation. *Applied Acoustics*, *36*(3–4), 259–291. https://doi.org/10.1016/0003-682X(92)90049-X

Lei, W., & Xiangyang, Z. (2016). New method for synthesizing personalized head-related transfer function. *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 1–5. https://doi.org/10.1109/IWAENC.2016.7602913

Lentz, T., Assenmacher, I., Vorländer, M., & Kuhlen, T. (2006). Precise Near-to-Head Acoustics with Binaural Synthesis. *JVRB - Journal of Virtual Reality and Broadcasting*, *3(2006)*(2). https://doi.org/10.20385/1860-2037/3.2006.2

Lerner, F., Tahar, G., Bar, A., Koren, O., & Flash, T. (2021). VR Setup to Assess Peripersonal Space Audio-Tactile 3D Boundaries. *Frontiers in Virtual Reality*, *2*(May), 1–16. https://doi.org/10.3389/frvir.2021.644214

Levitt, H. (1971). Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477. https://doi.org/10.1121/1.1912375

Lindau, A., Hohn, T., & Weinzierl, S. (2007). Binaural resynthesis for comparative. *Audio Engineering Society Convention 122*.

Lindau, A., Maempel, H., & Weinzierl, S. (2008). Minimum BRIR grid resolution for dynamic binaural synthesis. *The Journal of the Acoustical Society of America*, *123*(5), 3498–3498. https://doi.org/10.1121/1.2934364

Lindau, A., & Weinzierl, S. (2012). Assessing the plausibility of virtual acoustic environments. *Acta Acustica United with Acustica*, *98*(5), 804–810. https://doi.org/10.3813/AAA.918562

Little, A. D., Mershon, D. H., & Cox, P. H. (1992). Spectral content as a cue to perceived auditory distance. *Perception*, *21*(3), 405–416. https://doi.org/10.1068/p210405

Lokki, T., & Pätynen, J. (2011). Lateral reflections are favorable in concert halls due to binaural loudness. *The Journal of the Acoustical Society of America*, *130*, 345. https://doi.org/10.1121/1.3647866

Lopez-Poveda, E. A., & Meddis, R. (1996). A physical model of sound diffraction and reflections in the human concha. *The Journal of the Acoustical Society of America*, *100*(5), 3248--3259.

Majdak, P., Baumgartner, R., & Laback, B. (2014). Acoustic and non-acoustic factors in modeling listener-specific performance of sagittal-plane sound localization. *Frontiers in Psychology*, *5*(APR), 319. https://doi.org/10.3389/fpsyg.2014.00319

Majdak, P., Iwaya, Y., Carpentier, T., Nicol, R., Parmentier, M., Roginska, A., Suzuki, Y., Watanabe, K., Wierstorf, H., Ziegelwanger, H., & Noisternig, M. (2013, May 4). Spatially Oriented Format for Acoustics: A Data Exchange Format Representing Head-Related Transfer Functions. *Audio Engineering Society Convention 134th*.

Malham, D. G. (1999). Higher Order Ambisonic Systems for the Spatialisation of Sound. *ICMC*. https://quod.lib.umich.edu/i/icmc/bbp2372.1999.451/--higher-order-ambisonic-systems-for-the-spatialisation?rgn=main;view=fulltext

Malham, D. G., & Myatt, A. (1995). 3-D Sound Spatialization using Ambisonic Techniques. *Computer Music Journal*, *19*(4), 58. https://doi.org/10.2307/3680991

Martin, R. L., McAnally, K. I., & Senova, M. A. (2001). Free-field equivalent localization of virtual audio. *AES: Journal of the Audio Engineering Society*, *49*(1–2), 14–22.

Masiero, B., & Fels, J. (2011). Perceptually Robust Headphone Equalization for Binaural Reproduction. *Audio Engineering Society Convention 130th*, 1–7.

Mayr, S., Buchner, A., Erdfelder, E., & Faul, F. (2007). A short tutorial of GPower. *Tutorials in Quantitative Methods for Psychology*, *3*(2), 51–59. https://doi.org/10.1037/0096-1523.32.4.932

Mccormack, L., & Politis, A. (2019). SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods. *Conference on Immersive and Interactive Audio of Audio Engineering Society*, *March*.

McKeag, A., & McGrath, D. S. (1996). Sound field format to binaural decoder with head tracking. *Audio Engineering Society 6th Australian Reagional Convention*, 1–9. http://www.aes.org/e-lib/browse.cfm?elib=7477

Meddis, R., & Lopez-Poveda, E. a. (2010). Auditory Periphery: From Pinna to Auditory Nerve. In *Hearing Research* (Vol. 35, Issue June). https://doi.org/10.1007/978-1-4419-5934-8

Meesawat, K., & Hammershøi, D. (2008). Prediction of binaural speech intelligibility against noise in rooms. *The Journal of the Acoustical Society of America*, *124*(2), 259–271. https://doi.org/10.1121/1.2996336

Mershon, D. H., & Bowers, J. N. (1979). Absolute and relative cues for the auditory perception of egocentric distance. *Perception*, *8*(3), 311–322. https://doi.org/10.1068/p080311

Microsoft. (2020). *Microsoft spatializer*. https://docs.microsoft.com/en-us/windows/mixed-reality/spatial-sound

Middlebrooks, J. C., & Green, D. M. (1991). *Sound Localization by human listeners*. www.annualreviews.org

Miller, J. D., & Wenzel, E. M. (2002). Recent developments in SLAB: A software based system for interactive spatial sound synthesis. *International Conference on Auditory Display*. https://humansystems.arc.nasa.gov/publications/20051025102054_Miller_2002_ICAD2002_SLAB.pdf

Moller, H. (1992). Fundamentals of Binaural Technology. *Applied Acoustics*, *36*(December 1991), 171–218.

Møller, H., Hammershøi, D., Jensen, C. B., & Sørensen, M. F. (1995). Transfer Characteristics of Headphones Measured on Human Ears. *Journal of the Audio Engineering Society*, *43*(4), 203–217.

Moller, H., Sorensen, M. F., Hammershoi, D., & Jense, C. B. (1995). Head-Related Transfer Functions of Human Subjects. *J. Audio Eng. Soc*, *43*(5).

Møller, H., Sørensen, M. F., Jensen, C. B., & Hammershøi, D. (1996). Binaural Technique: Do We Need Individual Recordings? *Journal of the Audio Engineering Society*, *44*(6), 451–469.

Moore, B. C. (2012). An introduction to the Psychology of Hearing. In *Brill*.

Moore, B., & Glasberg, B. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *Journal of the Acoustical Society of America*, *74*(3), 750–753. https://doi.org/10.1121/1.389861

Moreau, S., Daniel, J., & Bertet, S. (2006). 3D Sound Field Recording with Higher Order. *120th Convention of the AES*, 20–30.

Morimoto, M., & Ando, Y. (1980). On the simulation of sound localization. In *J. Acoust. Soc. Jpn. (E)* (Vol. 1).

Murphy, D., & Neff, F. (2010). Spatial sound for computer games and virtual reality. *Game Sound Technology and Player Interaction: Concepts and Developments*, *December 2013*, 287–312. https://doi.org/10.4018/978-1-61692-828-5.ch014

Nam, J., Abel, J. S., & Smith, J. O. (2008). A method for estimating interaural time difference for binaural synthesis. *Audio Engineering Society - 125th Audio Engineering Society Convention 2008*, *2*, 1343–1349.

Nicol, R. (2010). Binaural Technology. AES Monograph. In *Audio Engineering Society Inc.* http://www.aes.org/blog/2010/4/aes-publishes-monograph-on-binaural-technology

Nishino, T., Hosoe, S., Takeda, K., & Itakura, F. (2014). Measurement of the Head Related Transfer Function using the Spark Noise. *ICA2004*, *October*.

Nishino, T., Kajita, S., Takeda, K., & Itakura, F. (1999). Interpolating head related transfer functions in the median plane. *1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. WASPAA'99*, 167–170. https://doi.org/10.1109/ASPAA.1999.810876

Noisternig, M., Musil, T., Sontacchi, A., & Höldrich, R. (2003). 3D binaural sound reproduction using a virtual ambisonic approach. *VECIMS 2003 - 2003 International Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems*, 174–178. https://doi.org/10.1109/VECIMS.2003.1227050

Noisternig, Markus, Sontacchi, A., Musil, T., & Holdrich, R. (2003, June 1). A 3D Ambisonic Based Binaural Sound Reproduction System. *AES 24th International Conference: Multichannel Audio, The New Reality*.

Oberem, J., Richter, J.-G., Setzer, D., Seibold, J., Koch, I., & Fels, J. (2020). Experiments on localization accuracy with non-individual and individual HRTFs comparing static and dynamic reproduction methods. *Preprint Version*.

Oculus VR. (2020). *Oculus audio SDK*. https://developer.oculus.com/documentation/native/audio-spatializer-features//

Ooura, T. (2001). *General Purpose FFT (Fast Fourier/Cosine/Sine Transform) Package*. http://www.kurims.kyoto-u.ac.jp/~ooura/fft.html

Oppenheim, A. V. (1999). *Discrete-time signal processing*. Pearson Education India.

Ortiz, J. A., & Wright, B. A. (2009). Contributions of Procedure and Stimulus Learning to Early, Rapid Perceptual Improvements. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(1), 188–194. https://doi.org/10.1037/a0013161

P. Majdak, Balazs, P., & Laback, B. (2007). Multiple Exponential Sweep Method for Fast Measurement of Head-Related Transfer Functions. *J. Audio Eng. Soc*, *55*(7/8), 623–637.

Parker, S. P. A., Eberle, G., Martin, R. L., & Mcanally, K. I. (2008). *Construction of 3-D Audio Systems: Background, Research and General Requirements*.

Parseihian, G., & Katz, B. F. G. (2012). Rapid head-related transfer function adaptation using a virtual auditory environment. *The Journal of the Acoustical Society of America*, *131*(4), 2948–2957. https://doi.org/10.1121/1.3687448

Patterson, R. D., Allerhand, M. H., & Giguere, C. (1995). Time-domain modeling of peripheral auditory system - patterson. *The Journal of the Acoustical Society of America*, *98*(4), 1890–1894.

Paul, S. (2009). Binaural recording technology: A historical review and possible future developments. In *Acta Acustica united with Acustica* (Vol. 95, Issue 5, pp. 767–788). https://doi.org/10.3813/AAA.918208

Peissig, J., & Kollmeier, B. (1997). Directivity of binaural noise reduction in spatial multiple noise-source arrangements for normal and impaired listeners. *The Journal of the Acoustical Society of America*, *101*(3), 1660–1670. https://doi.org/10.1121/1.418150

Pellegrini, R., & Kuhn, C. (2004). *Wave Field Synthesis: Mixing and Mastering Tools for Digital Audio Workstations*. Audio Engineering Society.

Perrett, S., & Noble, W. (1997). The effect of head rotations on vertical plane sound localization. *The Journal of the Acoustical Society of America*, *102*(4), 2325–2332. https://doi.org/10.1121/1.419642

Picinali, L., Afonso, A., Denis, M., & Katz, B. F. G. (2014). Exploration of architectural spaces by blind people using auditory virtual reality for the construction of spatial knowledge. *International Journal of Human-Computer Studies*, *72*(4), 393–407. https://doi.org/10.1016/j.ijhcs.2013.12.008

Picinali, L., Wallin, A., Levtov, Y., & Poirier-Quinot, D. (2017). Comparative perceptual evaluation between different methods for implementing Reverberation in a binaural context. *142nd Audio Engineering Society International Convention 2017, AES 2017*, *May*.

Poirier-Quinot, D., & Katz, D. F. G. (2018). The Anaglyph binaural audio engine. *144th AES Convention*, 1–4.

Pompidou. (2014). *LibSOFA C++ Library*. http://sofacoustics.org/data/database/

Pralong, D., & Carlile, S. (1996). The role of individualized headphone calibration for the generation of high fidelity virtual auditory space. *The Journal of the Acoustical Society of America*, *100*(6), 3785–3793. https://doi.org/10.1121/1.417337

Products | mhacoustics.com. (n.d.). Retrieved June 15, 2020, from https://mhacoustics.com/products

Pujol, R. (2020). *Auditory Brain | Cochlea*. http://www.cochlea.eu/en/auditory-brain

Pulkki, V. (1997). Virtual Sound Source Positioning Using Vector Base Amplitude Panning. *Journal of the Audio Engineering Society*, *45*(6), 456–466.

Pulkki, V., Karjalainen, M., & Huopaniemi, J. (1999). Analyzing Virtual Sound Source Attributes Using a Binaural Auditory Model. *Journal of the Audio Engineering Society*, *47*(4), 203–217. http://www.aes.org/e-lib/browse.cfm?elib=12110

Rayleigh, L., & Lodge, A. (1904). On the Acoustic Shadow of a Sphere. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *203*(359–371), 87–110. https://doi.org/10.1098/rsta.1904.0016

Rayleigh, Lord. (1907). XII. On our perception of sound direction. *Philosophical Magazine Series 6*, *13*(74), 214–232. https://doi.org/10.1080/14786440709463595

RealSpace3D. (2015). *VisiSonics-RealSpace3D Audio User Guide*. www.realspace3daudio.com.

Reed, M. C., & Blum, J. J. (1990). A model for the computation and encoding of azimuthal information by the lateral superior olive. *The Journal of the Acoustical Society of America*, *88*(3), 1442–1453.

Reijniers, J., Vanderelst, D., Jin, C., Carlile, S., & Peremans, H. (2014). An ideal-observer model of human sound localization. *Biological Cybernetics*, *108*(2), 169–181. https://doi.org/10.1007/s00422-014-0588-4

Reijniers, Jonas, Partoens, B., Steckel, J., & Peremans, H. (2020). HRTF Measurement by Means of Unsupervised Head Movements with Respect to a Single Fixed Speaker. *IEEE Access*, *8*, 92287–92300. https://doi.org/10.1109/ACCESS.2020.2994932

Reilly, A., & McGrath, D. (1995). *Convolution Processing for Realistic Reverberation*. Audio Engineering Society.

Reyes-Lecuona, A., Cuevas-Rodriguez, M., Gonzalez-Toledo, D., Molina-Tanco, L., & Picinali, L. (2021). Speech perception in VR: do we need individual recordings? *International Conference on Immersive and 3D Audio*, *101017743*, 1–1. https://doi.org/10.1109/i3da48870.2021.9610938

Reyes-Lecuona, A., Márquez-Moncada, A., Hauke Luis, B., González-Toledo, D., Cuevas-Rodriguez, M., & Molina-Tanco, L. (2021). Audio Binaural y Ganancia de Rotación en Entornos Virtuales Binaural Audio and Rotation Gain in Virtual Environments. *Interaccion - Revista Digital de AIPO*, *2*(2), 54–62.

Rocchesso, D., & Smith, J. O. (1997). Circulant and elliptic feedback delay networks for artificial reverberation. *IEEE Transactions on Speech and Audio Processing*, *5*(1), 51–63. https://doi.org/10.1109/89.554269

Roginska, A., & Geluso, P. (2017). *Immersive sound : the art and science of binaural and multi-channel audio*. Taylor & Francis.

Romblom, D., & Cook, B. (2008). *Near-Field Compensation for HRTF Processing*. Audio Engineering Society. http://www.aes.org/e-lib/browse.cfm?elib=14762

Romigh, G. D., Brungart, D. S., Stern, R. M., & Simpson, B. D. (2015). Efficient Real Spherical Harmonic Representation of Head-Related Transfer Functions. *IEEE Journal of Selected Topics in Signal Processing*, *9*(5), 921–930. https://doi.org/10.1109/JSTSP.2015.2421876

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, *1*(1), 43–46. https://doi.org/10.1097/00001648-199001000-00010

Rozenn, N., Laetitia, G., Cathy, C., Brian FG, K., & Laurent SR, S. (2014). a Roadmap for Assessing the Quality of Experience of 3D Audio Binaural Rendering. *EAA Joint Symposium on Auralization and Ambisonics*, 100–106.

Rumsey, F. (2001). Spatial Audio. *Spatial Audio*, 1–240. https://doi.org/10.4324/9780080498195

SADIE | Spatial Audio For Domestic Interactive Entertainment. (n.d.). Retrieved

November 4, 2018, from https://www.york.ac.uk/sadie-project/index.html

Sandvad, J. (1996). Dynamic Aspects of Auditory Virtual Environments. *100th AES Convention*, *4226*, Convention Paper 4226. http://www.aes.org/e-lib/browse.cfm?elib=7547

Saville, D. J. (2014). Multiple comparison procedures—cutting the gordian knot. *Agronomy Journal*, *107*(2), 730–735. https://doi.org/10.2134/agronj2012.0394

Savioja, L., Huopaniemi, J., Lokki, T., & Väänänen, R. (1999). Creating interactive virtual acoustic environments. *AES: Journal of the Audio Engineering Society*, *47*(9), 675–704.

Savioja, Lauri, Välimäki, V., & Smith, J. O. (2011). Audio signal processing using graphics processing units. *AES: Journal of the Audio Engineering Society*, *59*(1–2), 3–19.

Schärer, Z., & Lindau, A. (2009). Evaluation of equalization methods for binaural signals. *126th Audio Engineering Society Convention 2009*, *1*, 15–31.

Schimmel, S. M., Muller, M. F., & Dillier, N. (2009). A fast and accurate "shoebox " room acoustics simulator. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, *May 2009*, 241–244. https://doi.org/10.1109/ICASSP.2009.4959565

Schissler, C., Nicholls, A., & Mehra, R. (2016). Efficient HRTF-based Spatial Audio for Area and Volumetric Sources. *IEEE Transactions on Visualization and Computer Graphics*, *22*(4), 1356–1366. https://doi.org/10.1109/TVCG.2016.2518134

Schonstein, D., Ferré, L., & Katz, B. F. G. (2008). Comparison of headphones and equalization for virtual auditory source localization. *Proceedings - European Conference on Noise Control*, 4617–4622. https://doi.org/10.1121/1.2935199

Schörkhuber, C., Zaunschirm, M., & Höldrich, R. (2018). Binaural rendering of Ambisonic signals via magnitude least squares. *Fortschritte Der Akustik -- DAGA 2018*, *March*, 339–342.

Schroeder, M. R. (1965). New Method of Measuring Reverberation Time. *The Journal of the Acoustical Society of America*, *37*(6), 1187–1188. https://doi.org/10.1121/1.1939454

Schroeder, M. R., & Logan, B. F. (1961). "Colorless" Artificial Reverberation. *IRE Transactions on Audio*, *9*(6), 209–214. https://doi.org/10.1109/TAU.1961.1166351

Seeber, B. U., Fastl, H., & Others. (2003). Subjective selection of non-individual head-related transfer functions. *2003 International Conference on Auditory Display*, 1–4. https://smartech.gatech.edu/handle/1853/50488

Serafin, S., Geronazzo, M., Erkut, C., Nilsson, N. C., & Nordahl, R. (2018). Sonic Interactions in Virtual Reality: State of the Art, Current Challenges, and Future Directions. *IEEE Computer Graphics and Applications*, *38*(2), 31–43.

https://doi.org/10.1109/MCG.2018.193142628

Shinn-Cunningham, B. G. (2000). *Distance cues for virtual auditory space*. 13–15. http://cns.bu.edu/~shinn/pages/pdf/IEEE_distance.pdf

Shinn-Cunningham, B. G., Kopco, N., & Martin, T. J. (2005). Localizing nearby sound sources in a classroom: Binaural room impulse responses. *The Journal of the Acoustical Society of America*, *117*(5), 3100–3115. https://doi.org/10.1121/1.1872572

Sivonen, V. P., & Ellermeier, W. (2011). *Binaural Loudness* (pp. 169–197). Springer, New York, NY. https://doi.org/10.1007/978-1-4419-6712-1_7

Slaney, M. (1998). *Auditory Toolbox*.

Smith, J. O. (1985). A new approach to digital reverberation using closed waveguide networks. *Proc. Int. Computer Music Conf, Vancouver, 1985*, 47–53. https://ci.nii.ac.jp/naid/10026801357

Snow, W. B. (1953). Basic Principles of Stereophonic Sound. *Journal of the Society of Motion Picture and Television Engineers*, *61*(5), 567–589. https://doi.org/10.5594/j00963

Sodnik, J., Sušnik, R., Štular, M., & Tomažič, S. (2005). Spatial sound resolution of an interpolated HRIR library. *Applied Acoustics*, *66*(11), 1219–1234. https://doi.org/10.1016/J.APACOUST.2005.04.003

SOFA General Purpose Database. (2017). https://github.com/sofacoustics/API_Cpp

SOFA Matlab/Octave API (Github). (2007). https://github.com/sofacoustics/API_MO

Søndergaard, P. L., & Majdak, P. (2013). The Auditory Modeling Toolbox. In J. Blauert (Ed.), *The Technology of Binaural Listening* (pp. 33–56). Springer Berlin / Heidelberg. http://amtoolbox.sourceforge.net/notes/amtnote007.pdf

Sound, M. S. (2020). *MS HRTF Spatializer and Microsoft spatializer*. https://docs.microsoft.com/es-es/windows/mixed-reality/spatial-sound-in-unity

Spagnol, S., Geronazzo, M., & Avanzini, F. (2013). On the Relation Between Pinna Reflection Patterns and Head-Related Transfer Function Features. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(3), 508–519. https://doi.org/10.1109/TASL.2012.2227730

Spors, S., Spors, S., Rabenstein, R., & Ahrens, J. (2008). The theory of wave field synthesis revisited. *IN 124TH CONVENTION OF THE AES*. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.391.7275

Steadman, M. A., Kim, C., Lestang, J. H., Goodman, D. F. M., & Picinali, L. (2019). Short-term effects of sound localization training in virtual reality. *Scientific Reports*, *9*(1), 1–17. https://doi.org/10.1038/s41598-019-54811-w

SteamAudio. (2014). *Steam Audio API: Main Page*. https://valvesoftware.github.io/steam-audio/doc/capi/index.html

Stern, R. M. (1988). An overview of models of binaural perception. *1988 National Research Council CHABA Symposium, Washington, DC, USA.*

Stitt, P., Hendrickx, E., Messonnier, J.-C., & Katz, B. (2016, May 26). The Influence of Head Tracking Latency on Binaural Rendering in Simple and Complex Sound Scenes. *In Audio Engineering Society Convention 140th.*

Sunder, K., He, J., Tan, E.-L., & Gan, W.-S. (2015). Natural sound rendering for headphones: integration of signal processing techniques. *EEE Signal Processing Magazine*, *32*(2), 100–113. https://doi.org/10.1109/MSP.2014.2372062.

Takemoto, H., Mokhtari, P., Kato, H., Nishimura, R., & Iida, K. (2012). Mechanism for generating peaks and notches of head-related transfer functions in the median plane. *The Journal of the Acoustical Society of America*, *132*(6), 3832–3841. https://doi.org/10.1121/1.4765083

The Auditory Modeling Toolbox. (n.d.). Retrieved March 17, 2020, from http://amtoolbox.sourceforge.net/

Theile, G., Wittek, H., & Reisinger, M. (2003). *Potential Wavefield Synthesis Applications in the Multichannel Stereophonic World*. Audio Engineering Society.

Torger, A., & Farina, A. (2001). Real-time partitioned convolution for ambiophonics surround sound. *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, *October*, 195–198. https://doi.org/10.1109/aspaa.2001.969576

Torick, E. (1998). Highlights in the History of Multichannel Sound. *Journal of the Audio Engineering Society*, *46*(1/2), 27–31.

Välimäki, V., Parker, J. D., Savioja, L., Smith, J. O., & Abel, J. S. (2012). Fifty Years of Artificial Reverberation. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(5), 1421–1448. https://doi.org/10.1109/TASL.2012.2189567

Välimäki, V., Parker, J., Savioja, L., Smith, J. O., & Abel, J. (2016). More Than 50 Years of Artificial Reverberation. *Proc. AES 60th International Conference*, K-1. http://www.aes.org/e-lib/browse.cfm?elib=18061

Väljamäe, A., Larsson, P., Västfjäll, D., & Kleiner, M. (2004). Auditory Presence, Individualized Head-Related Transfer Functions, and Illusory Ego-Motion in Virtual Environments. *Seventh Annual Workshop Presence 2004*, *January*, 141–147. http://publications.lib.chalmers.se/publication/18536-auditory-presence-individualized-head-related-transfer-functions-and-illusory-ego-motion-in-virtual

Valve_Software. (n.d.). *Steam Audio SDK*. Retrieved October 8, 2019, from https://valvesoftware.github.io/steam-audio/

Völk, F., Konradl, J., & Fastl, H. (2008). Simulation of wave field synthesis. *Proceedings - European Conference on Noise Control*, 1165–1170. https://doi.org/10.1121/1.2933196

Völk, Florian. (2014). Inter- and intra-individual variability in the blocked auditory

canal transfer functions of three circum-aural headphones. *AES: Journal of the Audio Engineering Society*, *62*(5), 315–323. https://doi.org/10.17743/jaes.2014.0021

Vorländer, M., Schröder, D., Wefers, F., Pelzer, S., Rausch, D., & Kuhlen, T. (2010). Virtual reality system at RWTH Aachen University. *International Symposium on Room Acoustics, ISRA 2010*, *August*, 9. http://www.acoustics.asn.au/conference_proceedings/ICA2010/cdrom-ISRA2010/Papers/P4h.pdf

Warusfel, O. (2003). *LISTEN HRTF DATABASE*. http://recherche.ircam.fr/equipes/salles/listen/

Watanabe, K., Iwaya, Y., Suzuki, Y., Takane, S., & Sato, S. (2014). Dataset of head-related transfer functions measured with a circular loudspeaker array. *Acoustical Science and Technology*, *35*(3), 159–165. https://doi.org/10.1250/ast.35.159

Wefers, F. (2015). Partitioned convolution algorithms for real-time auralization [Aachen University]. In *RWTH Aachen University* (Vol. 20, Issue August). http://publications.rwth-aachen.de/record/466561

Wefers, F., & Vorländer, M. (2011). Optimal filter partitions for real-time FIR filtering using uniformly-partitioned FFT-based convolution in the frequency-domain. *14th International Conference on Digital Audio Effects, DAFx 2011*, *1*, 155–162.

Wenzel, E. M., Wightman, F. L., & Foster, S. H. (1988). A virtual display system for conveying three-dimensional acoustic information. *In Proceedings of the Human Factors Society Annual Meeting*, 86–90.

Wenzel, E., Miller, J., & Abel, J. (2000). Sound Lab: A real-time, software-based system for the study of spatial hearing. *108th AES Convention*, 1–27.

Wenzel, Elizabeth M. (1995). Relative contribution of interaural time and magnitude cues to dynamic sound localization. *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*. https://doi.org/10.1109/aspaa.1995.482963

Wenzel, Elizabeth M. (2001). Effect of Increasing System Latency on Localization of Virtual Sounds with Short and Long Duration. *2001 International Conference on Auditory Display*, ICAD01-185--190.

Wenzel, Elizabeth M., Arruda, M., Kistler, D. J., & Wightman, F. L. (1993). Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, *94*(1), 111–123. https://doi.org/10.1121/1.407089

Wenzel, Elizabeth M., & Foster, S. H. (1993). Perceptual consequences of interpolating head-related transfer functions during spatial synthesis. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 1993*, 102–105. https://doi.org/10.1109/ASPAA.1993.379986

Wenzel, Elizabeth M. (1998). The Impact of System Latency on Dynamic Performance In Virtual Acoustic Environments. *Article in The Journal of the Acoustical Society of America*, 2405–2406. https://doi.org/10.1121/1.422547

Werner, S., Klein, F., Mayenfels, T., & Brandenburg, K. (2016). A summary on acoustic room divergence and its effect on externalization of auditory events. *2016 8th International Conference on Quality of Multimedia Experience, QoMEX 2016*, *March 2019*. https://doi.org/10.1109/QoMEX.2016.7498973

Wiener, F. M., & Ross, D. A. (1946). The Pressure Distribution in the Auditory Canal in a Progressive Sound Field. *Citation: The Journal of the Acoustical Society of America*, *18*, 248. https://doi.org/10.1121/1.1902437

Wightman, F., Kistler, D., & Arruda, M. (1992). Perceptual consequences of engineering compromises in synthesis of virtual auditory objects. *The Journal of the Acoustical Society of America*, *92*(4), 2332–2332. https://doi.org/10.1121/1.404982

Wightman, F. L., & Kistler, D. J. (1989a). Headphone simulation of free-field listening. I: Stimulus synthesis. *The Journal of the Acoustical Society of America*, *85*(2), 858–867.

Wightman, F. L., & Kistler, D. J. (1989b). Headphone simulation of free-field listening. II: Psychophysical validation. *The Journal of the Acoustical Society of America*, *85*(2), 868–878.

Wightman, L., & Kistler, D. J. (1996). *Monaural Sound Localization revisited*. *2*, 1050–1063.

Woodworth, R. S., Barber, B., & Schlosberg, H. (1954). *Experimental psychology*. Oxford and IBH Publishing.

Woodworth, R. S., & Schlosberg, H. (1954). *Experimental psychology, Rev. ed*. https://psycnet.apa.org/record/1955-00027-000

Wu, R., & Yu, G. (2016). Improvements in HRTF dataset of 3D game audio application. *Nternational Conference on Audio, Language and Image Processing (ICALIP) IEEE*, 185–190.

Xie, B. (2013). *Head-Related Transfer Function and Virtual Auditory Display* (Second Edi). www.jrosspub.com

Xu, S., Li, Z., & Salvendy, G. (2007). Individualization of Head-Related Transfer Function for Three-Dimensional Virtual Auditory Display : A Review. *12th International Conference on Human-Computer Interaction (HCI International 2007)*, 397--407. https://doi.org/10.1007/978-3-540-73335-5_44

Yairi, S., Iwaya, Y., & Suzuki, Y. (2007). Estimation of detection threshold of system latency of virtual auditory display. *Applied Acoustics*, *68*(8), 851–863. https://doi.org/10.1016/j.apacoust.2006.12.005

Yao, S. N., Collins, T., & Liang, C. (2017). Head-related transfer function selection using

neural networks. *Archives of Acoustics*, *42*(3), 365–373. https://doi.org/10.1515/aoa-2017-0038

Yu, G., Wu, R., Liu, Y., & Xie, B. (2018). Near-field head-related transfer-function measurement and database of human subjects. *The Journal of the Acoustical Society of America*, *143*(3), EL194–EL198. https://doi.org/10.1121/1.5027019

Zacksenhouse, M., Johnson, D. H., & Tsuchitani, C. (1992). Excitatory/inhibitory interaction in the LSO revealed by point process modeling. *Hearing Research*, *62*(1), 105–123. https://doi.org/10.1016/0378-5955(92)90207-4

Zahorik, P. (2002). Assessing auditory distance perception using virtual acoustics. *The Journal of the Acoustical Society of America*, *111*(4), 1832–1846. https://doi.org/10.1121/1.1458027

Zahorik, P., Brungart, D. S., & Bronkhorst, A. W. (2005). *Auditory Distance Perception in Humans: A Summary of Past and Present Research. 91*, 409–420.

Zhang, M., Kennedy, R. A., Abhayapala, T. D., & Zhang, W. (2011). Statistical method to identify key anthropometric parameters in hrtf individualization. *2011 Joint Workshop on Hands-Free Speech Communication and Microphone Arrays, HSCMA'11*, 213–218. https://doi.org/10.1109/HSCMA.2011.5942401

Zhang, W., Samarasinghe, P., Chen, H., & Abhayapala, T. (2017). Surround by Sound: A Review of Spatial Audio Recording and Reproduction. *Applied Sciences*, *7*(5), 532. https://doi.org/10.3390/app7050532

Zotkin, D. N., Duraiswami, R., & Davis, L. S. (2002). Customizable Auditory Displays. *International Conference on Auditory Display.*

Zotkin, D. N., Duraiswami, R., & Davis, L. S. (2004). Rendering Localized Spatial Audio in a Virtual Auditory Space. *IEEE Transactions on Multimedia*, *6*(4), 553–564. https://doi.org/10.1109/TMM.2004.827516

Zotkin, D. N., Hwang, J., Duraiswaini, R., & Davis, L. S. (2003). HRTF personalization using anthropometric measurements. *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No.03TH8684)*, 157–160. https://doi.org/10.1109/ASPAA.2003.1285855

Zotter, F., & Frank, M. (2019). *Ambisonics. A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality.* Springer. http://www.springer.com/series/8109